RAPPORT DE RECHERCHE

L R I

BioGuideSRS : QUERING MULTIPLE SOURCES WITH A USER PERSPECTIVE

COHEN BOULAKIA S / BITON O / FROIDEVAUX C

Unité Mixte de Recherche 8623
CNRS-Université Paris Sud – LRI

03/2006

Rapport de Recherche N° 1436

CNRS – Université de Paris Sud
Centre d'Orsay
LABORATOIRE DE RECHERCHE EN INFORMATIQUE
Bâtiment 490
91405 ORSAY Cedex (France)

BioGuideSRS: Querying Multiple Sources with a User Perspective

Sarah Cohen-Boulakia[1,2], Olivier Biton[2**], and Christine Froidevaux[1]

[1] Laboratoire de Recherche en Informatique, UMR CNRS 8623,
Universite Paris-Sud XI, Bat 490, 91405 Orsay Cedex, France
cohen@lri.fr, chris@lri.fr
[2] Department of Computer Science
University of Pennsylvania, 3330 Walnut St, PA-19103 Philadelphia, USA
sarahcb@seas.upenn.edu, biton@seas.upenn.edu

Abstract: *Biologists are faced with the problem of integrating information from multiple heterogeneous public sources with their own experimental data contained in individual sources. It is difficult for them to choose between the numerous sources without assistance. Interestingly, biologists not only express preferences concerning the sources to be queried but also differ in their strategies i.e. the querying process they follow for navigating through the sources. In [3] we have introduced BioGuide, as a user-centric framework that allows to specify various querying processes and takes into account user preferences on the sources. In this paper, we present BioGuideSRS, a user-friendly system which accesses instances of data by using BioGuide on top of the SRS system. We show how BioGuideSRS meets the needs collected during our study of user's requirements and illustrate its use with some examples.*
Availability: *http://www.lri.fr/∼cohen/bioguideSRS/bioguideSRS.html*

Keywords: Integrating and querying data sources, user's preferences, browsing, navigating.

1 Introduction

The number and size of biological data sources have increased exponentially in the last few years. Providing solutions to select relevant sources is one of the current challenges in bioinformatics since these sources are complementary, focus on different objects, and reflect various experts' points of view.

For the past ten years, several solutions have been designed to help scientists to analyze their data. For instance, the European HKIS[3] project (2002-2004) developed an integrative software platform for biological data processing in oncology. The HKIS-platform provides a set of predefined scenarios reflecting several possible ways for analyzing biological data. The global approach of a HKIS-platform user is to select a scenario and to adapt it to her own need (by modifying and/or completing it). However, at each step of a scenario, the user may need to consult external sources to get data. As the number of sources a scientist uses regularly is small compared to the number of sources that are available on the Web, the biologist may thus not exploit the full richness of the data contained in these sources. In the context of the HKIS-project, we especially perceived how useful it is to consult

** Current affiliation.
[3] www.hkis-project.com

alternative sources (i) to get complementary results but also (ii) to detect and resolve conflicts.
Let us consider the following query *Where are all the BACs of my CGH array located on the genome sequence?*. This query, borrowed from the HKIS BAC augmentation scenario [4], aims to find information about BACs used in a CGH array experiment. By using alternative sources we get complementary results concerning instances of BACs (cf. (i)). For example, accessing only MapViewFish[4] allows to locate BAC RP11-89F21 on a given chromosome band whereas accessing UCSCGenome[5] gives the exact position of this BAC on the chromosome sequence [4]. We also highlight that using a single or a few sources can hide conflicts that would be detected by using alternative sources (cf. (ii)). Indeed GenBank and MapView localize the BAC CTD-2012D15 on chromosome X, while UCSC Genome and MapViewFish localize it on chromosome 11. In such a situation, the selection of a single path, *i.e.* a sequence of sources to be consulted and links to be considered, will lead to a result potentially unsatisfactory. If the user knows the two paths she will choose one rather than another depending on her relative confidence into the sources.

Consequently, it is necessary to help the user choose among the numerous sources. It is critical to explore alternative selections of biological sources in order to obtain complementary pieces of information. Moreover, it is also important to take into account the confidence the user has in the sources (e.g. reliability) during this process.

To generalize this work to other biological domains, we performed a thorough analysis of scientists' needs during the querying process by developing a questionnaire and conducting interviews with scientists working in various domains (especially *studies of diseases*, *functional* and *structural genomics*). Our study emphasized that scientists express *preferences* concerning the sources queried. Moreover, this study revealed that the process of querying itself – the *strategy* followed for navigating through the sources – varies from one scientist to another.

In response to these findings, we have designed the BioGuide [3] framework, which provides scientists with support during the querying process. BioGuide assists the scientist in data searches within sources, providing information concerning the sequences of sources to be consulted and the links to be considered thus giving the *paths* between sources to be followed. In this paper, we present the BioGuideSRS system, which accesses instances of data by using the BioGuide framework on top of the SRS system. BioGuideSRS allows the user to automatically select sources relevant to her query and to automatically get the corresponding answers, in a uniform way. We show how BioGuideSRS meets the needs collected during our study of user's requirements and provides interesting answers.

This paper is organized as follows: We first present how BioGuideSRS supports the user in the querying process (Section 2). We then emphasize the flexibility of BioGuideSRS as a user-centric approach (Section 3). Section 4 is dedicated to the architecture of BioGuideSRS and gives some examples of its use. Section 5 compares our work to previous works and concludes the paper.

[4] The NCBI MapView bank is available at http://www.ncbi.nlm.nih.gov/mapview/. It is split into two different sources: MapViewFish and MapView (Fish mapping or not).
[5] The UCSC genome is available at http://www.genome.ucsc.edu/cgi-bin/hgGateway

2 Support for questions

As a result of our study of user's requirements, a set of 156 scientific questions expressed in natural language have been collected[6].

Examples of such questions are:

1. *Which are the functional domains of this set of proteins that are involved in the same pathway?*
2. *What is known about the gene BCR1 and the diseases it may cause in the OMIM source?*
3. *Which information may I get about genetic sleep disorders such as narcolepsy and the genes related to these diseases?*

The analysis of the answers to this questionnaire led us to distinguish two kinds of queries in BioGuideSRS depending on whether the biological question explicitly mentions sources (expressed as mixed query) or not (expressed as transparent query). It also raised the need for knowing the origin of the data (traceability) and we noticed that various ways of navigating through the sources do exist.

2.1 Transparent queries

Our analysis of the collected scientific questions revealed that, in many cases, neither the data sources nor the links to be used were specified by the biologists. The scientists rather poses a question referring to biological *entities* and *relationships* between them. Entities are often associated with specific names or ids. For instance in question 2, the underlying entities of the question are GENE and DISEASE and the underlying relationship is *causes*. The name of the gene for which information is sought is specified (BCR1).

We extracted entities and relationships from the collected questions and used the answers given during interviews to build the **graph of entities** (Fig. 1 left hand side). This graph provides the user with a graphical representation of the biological domain and supports her in the expression of her question at conceptual level. She can map the various components of her question (e.g. "gene BCR1", "this set of proteins") to higher level biological objects (GENE, PROTEIN).

In this graph, nodes are biological entities and labeled edges are biological relationships between them (relationships are symmetric). This graph models biological knowledge (e.g. *proteins are encoded by genes*, *genes cause diseases* etc.). We will see in the next section how this graph can be adapted to each user's needs.

In BioGuideSRS, scientists can make use of this graph to build *transparent queries* by selecting entities and, possibly, relationships between these entities. They can specify ids, names or keywords to be considered for each entity.

2.2 Querying Strategies

Interviews revealed that scientists are used to follow paths between sources to get information about the biological entities of interest for their question.

However, the scientists differ considerably in some aspects of querying, in particular whether or not they follow an order on the entities searched and they explore additional entities. In [3], we termed these querying criteria *Ordered* and *OnlyGivenEntities*, respectively.

[6] The questionnaire and survey results are available at BioGuideSRS web site.

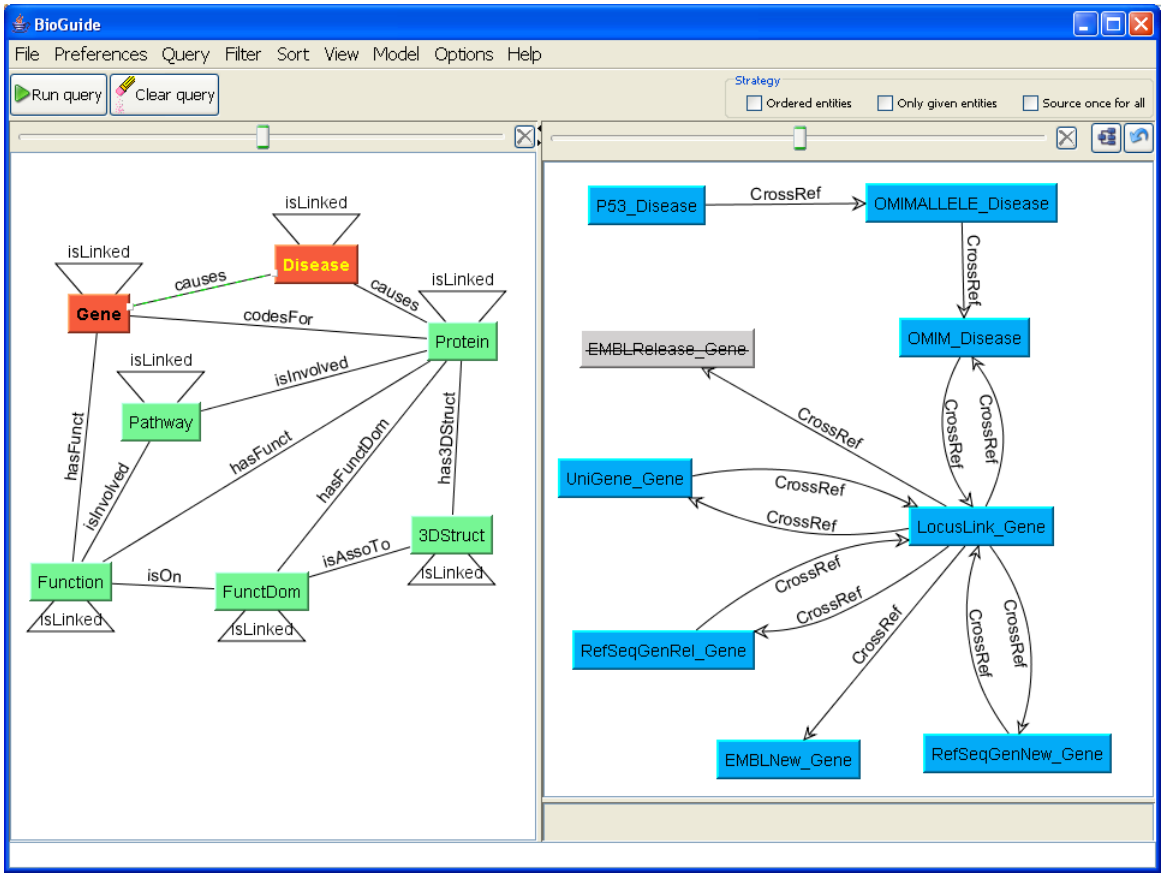

Figure 1. BioGuideSRS main interface. On the left hand side: the graph of entities where two entities have been selected by the user (DISEASE, GENE). The name of the searched disease has been indicated (consequently, DISEASE appears in brighter font). On the right hand side: a sub-part of the graph of sources-entities corresponding to the causes relationship between DISEASE and GENE. The user has specified she does not want the EMBLRelease source to be used (consequently, strikethrough formatting is applied to the name of this source). See section 4 for the complete query. Note that initial colors have been modified to make the graph easier to read when printed in black and white.

The first criterion, **Ordered**, distinguishes the situation where the scientist searches for a specific piece of information from the case where she rather wants to explore all the available data. For example, if the scientist searches for the function of a known protein, she may access a few sources containing the protein (e.g. SwissProt) and then follow some cross-references to sources providing function (e.g. Enzyme, Brenda). In such a situation, the scientist **orders** the entities of her query and thus follows only links from a given entity (e.g. PROTEIN) to another (e.g. FUNCTION). A completely different situation occurs if the protein of interest has not been already studied by others scientists. In that case, the biologist browses the sources by looking for as much information as possible. She thus not only consults some sources providing PROTEIN information and follows cross-references to sources providing FUNCTION as previously, but she can also consider the two entities in the inverse order (*i.e.* from FUNCTION to PROTEIN). Indeed, she may also search for information on FUNCTION in sources (e.g. GO) linked to sources (e.g. TrEMBL) containing information on her protein of interest. In such a case, the scientist has no precise idea of the order in which she is going to deal with entities and rather explores all the possible links between entities in sources. She thus considers all the orderings between entities.

The second criterion, **OnlyGivenEntities**, corresponds to the distinction between situations where the scientist is interested in finding information about some entities (and only these entities) from the case where she is also interested in exploring *additional* entities which are *semantically linked* to the ones she explicitly looks for. For example, if the scientist is interested in finding data on PROTEINS encoded by a given GENE, she may query sources providing information about these two entities. If she does not find information about her gene of interest, she may consult sources providing other entities, such as DISEASE, and search if some diseases are known to be related to this gene. She then tries to find a set of proteins associated to these diseases. This way of doing is thus *navigating* as far as *additional* entities (e.g. DISEASE) can be used to get information on a given set of entities (e.g. GENE and PROTEIN).

Interestingly, a variety of querying processes discovered through the interviews can be generated by combining these criteria. If the results obtained are not satisfactory, the scientists may then drop one of these criteria, e.g. allow the entities to be queried in any order. Section 4 shows how following strategies in BioGuideSRS allows the scientists to find complementary data.

We call the combination of criteria the querying **strategy** [3].

2.3 Mixed queries and Traceability

Answers to questionnaires have also revealed that users occasionally need to cite some sources (e.g. in question 2, OMIM is cited).

In BioGuideSRS they are supported in this task by a graphical representation of sources, offered by the **graph of sources-entities** (Fig. 1, right hand side), in which each node represents an entity in a source and arrows indicate the links between two entities in the same source or in another. Labels on arrows specify the kind of link: cross-reference (*CrossRef*) and internal link (*Internal*) that is, links between entities in the same source. For instance, the link $LocusLink\_Gene \overset{CrossRef}{\rightarrow} UniGene\_Gene$ means that the $LocusLink$ source provides cross-references from its genes to the genes in $UniGene$.

Using the graph of sources-entities, scientists can express *mixed queries* in which they specify the sources to access or to avoid.

Although some degree of transparency is often needed in queries, scientists also expect to be aware of the *provenance* of the answers. That is, they need to know which data sources and links have

been used to generate the answer to their question (this is called *why-provenance* in [1]). Traceability of results is crucial for verifying results, drawing conclusions, and testing biological hypotheses [14].

BioGuideSRS meets this need in two ways. First, BioGuideSRS allows the user to visualize the correspondence between the graph of entities and the graph of sources-entities. By selecting an entity, the user visualizes the sources which provide information about this entity; similarly, by selecting a relationship, the user visualizes the links between sources which achieve this relationship. Second, the data obtained as a result yielded by BioGuideSRS to the user is systematically associated with the path which has been used to obtain it. In this way, the user knows the exact sequence of sources and links used.

Section 4 will illustrate these two points.

3 A user-centric approach

3.1 Preferences values

From the answers to the questionnaire, the need of taking into account the *preferences* expressed by the biologists appeared as a necessity.

Responses to our questionnaire showed that the reason why a source or link is preferred varies between scientists. Interviews revealed that about 30 criteria determine preferences (e.g. reliability, completeness and ease of use). BioGuideSRS aims to help the users to quantify the confidence they have in entities in sources and links between them. Thus each component of a path is associated with some confidence level.

Initial confidence values for entities in sources and links are provided by using information from the SRS platform (e.g. about the number of instances each source contains for some given entity, the number of links between two sources etc.).

The user can then adjust the initial values by using the user-friendly interface (Fig. 2) provided by BioGuideSRS.

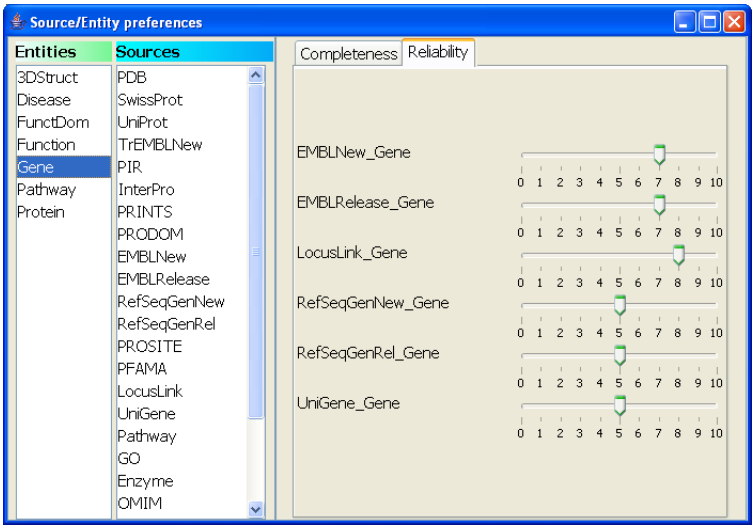

Figure 2. Initializing Preferences. Reliability values of the sources containing the Gene entity.

3.2 Filtering and Ordering the paths

To help the user in the process of selecting data, the values of preference criteria are used by BioGuideSRS to filter and order the paths it provides.

Firstly, we introduce the notion of **level of filter preference** and distinguish three different levels: (i) *global*, (ii) *intermediate* and (iii) *local*. The *global level* corresponds to a filter on a path, *i.e.* on the sequence of sources and links taken as a whole. For example, BioGuideSRS allows the user to select the paths shorter than (or longer than) a given length (the length being defined as the number of cross-references). It is also possible to search exclusively for data in complete sources (exploiting completeness preferences) etc.

Filters at the *intermediate level* focus on a given entity or relationship. For example, BioGuideSRS allows the user to specify that information about a given entity (e.g. Protein) must come from highly reliable sources.[7] At the *local level*, filters relate to a given source-entity or a given link, allowing the biologist to name the source/link to use or to avoid. For example, the biologist may specify that information about proteins have to be provided by the SwissProt source.

Secondly, we provide ways of **sorting** paths according to the biologist's preferences [4]. To do this, we associate a value with each path. The global value of a path is computed from the confidence assigned to its components (source-entities and links). The way it is computed, *i.e.* the *sort-operation* used (e.g. *Weighted sum*, *Best Source* operations), can vary. For example, in the *Weighted Sum* operation, the value of the path is the average of the confidence values of all of the nodes and arrows of the path whereas in the *Best Source* operation, the confidence value of the whole path is the value of the node having the highest confidence value. More information about these operations can be found on the BioGuideSRS Web site.

3.3 Adapt BioGuideSRS to your needs

We have seen that BioGuideSRS is very flexible and can be adapted to each user who can manage her preferences by creating new kinds of preferences and/or may adjust or rectify the preferences values provided by default. Also, the user can graphically manage the graph of entities and the graph of sources-entities: she can add/remove/modify links and nodes. Each user can save her own configuration though an XML file.

4 From paths to instances of data

BioGuide builds paths which are alternative ways of querying the sources. Paths can be interpreted as query plans [8] suitable for being used by any integration system. To get answers (instances of data) from a given biological question, paths have then to be rewritten into queries against the sources. We are currently investigating the use of BioGuide in different integration systems. In this paper, we introduce BioGuideSRS as such a solution. It is grounded on the Sequence Retrieval System (SRS, [5]). SRS was chosen because (i) it performs a uniform and fast access to more than 200 sources, (ii) its query language, Getz, allows to exploit cross-references, and (iii) it is daily used by many biologists as far as it is one of the main entry point to numerous sources at the European Bioinformatics Institute (e.g. EMBL, UniProt, Ensembl). BioGuideSRS has been sooner developed to offer a system accessible by any user having an Internet connexion.

[7] In this case, the sources used to provide information about entities different from Protein are not necessarily reliable.

4.1 BioGuideSRS Architecture

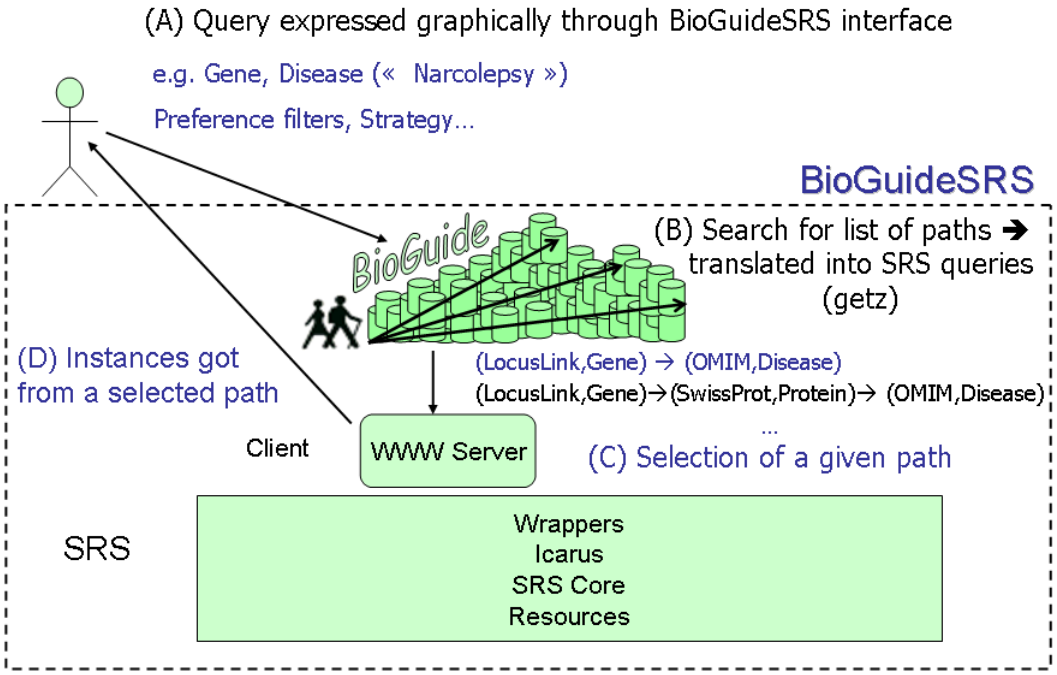

Figure 3. BioGuideSRS Architecture

Figure 3 presents the BioGuideSRS architecture. The BioGuideSRS querying process consists in four main steps. (A) First, the user asks for a query through the BioGuideSRS interface as described previously (*i.e.* composed of entities, preferences, strategies etc.). (B) Then, BioGuideSRS generates a list of paths between sources as alternative ways of obtaining data. BioGuideSRS builds for each of these paths a corresponding SRS query (in the Getz query language). (C) The user selects the path she would like to follow and (D) she automatically gets a set of corresponding instances of data obtained by the SRS Web server. Steps (C) and (D) can be done for each path yielded.

4.2 Example of queries and results

As an example, let us consider the following question: *Which information may I get about genetic sleep disorders such as narcolepsy and the genes related to these diseases?*[8]. In this question the user expresses the need for getting information about the correlation between diseases and sets of genes.

This question can be expressed by selecting two entities (GENE and DISEASE) through the BioGuideSRS interface and by specifying ”narcolepsy” as a keyword to be searched for the DISEASE entity.

Assume that the user exploits the rank-operation and the strategy provided by default (i.e. the *Weighted sum* operation, and the strategy considering every ordering between entities and allowing intermediate entities). Assume that the user specifies some preference-filters: (i) no more than 3 cross-references must be followed per path, (ii) only very reliable sources should be consulted, (iii) concerning the gene entity, only complete sources should be consulted and (iv) the EMBLRelease source (the full EMBL source) should not be considered.

[8] Complete information about this example can be found in BioGuideSRS Web site.

Let us notice that different levels of preferences have been used here: global level (cf. (i) and (ii)), intermediate level (cf. (iii)) and local level (cf. (iv)).

As a result, BioGuideSRS provides five paths as five alternative ways to get data.

- (1) LocusLink_Gene $\overset{CrossRef}{\rightarrow}$ OMIM_Disease
- (2) OMIM_Disease $\overset{CrossRef}{\rightarrow}$ LocusLink_Gene
- (3) LocusLink_Gene $\overset{CrossRef}{\rightarrow}$ SwissProt_Protein $\overset{CrossRef}{\rightarrow}$ OMIM_Disease
- (4) OMIMALLELE_Disease $\overset{CrossRef}{\rightarrow}$ OMIM_Disease $\overset{CrossRef}{\rightarrow}$ LocusLink_Gene
- (5) OMIM_Disease $\overset{CrossRef}{\rightarrow}$ LocusLink_Gene $\overset{CrossRef}{\rightarrow}$ EMBLNew_Gene

By selecting these paths, the user obtains instances of data (see Fig. 4). First, it is important to see that the user had neither to specify the sources to be queried nor to indicate the links to be followed. Alternative ways of finding data have been automatically provided by BioGuideSRS. Second, getting instances of data from SRS for each path has also been done automatically by our system: querying the SRS system is an automatic process from the beginning to the very end.

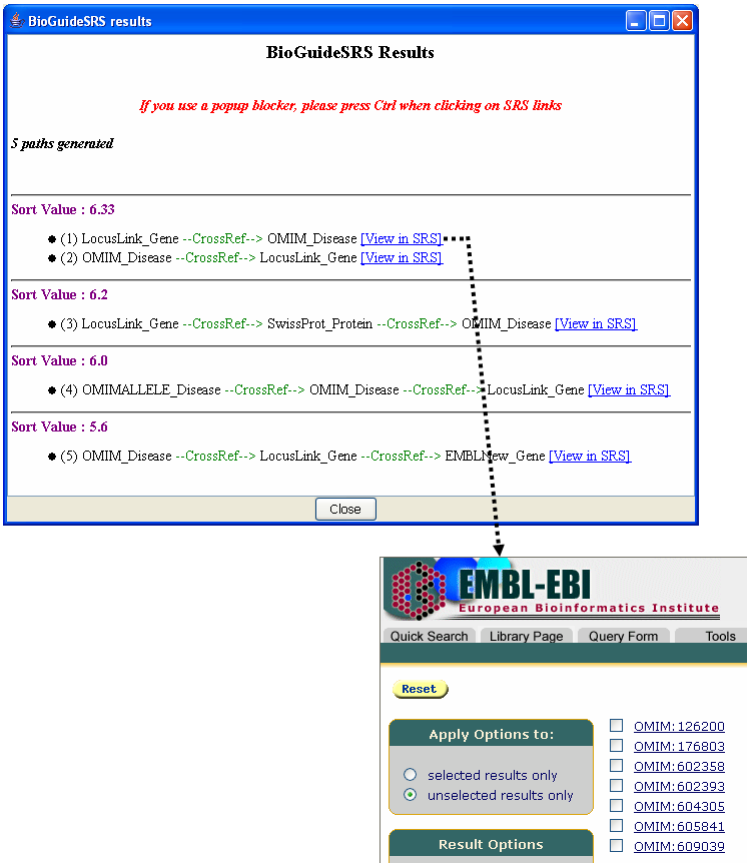

Figure 4. Example of BioGuideSRS results: list of paths (top) and EBI SRS screen with data answers from the first path (bottom).

Let us analyze the results obtained.

First of all, let us see that paths (1) and (3) provide information about disease from OMIM whereas paths (2), (4) and (5) provide information about genes, (2) and (4) from LocusLink and (5) from EMBLNew (which contains only fresh entries: the updates to the latest full release of EMBL).

On the one hand, path (1), which searches for information in LocusLink related to the narcolepsy disease in OMIM, provides 7 OMIM entries while path (3), which links genes to diseases by passing through the proteins of SwissProt, yields 8 OMIM entries. Four entries are common between paths (1) and (3) whose ids are 176803, 602358, 602393, 604305. These entries give information about clinical studies conducted with narcoleptic patients and information about another sleep disease named NREM sleep (NonRapid Eye Movement). However, some entries differ between paths (1) and (3). First, the entry 126200, which is the first one found in path (1), is not yielded by path (3). It interestingly describes a familial form of multiple sclerosis associated with narcolepsy. Second, the entry 161400 is found by path (3) but not by path (1). This entry, named Narcolepsy 1, is crucial to be known by the scientist as far as it gives the precise knowledge about the general form of the narcolepsy disease.

By passing through SwissProt more information linking genes and the narcolepsy disease is found. Thanks to BioGuideSRS, the user can thus exploit the added-value information given by annotators of sources.

On the other hand, path (4) provides a single LocusLink entry which is one of the 7 entries yielded by path (2). This single entry, whose LocusLink id is 3060, describes the HCRT gene which is well-known to be responsible for the narcolepsy disorder. Path (2) gives access to other LocusLink entries providing information about more recent narcolepsy loci which have been mapped to chromosome 4 (LocusLink id: 100918) and chromosome 21q (LocusLink id : 494446) and are very probably caused by the genes NRCLP3 and NRCLP, respectively. It also provides the entry 5730 giving information about the gene PTGDS but without indicating in the annotation any possible link with the narcolepsy disease. Finally, path (5) returns one EMBLNew entry, AK075333, which is annotated as associated to the PTGDS gene.

As a result, the user obtains information about several forms of the narcolepsy and for each form, the genes known to be involved in the disorder. The LocusLink entry 5730 does not give information about narcolepsy but its annotation mentions the PTGDS gene as well as the EMBLNew entry. The user thus knows that the gene PTGDS may be correlated with sleep disorders, probably different from narcolepsy. The next step of the user study may be to explore the links between the gene PTGDS and other sleep diseases such as the NREM sleep. Thanks to BioGuideSRS the user has new hints for her research.

Obviously, BioGuideSRS allows the user to make the most of the data available in the sources by automatically exploring alternative and complementary ways of obtaining data.

5 Conclusion

The study of transparent queries answered by providing paths of cross-references has increased among the past few years (e.g see [10,7,9,6]). The BioMediator [10] project was the first one in this field and has recently provided access to instances of data. This project focuses on an XML mediator approach and the current query language is XQuery. BioMediator considers some NCBI sources for which the programmers are maintaining wrappers. BioMediator is thus dedicated to users who know the XQuery language and is not willing to be used by external research groups.

Unlike BioMediator, BioGuideSRS aims to provide a graphical query language and to be used by anyone who has a Web access (BioGuideSRS is a Java Applet). Moreover a formal (internal) query language has been provided for BioGuide and studied in [2].

In other related works such as [7,9,6], no automatic access to instances of data is provided.

We summarize below the key advantages of using BioGuideSRS:

- Queries can be expressed in terms of biological entities through the BioGuideSRS user-friendly interface in a transparent way: the user is freed from having to find the sources and answers are automatically searched;
- *Filters* on the kind of sources to be accessed can be easily specified, reflecting user's *preferences*;
- *Links* between the sources are systematically followed according to the own strategy of the user in order to explore *alternative* and *complementary* ways of finding data;
- An intermediate level between queries and data is offered: to each path corresponds a given set of data thus (i) the user always knows the *origin* of the data got (sources and links followed), (ii) paths can be explored one after the other following their order in the list (corresponding to their order of *preference*).

We have shown that BioGuideSRS is very flexible and can be adapted to each user's needs. BioGuideSRS user interface is progressive because it moves complex and less frequently used options out of the main user interface, into secondary screens. In this way, BioGuideSRS allows non-experienced users to exploit default values while permitting experienced-users to customize the system according to their needs (adjusting preferences values, managing graphs etc.).

BioGuideSRS is currently used by members of the Children hospital of Philadelphia. In this context we are working on some evolutions of the system. For example, BioGuideSRS does not exploit the full expressive power of BioGuide: especially, it does not consider the use of tools (such as Blast) while BioGuide was designed for it. This choice was made because the current Getz query language does not support the call for tools in a straightforward way. BioGuideSRS will include the ability to use bioinformatics tools as far as the next version of SRS will support web-service calls [13]. Also, the current version of BioGuideSRS uses the SRS Web interface to present the answers to the user; we are working on a concise way to display these answers in order to allow the user to quickly understand the obtained results.

More generally, we are investigating the use of BioGuide on top of other kinds of integration approach. Particularly, we are working on using BioGuide on top of huge warehouses to guide the scientists for finding relevant data.

Acknowledgements

Authors thank Frédérique Lisacek (SIB, Swiss Institute of Bioinformatics), Philippe Bessières (MIG, Mathématique, Informatique et Genome, INRA, Jouy-en-Josas) for their valuable comments on BioGuideSRS as well as Susan Davidson (Database group, UPenn), Hijun Qui, Long Qu, and Pete White (Children hospital of Philadelphia), Shailesh Date, Kobby Essien, Jonhathan Schug, Howard Bilofsky, and Chris Stoeckert (Computational Biology and Informatics Laboratory, UPenn), for fruitful discussions they had with them about BioGuideSRS.

References

[1] P. Buneman, S. Khanna, W. Tan, Why and Where: A Characterization of Data Provenance. *Proceedings of the International Conference on Database Theory (ICDT)*, pp. 316-330, 2001.

[2] S. Cohen-Boulakia, C. Froidevaux and E. Pietriga, Selecting Biological Data Sources and Tools with XPR, a Path Language for RDF. *In Biocomputing 2006, Proceedings of the Pacific Symposium on Biocomputing (PSB)*, pp. 116-127, 2006.

[3] S. Cohen-Boulakia, S. Davidson and C. Froidevaux, A User-centric Framework for Accessing Biological Sources and Tools. *Proceedings of the International Workshop on Data Integration in the Life Sciences (DILS)*, Springer-Verlag, Lecture Notes in Computer Science (LNCS) series, pp. 3-18, 2005.

[4] S. Cohen-Boulakia, S. Lair, N. Stransky, S. Graziani, F. Radvanyi, E. Barillot, C. Froidevaux, Selecting biomedical data sources according to user preferences. *Bioinformatics*, 20:i86-i93, 2004.

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

[6] S. Heymann, F. Naumann, L. Raschid and P. Rieger, Labeling and Enhancing Life Sciences Links. *Proceedings of the Computational Systems Bioinformatics Conference (CSB2006)*, pp. 598-599, 2004.

[7] A. E. Lash, W.-J. Lee and L. Raschid, A Methodology to Enhance the Semantics of Links between PubMed Publications and Markers in the Human Genome. *Proceedings of the International Symposium on BioInformatics and BioEngineering (BIBE)*, pp. 185-192, 2005.

[8] A.Y. Levy, Combining Artificial Intelligent and Databases for Data Integration. *Artificial Intelligence Today*: 249-268, 1999.

[9] G. A. Mihaila, F. Naumann, L. Raschid and M.-E. Vidal, A Data Model and Query Language to Explore Enhanced Links and Paths in Life Science Sources. *Proceedings of the Workshop on the Web and databases (WebDB)*, pp. 133-138, 2004.

[10] P. Mork, A. Halevy, P. Tarczy-Hornoch, A model for data integration systems of biomedical data applied to online genetic databases. *Proceedings of the AMIA Symposium, American Medical Informatics Association*, pp. 473-477, 2001.

[11] H. Muller, F. Naumann, Data Quality in Genome Databases. *Proceedings of the International Conference on Information Quality*, 269-284, 2003.

[12] F. Naumann, U. Leser, J.C. Freytag, Quality-driven Integration of Heterogenous Information Systems. *Proceedings of the International Conference on Very Large DataBases (VLDB)*, pp. 447-458, 1999.

[13] D. Staines, Web services-based access to SRS. *Proceedings of the Tools and Applications in Biology workshop (NETTAB)*, electronic proceedings, 2005.

[14] J. Zhao, C. Wroe, C. Goble, R. Stevens, D. Quan and M. Greenwood, Using Semantic Web Technologies for Representing e-Science Provenance. *Proceedings of the International Semantic Web Conference (ISWC)*, pp. 92-106, 2004.