Open access • Journal Article • DOI:10.1142/S0219720006002375

Path-based systems to guide scientists in the maze of biological data sources

Sarah Cohen-Boulakia, Sarah Cohen-Boulakia, Susan B. Davidson, Christine Froidevaux ...+3 more authors

Institutions: University of Paris-Sud, University of Pennsylvania, University of Maryland, College Park, Arizona State University ...+1 more institutions

Published on: 01 Oct 2006 - Journal of Bioinformatics and Computational Biology (Imperial College Press)

Topics: Biological data

August 2006

Path-based systems to guide scientists in the maze of biological data sources

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

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

Christine Froidevaux
Université Paris-Sud

Zoe Lacroix
Arizona State University/University of Maryland

Maria-Esther Vidal
Universidad Simon Bolivar/University of Maryland

Electronic version of an article published as *Journal of Bioinformatics and Computational Biology*, Volume 4, Issue 5 (October 2006), pp. 1069-1095.
Article DOI: http://dx.doi.org/10.1142/S0219720006002375 © copyright World Scientific Publishing Company
Journal URL: http://www.worldscinet.com/jbcb/jbcb.shtml

Path-based systems to guide scientists in the maze of biological data sources

Abstract

Fueled by novel technologies capable of producing massive amounts of data for a single experiment, scientists are faced with an explosion of information which must be rapidly analyzed and combined with other data to form hypotheses and create knowledge. Today, numerous biological questions can be answered without entering a wet lab. Scientific protocols designed to answer these questions can be run entirely on a computer.

Biological resources are often complementary, focused on different objects and reflecting various experts' points of view. Exploiting the richness and diversity of these resources is crucial for scientists. However, with the increase of resources, scientists have to face the problem of selecting sources and tools when interpreting their data.

In this paper, we analyze the way in which biologists express and implement scientific protocols, and we identify the requirements for a system which can guide scientists in constructing protocols to answer new biological questions. We present two such systems, BioNavigation and BioGuide dedicated to help scientists select resources by following suitable paths within the growing network of interconnected biological resources.

Keywords

bioinformatics, querying biological resources, paths between sources, metadata and user's preferences

Comments

Electronic version of an article published as *Journal of Bioinformatics and Computational Biology*, Volume 4, Issue 5 (October 2006), pp. 1069-1095.
Article DOI: http://dx.doi.org/10.1142/S0219720006002375

Journal URL: http://www.worldscinet.com/jbcb/jbcb.shtml

Path-based systems to guide scientists in the maze of biological data sources

Sarah Cohen-Boulakia[1,2], Susan Davidson[1]

[1] *Department of Computer and Information Science, University of Pennsylvania, 3330 Walnut St, PA-19104 Philadelphia, USA*

sarahcb@seas.upenn.edu, susan@seas.upenn.edu

Christine Froidevaux[2]

[2] *Laboratoire de Recherche en Informatique, CNRS UMR 8623, Université Paris-Sud 11, 91405 Orsay, France*

cohen@lri.fr, chris@lri.fr

Zoé Lacroix[3,4]

[3] *Arizona State University, Tempe AZ, USA*
[4] *University of Maryland, College Park MD, USA*

zoe.lacroix@asu.edu

Maria-Esther Vidal[5,4]

[5] *Universidad Simón Bolívar, Caracas, Venezuela*

mvidal@umiacs.umd.edu

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Fueled by novel technologies capable of producing massive amounts of data for a single experiment, scientists are faced with an explosion of information which must be rapidly analyzed and combined with other data to form hypotheses and create knowledge. Today, numerous biological questions can be answered without entering a wet lab. Scientific protocols designed to answer these questions can be run entirely on a computer.

Biological resources are often complementary, focused on different objects and reflecting various experts' points of view. Exploiting the richness and diversity of these resources is crucial for scientists. However, with the increase of resources, scientists have to face the problem of selecting sources and tools when interpreting their data.

In this paper, we analyze the way in which biologists express and implement scientific protocols, and we identify the requirements for a system which can guide scientists in constructing protocols to answer new biological questions. We present two such systems, BioNavigation and BioGuide dedicated to help scientists select resources by following suitable paths within the growing network of interconnected biological resources.

Keywords: Querying biological resources; Paths between sources; Metadata and user's preferences.

1. Introduction

Due to the explosion of high-throughput scientific data that is available over the web, an increasing number of biological questions can be answered without entering a wet lab. That is, the scientific protocols designed to answer these questions are *digital* and can be run entirely on a computer. Each step of a digital scientific protocol is a *bioinformatics task*, and is defined by a description which captures its scientific aim and an implementation which specifies the resources selected to execute the task[44,1]. Thus the same task description may admit multiple implementations.

Although many scientists currently use scripting languages such as Perl to express and execute digital scientific protocols, there is an increasing recognition of the need for workflow systems to manage them, i.e. systems which allow the expression of protocols, can invoke protocol implementations, record results, and be used to query results as well as the reasoning that produced those results; examples include myGrid/Taverna[37], Kepler[3], Chimera[16], DiscoveryNet[40], MHOLline[46], HKIS-Amadea[12], and AdaptFlow[20] (see the web page survey[42] for other examples of workflow systems).

These systems are very good at describing implementations of scientific protocols and managing the resulting digital results. However, they do not help scientists with the overwhelming task of selecting an implementation for a given scientific protocol.

As an illustration, Figure 1 shows an example of a protocol called the *bacterial artificial chromosomes (BAC) augmentation protocol*, which has been designed in the context of the HKIS project.[a12] The steps of the protocol are indicated with boxes, and the data flow by ovals and arrows. Within comparative genomic hybridization (CGH) array experiments, BACs are used to detect gains and losses in the DNA of tumoral samples, thus allowing the identification of new cancer-related genes. Thus, loosely speaking, this protocol answers the question *"What genes in my CGH array can be related to cancer?"*. Each step of a protocol corresponds to an intermediate question; for example, the question associated to the first step (Position BAC) is *"Where are all the BACs of my CGH array located on the genome sequence?"* Some of these steps require importing data from external data sources. Note that external data is required in two of the bioinformatics tasks in the sample protocol.

Public biological resources form a complex maze of heterogeneous data sources, interconnected with links and applications. This valuable network offers scientists potential answers to a wide variety of scientific questions. However selecting the appropriate resources for obtaining the data of interest is a tedious task: While scientific questions are posed at a conceptual level, their implementation entails determining which data resources and tools to use, that is, which *paths* in the network of sources to follow.

[a]For more information about the HKIS project and its Web-based platform see http://isoft.free.fr/hkis/.

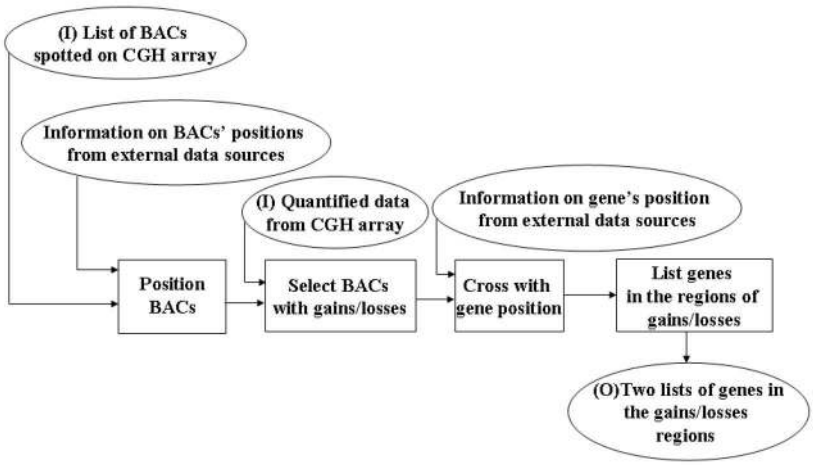

Fig. 1. BAC augmentation protocol

For example, the first step of the BAC augmentation protocol (*Position BAC*) requires the scientist to know which data resources have information about the positions of BACs. Frequently, this choice is based on what resources the scientist is most familiar with, rather than what resources are best suited to the question.

The number of resources a scientist uses regularly is very small compared to the number of resources that are available on the Web; scientists tend to access a core of well-known resources (e.g., GenBank, Swiss-Prot) to which they add a few specialized resources depending on their specific needs. A survey conducted in 2003[22] showed that 29 scientists belonging to three different research teams used on the average six public resources (including databases and applications) of a total of 57 cited distinct resources. In contrast, there are 719 public data sources listed in the Molecular Biology Database Collection[18], and this number is increasing exponentially: since 2004 there has been a 31% increase, and a total increase of 351% since the list was compiled in 1999[5].

Unfortunately, it is unreasonable to expect scientists to use more resources due to the complexity of evaluating and mastering them, combined with the speed at which they evolve. In evaluating the usefulness of a resource for a biological question, the scientist must understand a number of things, including the content of the resource; the state of its curation; the format of its entries; the capabilities made available to access, analyze, and display the data; and its relative position within the network of interconnected public biological resources.

The goal of this paper is to analyze the way in which scientists currently construct their protocols, and identify the requirements for a system which can guide them in constructing new protocols to answer new biological questions. In particular, the system must help scientists select resources and find paths between resources within the growing network of interconnected biological resources.

The paper is organized as follows: Section 2 motivates the need for alternative resource selections. Challenges for designing scientific protocols, and user requirements collected from a questionnaire are discussed in Section 3. Section 4 focuses on features which should be satisfied by a guidance system. Path-based guiding systems are then introduced and compared in Section 5. Section 6 discusses additional challenges in designing workflow systems to manage digital scientific protocols.

2. Alternative resource selection

The network formed by biological resources is diverse, and offers multiple orthogonal viewpoints on scientific data. Each viewpoint is expressed by the way in which the data are organized (e.g., GenBank is sequence-centric while GeneCards is gene-centric), the access capabilities offered to scientists (e.g., to access gene descriptions in GeneCards, one can use a full-text search engine or provide a HUGO symbol), as well as the applications, annotations, and links to other relevant resources. In addition to this structural diversity, biological resources offer a rich semantic diversity. This semantic diversity may be characterized by the number of entries present in the data source, the number of attributes pertaining to each entry, the number of links between entries, as well as the quality, consistency, and reliability of these resources. All these semantic characteristics offer metrics that may be used in determining whether or not to use the resource in an implementation of a scientific protocol, and dramatically affect the collected dataset instance[28]. In addition, combining several similar resources may provide complementary pieces of information, thus generating a more complete dataset.

An experiment conducted in February 2005[23] demonstrates how the selection of resources may result in dramatically different datasets. Consider the simple query *"Retrieve bibliographic references related to [a particular genomic disorder]."* To execute this query, a selection of resources could include OMIM[b] and the PubMed Links provided by NCBI to retrieve PubMed entries related to each of the OMIM entries. These links offer a valuable contribution to the scientists as curated precomputed joins between heterogeneous data sources. For the disease *diabetes*, 48,941 entries (without duplicates) were retrieved from PubMed that were linked to OMIM entries retrieved by conducting a search with 17 keywords related to diabetes. Alternative resource selections include using the same data sources (OMIM and PubMed) but using different links; using alternative data sources (e.g., an alternate resource for OMIM is GeneDis[c]); or using alternative paths that may include additional resources. An alternative link between OMIM and PubMed may be found by parsing the retrieved OMIM entries and extracting all PubMed references. This alternate execution of the same query retrieved 50,843 PubMed entries from the same set of OMIM entries, that is 1,902 more entries than the previous selection. Therefore, a

[b]OMIM is available at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM.
[c]GeneDis is available at http://life2.tau.ac.il/GeneDis.

scientist whose goal is to retrieve as many entries as possible while accessing the fewest resources should favor the latter option. However, a scientist wishing to retrieve all possible entries should collect entries from both options because the first option retrieved 21 entries not included in the second option; whereas a scientist wishing to retrieve entries with the most evidence might take the intersection of entries retrieved in the first and second options. This example illustrates that the selection and use of resources is critical to the quality and completeness of the retrieved dataset, and is linked to the goal of the scientist who is posing the question. Section 3.2.4 explains how a semantic description of resources should be used to define the approach used for biological resource selection.

As another example of why alternative resources should be considered, recall the first step of our sample workflow query introduced in Section 1 which asked the question *"Where are all the BACs of this CGH array located on the genome sequence?"*. Alternative selections of resources give complementary results concerning single instances of BACs. For example, accessing only MapViewFish[d] locates BAC RP11-89F21 on a particular chromosome band, whereas accessing UCSCGenome[e] gives the exact position of this BAC on the chromosome sequence[10].

However, alternative resource selections may lead to conflicting results. For the BAC CTD-2012D15, GenBank and MapView locate it on chromosome X, while UCSCGenome and MapViewFish locate it on chromosome 11. In this situation, the selection of a single path would lead to a result which is potentially inaccurate. However, if the user selects both paths, the system must resolve the conflict in chromosomal location for this query: X or 11. One solution would be to exploit the user's evaluation of the reliability of the resources involved to resolve the conflict; another would be to report both locations and give the provenance of the answer, e.g. the data source providing the answer.

For these reasons, it is critical to explore alternative selections of biological resources to execute scientific protocols and use various paths to obtain complementary pieces of information. In the next section, we give an overview of a study of how users go about selecting resources to use in implementing their protocols, and explore their motivations. The analysis of these use cases lays a foundation for how guidance systems should support the selection of scientific resources to execute protocols.

3. User requirements

There are many challenges that bioinformaticians face while supporting scientists in creating digital scientific protocols. First, the statement of a protocol is frequently abstract, e.g. *"Retrieve bibliographic references related to [a genomic disorder]."*, admitting many different implementations each of which consults different data

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

sources, links, and tool-links. Second, the choice of implementation is subjective and varies from scientist to scientist. The bioinformatician must therefore know what information to give to the scientist as well what subjective information to elicit from the scientist in order to help formulate an appropriate implementation. In particular the bioinformatician must locate relevant resources and extract information about these resources, such as what entities they contain and associated metadata.

In this section, we start by describing a case study that was performed to determine how scientists formulate their protocols. We then discuss the need for transparent and mixed queries, and the kind of information that must be obtained in order to implement those queries.

3.1. *Collecting user requirements*

Over the past few years, several surveys have been performed to study how scientists search for information within data sources[44,45,13]. More recently, in the context of the BioGuide project[10], a study was done to determine the reasons why biologists select one source to query rather than another, and to identify the ways in which biologists collect information. This study was performed by developing a questionnaire and performing interviews, and involved 20 researchers whose interests included disease studies, functional genomics and structural genomics.

The questionnaire was developed using standard guidelines[21,17,38], and built on previously recognized considerations such as user preferences[44,45,13] and data quality[35,33,36]. The questionnaire consisted of 28 questions organized around three main topics: (A) queries and sources accessed, (B) ways of accessing sources, and (C) bioinformatics tools used. A few of the questions are provided below as an illustration:

- (Part A) Select a context from your own area of study and list some biological questions that you frequently ask. If several sources yield answers for your question, do you access them all or do you select only a few?
- (Part B) When you are collecting data related to two linked entities (e.g., a gene and the protein it encodes), how do you proceed (e.g., do you follow a particular order of entities)?
- (Part C) How many bioinformatics tools (e.g., Blast, Fasta) do you use when you want to perform a given bioinformatics task (e.g., similarity search)? How often do you change the parameters of a given tool?

Interviews were then conducted using classical techniques[15,39]. More precisely, two pools of individuals were considered. The first pool of five individuals were asked to fill in the questionnaire and were then interviewed. For this pool, we used the *debriefing* technique[39] in which answers are analyzed with the interviewee. We used this technique to gain a better understanding of how respondents interpreted the questions asked of them. We were thus able to identify words, terms or concepts that respondents did not interpret as we intended, and to obtain suggestions for revising

the questionnaire. By using this technique, we not only improved our questionnaire but also collected clear answers. The second pool consisted of fifteen individuals who were interviewed using the *incident critics* technique[15]: we asked scientists to describe their area of study, and show us what they do when they want to find information in their data sources of interest. In this way, we obtained answers to the questionnaire by having a discussion with the scientist.

As a result of this study, a set of 156 scientific questions expressed in natural language were collected, as well as a set of sources, tools and preference criteria. The collected questions are analyzed in the next subsection.[f]

3.2. *Analyzing responses to survey*

The scientific questions that were collected can be classified as *transparent queries* and *mixed queries*, which are respectively explained in Sections 3.2.1 and 3.2.2. In addition to the wording of the scientific question itself, scientists expressed preferences about the resources used to implement a protocol and how the protocol was to be executed, i.e. the *approach* followed. The preferences expressed by the scientists are developed in Section 3.2.3 while Section 3.2.4 introduces the different kinds of approaches which can be followed to implement a given protocol. The results we present hereafter were obtained by working in close collaboration with biologists who systematically validated our results.

3.2.1. *Transparent queries*

Our analysis of the collected scientific questions revealed that, in many cases, neither the data sources nor the tools to be used were specified by the biologists. That is, scientists designed their questions at a conceptual level[24] referring to underlying biological *entities* and the *relationships* between these entities. Examples of such questions are:

(1) *Return all contigs that map "close" to the marker M on chromosome 19.*
(2) *Where are all the BACs of this CGH array located on the genome sequence?*
(3) *What gene(s) result(s) in the Long QT syndrome disease and codes for an inactive protein?*
(4) *From this set of structural motifs, which are the corresponding proteins having the same 3Dstructure (using the FSSB source) and their ORFs?*

The biologists we worked with were then asked to identify the underlying entities and relationships present in their questions. As an example, the term "ORF" in question (4) was associated with the entity "Gene", and in question (3) the fact that a set of genes may "result in" a given disease was associated with a "causes"

[f]The questionnaire and survey results are available at BioGuide web site: http://bioguide-project.net.

relationship. Occasionally, we also asked biologists to associate entities and relationships with terms in questions listed by other scientists.

More generally, the questions above refer to the following biological entities: (1) CONTIG, MARKER, and CHROMOSOME, (2) CHROMOSOME and BAC, (3) GENE, DISEASE, and PROTEIN (4) 3D-DOM-STRUCTURE, PROTEIN, 3D-STRUCTURE, and GENE

They also refer to the following relationships (1) "mapsWith" and "isOn", (2) "isOn", (3) "causes" and "codesFor", (4) "hasStruct", and "hasSimilarStruct".

Discussions with the biologists revealed that they consider different types of relationships between entities. Some relationships are precalculated, that is, explicitly represented and stored in the sources consulted; others are implicit and must be calculated on-the-fly using some bioinformatics tool. For example, in query (1) the fact that *"Marker M is on chromosome 19"* may be stored in the data sources queried by the biologist. In contrast, the relationship "close mapping" can be calculated (e.g., using *Blastn*).

The fact that many of the scientific questions were abstract points to the need to be able to express scientific protocols as *transparent queries*[31] which refer to scientific entities and relationships rather than physical data sources and bioinformatic tools.

3.2.2. *Mixed queries*

Although many of the questions collected were transparent, others evoked specific biological resources. We refer to these questions as *mixed* queries. For example, the question *"Retrieve proteins involved in breast cancer from OMIM"* specifies OMIM as the data source to be consulted to retrieve information about the disease "breast cancer", although no source is specified to retrieve the proteins involved (such a data source could be Swiss-Prot, TrEMBL, PIR, etc.).

Biological resources evoked in a scientific question may be data sources or links. *Data sources* are repositories of entries which are instances of scientific entities; for example, OMIM is a repository of instances of the DISEASE entity. *Links* connect different entries, possibly from different data sources. They express the relationships between scientific entities and include cross-references (or internal links), and tool-links.[g]. *Cross-references* are hypertext links (hyperlinks) from an entry in one source to complementary information (another entry or a set of entries) in another source. Cross-references are not necessarily symmetric; for example, although numerous specialized sources cross-reference GenBank, these resources are not referenced by NCBI in return. *Internal links* are used to join entries within a data source, and correspond to foreign keys in relational databases. Finally *tool-links* are (Web) services provided by a source to link entries from different sources. Each source may provide several different services achieving a given relationship between

[g]A cross-reference or a tool-link is also referred to as a *capability*[28]

entities. For example, ENTREZ provides various tool-links across the NCBI data sources. Tool-links such as PubMed Links, Nucleotide Links, Protein Links, etc. are implemented as indices, whereas others (such as "sequence similarity search") are implemented by tools (such as *MegaBlast* or *BlastN*).

Although some degree of transparency is often needed in queries, scientists also expect to be aware of the *provenance* of the answers[19,50]. That is, they need to know which data sources and tools have been used to generate the answer to their questions (this is called *why-provenance*[4]). Traceability of results has been identified as a requirement for systems which execute scientific protocols[37,32] and is crucial for verifying results, drawing conclusions, and testing biological hypotheses[50].

3.2.3. *Preferences*

Answers to the questionnaires also revealed that biologists express preferences about biological resources to be used in the implementation of a protocol. These preferences may be subjective, or based on measurable qualities captured in the metadata associated with the data sources and tools.

Subjective preferences
Many preferences are subjective, and are related to the management or curation of the data source, the ease-of-use of the data source, the completeness of the entries in the data source with respect to the question being asked, the the richness of information given about the entity of interest, and the quality of cross-references from entries in the data source. The confidence a user has in a tool-link may depend on its ease-of-use, the reputation of the tool, the completeness of answers generated by the tool, as well as what source is providing it. For example, a user may consider a Blast tool as reliable in general but assign to it a different reliability level depending on the source which provides it (e.g., *BlastN* from NCBI vs. *BlastN* from Expasy). Although preferences values should be given by each user, default values established by other scientists or by means of metadata (see below) are often very useful, especially for less experienced users.

Metadata-driven preferences
Other preferences are based on quantitative measures that are captured as metadata associated with a data source or tool. In contrast to subjective preferences, metadata does not vary from user to user. Examples of metadata include data source cardinality; link cardinality, i.e. the number of pairs of entries for each type of link; the number of entries in the data source with at most one outgoing link; the number of entries in the data source with at least one incoming link; how many attributes for an entity are contributed by the source; and measures of the curation of the data source and links. Note that the measures of curation may be somewhat subjective, but not at the level of the user.

3.2.4. *Approaches*

When scientists search for answers to a given biological question, they usually follow paths of links between sources as implementations of their question. For example, to answer the question *"Retrieve the nucleotide sequences, proteins, and references to published articles related to [a genetic disorder]"*, one implementation could be:

(1) Retrieve all OMIM entries related to the generic disorder (with a list of keywords characterizing the disorder).
(2) Follow all links from the entries collected from OMIM to the corresponding sequences in NCBI Nucleotide.
(3) Follow all links from the entries retrieved from NCBI Nucleotide to entries in NCBI Protein.
(4) Follow all links from the entries retrieved from NCBI Protein to entries in PubMed.

The implementation of the question is a path over the biological resources OMIM → NCBI Nucleotide → NCBI Protein → PubMed. The links not specified in the query may be the NCBI links (indices computed and provided by NCBI via the *EUtils*), or a local parser that extracts from each entry the needed identifiers from the retrieved entries combined with a call to the resource to retrieve the linked entries (e.g., to retrieve the PubMed entries linked to the protein entries, one can either use the PubMed Links or extract the PubMed ids from the protein entries and retrieve the corresponding PubMed entries).

Scientists may therefore wish to characterize the set of paths used in the implementation of their question, that is, to specify the *approach* to be followed. There are several ways in which this might be done. First, approaches may exploit preferences; for example, by considering the set of paths that find the greatest number of entries, entries that are supported by the largest number of paths, paths that maximize the information about each entry, or paths that take the shortest time possible. Second, approaches may exploit some characteristics of the paths, for example, whether a data source can or should be visited more than once, or the order in which data sources are visited.

4. Features of path-based systems

In this section, we draw on the user requirements presented in the previous section – the need for transparency in queries, subjective as well as metadata-driven preferences, and expression of approach – and discuss how they impact the design of a system that assists scientists in constructing digital protocols.

4.1. *Graphical support*

Scientific questions are frequently expressed as paths between scientific entities whereas the implementation of the question involves specifying the physical re-

sources involved. For these two reasons, a system that guides scientists in constructing digital protocols must provide two layers of representation.

Graph-based representation of biological entities

In order to allow scientists to formulate meaningful (and transparent) queries, a path-system should provide a logical level to represent the *biological knowledge* involving scientific entities (e.g., GENE, PROTEIN) and relationships between them (e.g., a GENE *codes for* a PROTEIN, a GENE *maps with* a PROTEIN). The most natural way to represent biological knowledge is to use a *conceptual graph* in which each node represents a biological entity (or scientific class) considered at a conceptual level. The edges connecting these nodes represent biological relationships between the corresponding entities. As two scientific entities may be linked by multiple relationships, each expressing a scientifically meaningful property, the graph should allow multiple labeled edges between two nodes.

Graph-based representation of resources

In order to allow scientists to be aware of the relevant resources to be selected for answering their queries, a path system should provide a logical level to represent the network of resources.

As seen in Section 3.2.2, data sources (or physical sources) are connected through different kinds of links: *internal links*, *cross-references* and *tool-links*. These links express effective ways to navigate from one data source to another. A straightforward way to represent biological resources is a graph in which each node represents a data source (e.g., OMIM) and each edge between two nodes represents a link (e.g., Entrez Nucleotide links). To adequately represent biological resources, edges of the *physical graph* should be directed and labeled, and there may be multiple edges between two nodes. As seen previously, in most cases links between two biological data sources are not symmetric, and multiple tool-links or cross-references may be available between two physical data sources because different physical links may carry significantly different scientific meanings or possess different properties.

Mapping between graphs

A mapping between the conceptual graph and the physical graph is also needed. Each node of the conceptual graph (e.g., GENE) is mapped to the data sources that provide information about the scientific entity (e.g., OMIM, NCBI Gene, GeneCard). Similarly, each relationship between two concepts (e.g., *is published in* between the concept GENE and the concept PUBLICATION) is mapped to the physical links that implement it (e.g., *NCBI PubMed Links*). For each calculated relationship between entities, the system should determine which tools can be used by the scientist to calculate the relationship (e.g., the tool *BlastN* calculates the

relationship *mapsWith*).

Since the conceptual graph captures knowledge about a domain (biological entities and their relationships) and creates a shared understanding of the domain that can be used by both humans and computers, it may be considered an *ontology*[43]. Ontologies of resources may then be considered based on the mapping between the physical and conceptual graphs and/or based on the metadata and preferences associated to the resources.

Designing graphs and mappings should be performed in close collaboration with biologists. Entity and relationship names should also be carefully chosen to reflect the scientists' knowledge of the domain.

A good path system should also provide a flexible graphical interface in which it is easy to support updates (modification, insertion or deletion) to both the conceptual and physical graphs. This is especially important as the network of biological resources is rapidly changing: schemas of data sources evolve, and new resources become available. Similarly, the mapping between the two graphs must be easily modifiable in response to change.

4.2. *Browsing and querying*

A guiding system should provide scientists transparency combined with an active control of the selection of resources. To allow informed decisions, a guiding system should provide the ability to access and navigate through both layers of representation and their mapping. A *browsing mode* allows users to navigate through the conceptual and physical graphs and access the information known about the resources and used by the system to select resources. With the *querying mode*, scientists express their scientific protocols and the guiding system returns a selection of resources that may be used for their implementation. A user-friendly interface should exploit graphical interactions for both modes.

4.2.1. *Specifying the approach*

As each transparent scientific query may result in multiple path and resource selections, users must also be able to influence the implementation of the query by specifying the approaches to be followed. In particular this can be done by exploiting the preferences they have in the resources. More precisely, because the set of all possible paths that match a transparent query can be extremely large, the system must provide a *filtering* mechanism for selecting path and resource selections. There are three categories of filters: (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. Examples of such filters include constraints such as the maximum path length, the calculated reliability of a path, or the size of expected output (e.g., the number of returned entries expected at execution)[12]. Filters at the *intermediate level* focus on a given entity or relationship. For example, such filters could be used

to build paths where all the sources provide reliable information about proteins, or where the tool-links for calculating similarity are easy to use. At the *local level*, filters relate to a given source or link, allowing the biologist to name the resource used.

A path-based system should *rank* implementations with respect to the scientists' expectations, allowing them to choose how to schedule their execution. The feature vector used to compute the ranking should take into account the data, cross-reference and tool-based resources used in the path as well as path characteristics. Ranks can be calculated by weighting the feature vector according to the scientist's expectations.

4.2.2. *Evaluating and reusing queries*

Having decided on a set of paths to implement the scientist's query, the system must then implement it over the physical resources. In general, this will entail using an integration environment capable of running queries or pipelines over multiple, heterogeneous data and tool resources.

Efficiency is also a concern as the number of paths connecting a set of nodes may be extremely large if any ordering of nodes is allowed. However, it is worth pointing out that queries generally evoke only a small number of entities at the same time (only 8 % of the queries of the set of collected scientific questions had more than three entities) and the length of paths expressed by scientists' queries rarely exceeds 6 nodes. Therefore path-based systems can be expected to be practically efficient in real applications.

Finally, the system should provide the ability to store and reuse queries and their implementation (a per-user "history"), the ability to suggest default settings for preferences, filters and ranking, and the ability to share results between trusted sets of users.

5. Current path-based guiding systems

Having described user requirements and expected features of path-based guiding systems, we now summarize systems that are currently available for biological applications: BioMediator[34,41], Biozon[2], BioNavigation[25] and BioGuide[10]. All of these systems provide a graph-based representation of a biological domain and exploit cross-references to navigate within the biological network.

Table 1 gives an overview of the features of these four systems, and indicates for each: (i) whether it allows a graphical representation of the conceptual (C) and physical (P) graphs, and what the complexity of the mapping between these graphs is (x nodes in the conceptual graph correspond to y nodes in the physical graph), (ii) the query language of each system, (iii) whether it allows the user to express preferences and approaches, (iv) whether it is architecture-independent.

BioMediator[34] was the first system to consider queries based on a biological semantic network layer over physical data sources. In BioMediator, the conceptual

Systems	Graphs(C,P)(x-y)	Query Language	Preferences	Archi.-indep
BioMediator	(Yes,No)(1-n)	XQuery-like	Limited	No
Biozon	(Yes,No)(1-n)	Web-Forms	No	No
BioNavigation	(Yes,Yes)(1-n)	Graphical/LR(E)	Yes	Yes
BioGuide	(Yes,Yes)(n-n)	Graphical/XPR	Yes	Yes

Table 1. Current path-based systems

graph is a mediated schema to which the sources are mapped; each path in the graph is a query plan. The project focuses on an XML mediator approach using the query language XQuery. BioMediator considers NCBI sources for which wrappers are available. BioMediator is thus currently dedicated to users who know XQuery, and cannot be used online. Moreover, the use of preferences and metrics to filter and rank paths as well as the possible effect of divergence between alternative paths were not fully considered in the context of this project.

Biozon[2] is another recent and very interesting system that allows the user to ask queries by selecting entities from a conceptual graph. Web-forms related to these entities may then be filled in by the user to express his query. However, Biozon does not consider generating multiple and alternative paths between sources before accessing instances of data. The conceptual graph is fixed, only eight entities are considered, and there is only one edge between two entities. Instances of data are stored in a curated local data warehouse. There is therefore no physical graph of sources. Mixed queries are not allowed, and neither preferences nor approaches may be expressed.

BioGuide[10] and BioNavigation[25] are the only two systems providing a graphical representation of the two graphs, proposing both graphical and formal query languages, allowing users to express preferences and approaches, and having been designed without being associated with a given architecture. We therefore present these systems in more detail, and briefly compare them.

5.1. *BioNavigation*

In BioNavigation[25,26,29], a query is defined as a regular expression *LR(E)* over the alphabet V of scientific entities. Thus an expression e is defined recursively as follows:

$$e == c|e.e|e.(*).e|e.(+).e$$

where $c \in V$. The symbol "$(*)$" represents the Kleene's closure, and denotes that any number of entities may be considered; this operator is analogous to descendant traversal "//" in XPath[49]. The symbol "." expresses that a link must exist between two particular sources. Finally, the symbol "$(+)$" specifies that one or more occurrences of any scientific entity may appear. Regular expressions on the set V capture transparent queries.

The *semantics* of an expression $e = e_1.....e_m$ in this language is the set of paths in the physical graph which satisfies the expression. Recall that a node in the physical graph is a data source which is associated with one class in the conceptual graph via a conceptual-physical graph mapping.

Then, intuitively, a path $s_1.s_2.....s_n$ satisfies e if each s_i can be mapped to some e_j such that:

(1) if $e_j \in V$ then the data source s_i which maps to e_j is also associated with e_j in the conceptual-physical graph mapping (i.e. its scientific entity is "correct");
(2) if e_j is $(*)$ (resp. $(+)$) then zero (resp. one) or more consecutive s_i's are mapped to it; and
(3) the order between elements of the paths is preserved in this semantic mapping.

As an example, consider a scientist interested in retrieving citations related to a particular disease. This corresponds to the query, *Disease*.$(*)$.*Citation*, specifying paths that start at a particular source which provides information on diseases, traverses any number of other sources using links, and ends at a source which provides information on Citations. One physical path which conforms to this would be to start at OMIM and follow the NCBI PubMed Link from OMIM to PubMed. Here, OMIM is mapped to *Disease* and PubMed is mapped to *Citation*; no source is mapped to $(*)$. However, there are many other physical paths which conform to the query involving one or more intermediate data sources, the entity classes of which are unimportant.

Mixed queries are expressed with a specification of the resources that need to be included in (or removed from) the physical paths. Thus in the example above, the user could specify that *Disease* can only be mapped to OMIM.

The browsing mode of BioNavigation allows the user to navigate the conceptual and the physical graphs. The user can click on a node in the physical graph to learn more about its properties, including the scientific class it is mapped to, the URL of the source, the number of entries, the schema of data records, etc. Similarly, the user can click on a link between two nodes in the physical graph to view its properties. At the conceptual level, the user can explore what data sources each node represents.

The *querying mode* allows the user to enter a query, or graphically build a query by selecting the nodes in the conceptual graph, and run the guiding system. For mixed queries, users can specify which data sources they wish to use for a scientific entity by selecting nodes in the physical graph that are mapped to that entity.

Graphical User Interface

Figure 2 shows a screen shot of the first version of the BioNavigation graphical user interface.[h] Note that the nodes (ovals) at the top represent scientific entities (conceptual graph), while those at the bottom represent data sources (physical graph).

[h]BioNavigation is available at http://bioinformatics.eas.asu.edu/bionavigation.htm

An edge between two data sources represents a tool-link or cross-reference. An edge between a scientific entity and a data source represents the semantic mapping. The pane to the right is the query builder.

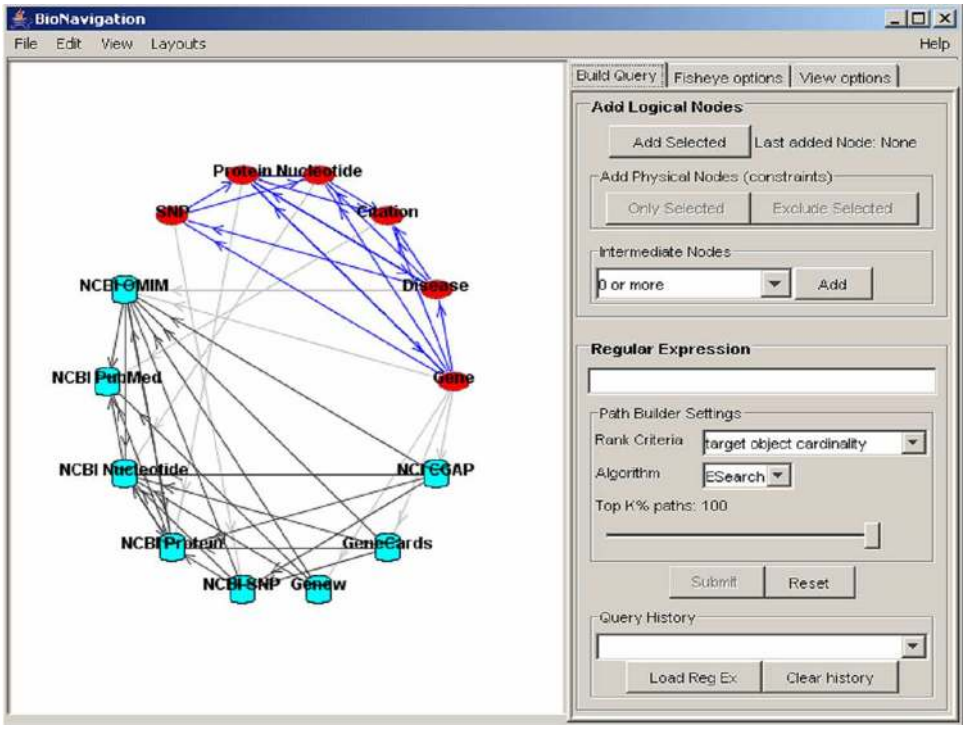

Fig. 2. The BioNavigation Interface

Once a query is built, BioNavigation generates the search space of all physical simple paths validating the regular expression. A *naive* evaluation returns all the physical paths that match the query. This process generates a potentially exponential number of paths, and is therefore neither efficient nor useful to the scientist. BioNavigation therefore performs an exhaustive breadth-first search and returns a list of simple physical paths ordered with respect to a selected user preference (algorithm *ESearch*[29]). A simple path repeats no nodes.

User preferences are handled by annotating each node (biological resource) and edge (tool-link or cross-reference) in the physical graph with metadata information. For biological sources, this information includes the number of entries and the number of attributes. For tool-links and cross-references, this includes the number of distinct pairs of entries between the two sources (*link cardinality*), the number of entries from the input source that are linked to at least an entry in the output source (*link participation* or *domain*), and the number of entries in the output source

linked from at least an entry from the input source (*link image*).

ESearch uses a deterministic finite automaton (DFA) that recognizes the regular expression query to traverse the physical graph, and produces only the physical paths that implement the paths in the conceptual graph that are accepted by the DFA. The ESearch algorithm runs in polynomial time in the size of the physical graph, when the physical graph is cycle-free and there is a path between any two nodes[48]. If d is the maximum number of sources that can precede a source in the physical graph, and b is the maximum length of (cycle free) physical paths, then $O(d^b)$ is an upper bound for ESearch.

The list of paths returned by the ESearch algorithm represents the different ways in which the user can navigate through the data sources in order to evaluate the query. The paths are ranked with respect to the following three metrics:

- *Path Cardinality*, the number of instances of paths of the result. For a path of length 1 between two sources s_1 and s_2, it is the number of pairs (e_1, e_2) where e_1 is an entry in s_1 linked to entry e_2 of s_2.
- *Target Object Cardinality*, the number of distinct objects retrieved from the final data source; and
- *Evaluation Cost*, the cost of the evaluation plan, which involves both the local processing cost and remote network access delays.

In Figure 2, this selection can be made by the user under "Path Builder Settings" in the pane to the right.

Although BioNavigation has many nice features and meets the requirements in Section 4, there are a few limitations in its current implementation (BioNavigation 1.0). First, only one tool-link or cross-reference between two data sources is allowed in the physical graph. Second, ESearch produces simple paths, and therefore will never visit the same source more than once in a path (although it may be useful for refining the final result).

5.2. *BioGuide*

As in BioNavigation, BioGuide[10,11] presents both a conceptual and physical graph to the user. However, the semantic mapping between the two is more complex.

Rather than mapping a data source to a single scientific entity, BioGuide allows a data source to be mapped to a *set* of entities. Thus each node in the physical graph is a "source-entity" – i.e., the view of an entity in a given source – and the semantic mapping associates sources-entities with scientific entities. There are labels on edges of the two graphs. Labels on arrows in the physical graph specify the kind of link: cross-reference (CrossRef), internal link (Internal, links between entities in the same source), or tools-link (e.g., Blast). In both the conceptual and physical graphs, there may be multiple edges between two nodes. For example, three different relationships are considered between the entities Gene and Protein: *similarSeq*, *encodedBy*, and *translated*. Similarly, to implement the cross-reference *similarSeq* several different

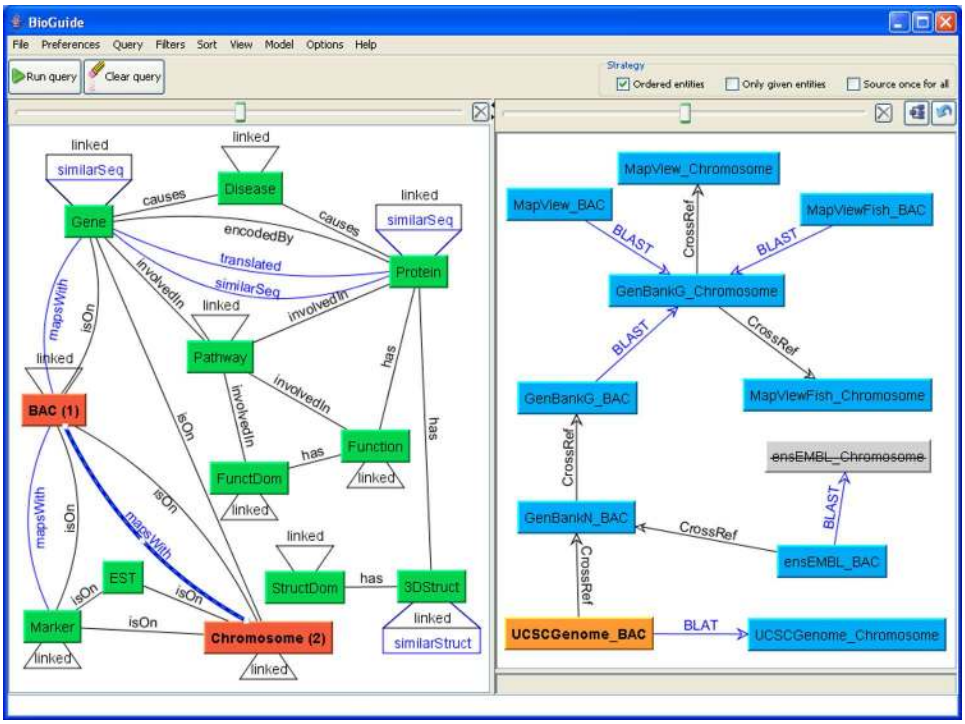

Fig. 3. BioGuide main interface. On the left hand side: the graph of entities where two entities (Bac, Chromosome) and a relationship (mapsWith) have been selected by the user. On the right hand side: the sub-part of the graph of sources-entities corresponding to the mapsWith relationship between Bac and Chromosome. The user has specified she does not want the ensEMBL source to be used to find Chromosome (consequently, ensEMBL_Chromosome is struck out). She also has specified that she wants the UCSCGenome source to be used to find BAC information (consequently, UCSCGenome_BAC appears in bright). The strategy with ordered entities is considered (as seen on top of the figure and indicated on the conceptual graph by numbered entities).

BlastX tools could be used.

BioGuide's graphical user interface is shown in Figure 3. The conceptual graph is shown on the left, and a portion of the physical graph on the right. The semantic mapping between the two graphs is many-to-many, and can be visualized through the BioGuide interface: By clicking on a given node in the conceptual graph the user can determine which sources contain this entity; similarly, by clicking on a relationship the user can determine which links achieve this relationship. BioGuide thus offers browsing capabilities.

The BioGuide user interface allows a progressive use of features by moving complex and less frequently used options out of the main user interface (Figure 3) into secondary screens (Figure 4). In this way, BioGuide allows non-experienced users to exploit default values while permitting experienced users to customize the system according to their needs: specifying strategies, defining filters, ranking methods,

modifying preferences values etc. The next paragraphs are dedicated to these points.

Querying
Users specify queries in BioGuide by selecting a set (possibly ordered) of entities in the conceptual graph. From this set of entities, a list of physical paths is enumerated using strategies and preferences.

Strategy criteria are alternative approaches for implementing paths. During interviews with scientists, it became apparent that they differ in whether or not they (i) followed an order on the entities; (ii) were willing to explore other, unspecified, entities (analogous to (*) in BioNavigation); and (iii) were willing to visit a source more than once. We term these elementary strategy criteria *Ordered*, *OnlyGivenEntities* and *SourceOnceForAll*, respectively. Note that *SourceOnceForAll* allows cycles in the enumerated paths. While it seems counter-intuitive that cycles should be allowed, they are used by scientists to validate information already obtained: Visiting a given source multiple times allows the biologist to check whether or not the information obtained has remained coherent. This process is particularly interesting when accessing data sources which are not curated.

The strategy criteria can be selected through the graphical user interface, and their combination forms the *query strategy*. Selecting one or several criteria ensures that only paths which meet the criteria are enumerated in the implementation. The usefulness of BioGuide's query strategy concept was shown[10]; an independent study has also underlined these results[23].

Preferences
BioGuide considers subjective preferences, such as the reliability of entities in a data source, the confidence the user has in links, and the completeness of the data source. Metadata driven preferences have not been explicitly taken into account, but could be used to guide the user in assigning subjective preference values. Initial values for preferences are set by the system, but can be adjusted by the user (see Figure 4).

Preference values are used to filter and order the paths (see 4th and 5th menu options on Figure 3). All three levels of filters are provided: global (e.g., length of the paths), intermediate (e.g., reliability of the sources providing a given entity), and local (e.g., particular sources or tools to be considered or avoided). Various path estimation operations are provided, including *Weighted Sum*, in which the confidence value of the path is the average of the confidence values of all of the nodes and arrows of the path, and *Best source* in which the confidence value of the path is the value of the node having the highest confidence value.

BioGuide can be customized to each user by creating new kinds of preferences, as well as by changing the content of the conceptual and physical graphs, i.e., adding/removing/modifying links and nodes. The resulting configuration can then be saved to an XML file for future use by the user.

Although not visible to the user, there is an underlying query language for

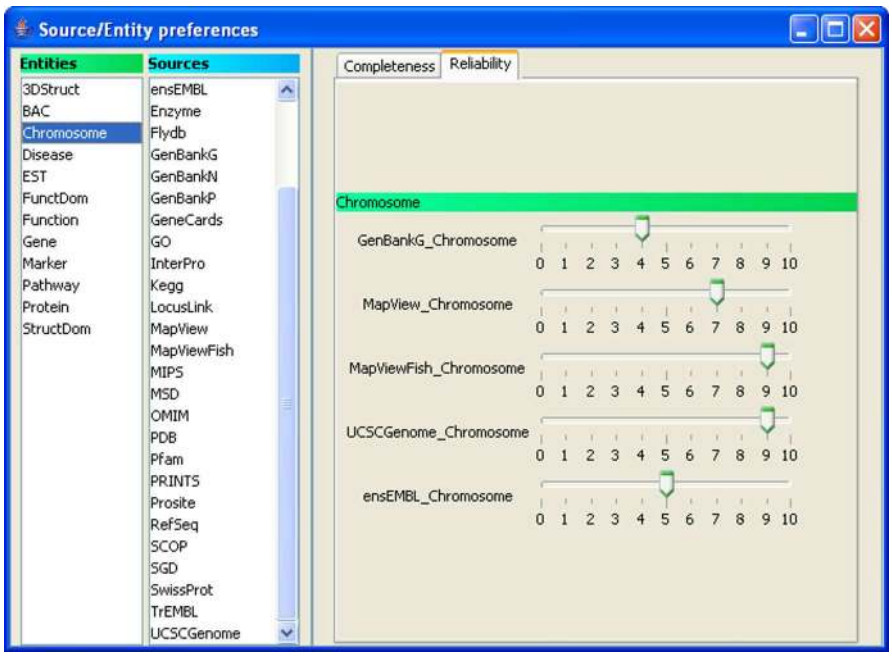

Fig. 4. Initializing Preferences. Reliability values of the sources containing the Chromosome entity.

BioGuide called XPR (eXtensible Path language for RDF)[11]. XPR is expressive enough to model all BioGuide queries, and takes into account selected elements of the graphs (entities, and possibly relationships, sources, tools), as well as preferences and strategies. XPR is based on a complete RDF representation of BioGuide, and is well-suited to the representation of biological entities and Web sources structured as multi-labeled graphs. XPR queries can be saved for reuse, exchanged in collaborations between experts, compared in terms of expressiveness, and efficiently evaluated.

Evaluating paths

BioGuide was initially used in the context of the HKIS-project to deal with the positioning of BACs on the genome, in particular for the "BAC augmentation protocol" introduced in Section 2. In this usage, BioGuide generated alternative paths which were then manually followed by biologists to get instances of data.

More recently, BioGuide has been placed on top of the SRS integration system[14] (*BioGuideSRS*). In BioGuideSRS[7,8], paths are created using preferences and query strategy, and automatically implemented using SRS links between sources. Building the SRS dedicated conceptual and physical graphs was very easy, and only took a few hours. Furthermore, the initial preference values were easily set using information given by the SRS system (e.g., the number of instances per source etc.). BioGuide

is available for use at http://bioguide-project.net.

5.3. *Discussion*

BioGuide and BioNavigation are two similar path systems that offer distinct viewpoints on how to build paths between resources to exploit the richness of the resources. Both systems use a formal, graph-based framework and provide the user with graphical interfaces to view the conceptual and physical layers. The two systems are user-friendly, and allow users to express both transparent and mixed queries. They also allow various ways of expressing preferences and approach, by providing means of filtering/ranking paths. Thus, BioNavigation and BioGuide meet the requirements and expected features introduced in Sections 3 and 4.

BioNavigation and BioGuide differ in some aspects, such as the underlying query language used (LR(E), a regular expression language vs. XPR, an RDF-based path language), and the structure of their graphs (single links vs multiple links between two nodes). However, the most salient differences between the two systems spring from their initial objectives. The goal of BioNavigation is to generate a limited, ordered and selected set of paths, and hence optimize the way in which data is searched for. This is done by returning the best paths according to the query and the metrics (metadata) considered. On the other hand, the goal of BioGuide is to generate an exhaustive set of alternative paths for finding data, carefully filtered and ordered according to users's preferences. BioGuide provides the user with a (potentially large) set of paths with an ordering which enables the user to favor some of them and to deal with divergent data.

This difference in initial goals has obvious consequences in the systems at two levels: the kind of filter criteria they provide and the complexity of their algorithms.

First, BioNavigation uses metadata based on properties of sources (e.g., number of instances contained, number of attributes, etc.) while BioGuide mainly uses *subjective* metadata (e.g., reliability of a source, ease-of-use, completeness, etc.).

The metadata in BioGuide can be initialized using objective properties (e.g., using information from SRS, as seen in Section 5.2) but the initial values can always be modified by the user. It is thus not surprising that in BioNavigation the metadata can be consulted by the user while in BioGuide several interfaces are given to allow the user to view and modify the preference values. Nevertheless, both systems allow filters to be specified at the three levels discussed in Section 3.2.4 and provide different sort methods to order these paths.

Second, the complexity of the algorithms is different. In BioNavigation, the ESearch algorithm runs in polynomial time in the size of the graph assuming some constraints on the graph (e.g., the physical graph is cycle-free and there is one path between any two nodes). In contrast, the worst case time complexity of BioGuide is high[10], as it searches for an exhaustive set of alternative paths, allows various strategies, and does not make any assumption on the topology of the graph. This complexity is unavoidable because the number of ordered paths in a graph can

be exponential[10,48]. However, whatever the chosen strategy is, the set of sources-entities paths is, in practice, very rapidly generated in the prototype version of BioGuide. Therefore, BioNavigation and BioGuide can be both used efficiently in practical cases.

6. Conclusion

In this paper, we have discussed the need for systems to support scientists in constructing digital scientific protocols. In particular, we have analyzed how scientists currently construct digital protocols, and highlighted the role of preferences and approach in selecting appropriate implementations. Building on this, we discussed essential features of a path-based system, whose goal is to generate paths composed of biological resources where each path provides an alternative way to obtain data. Alternative paths are especially useful for integrating biological data because they exploit the complementarity of biological sources and provide opportunities for dealing with divergent data.

More generally, we claim that path-based systems should be architecture-independent and should consider different degrees of coverage at the conceptual level. The conceptual level presumes that the biological entities together with the relationships between them can be listed as exhaustively as possible, i.e. that an ontological representation of all biological entities structured by the relationships between them can be provided. A complete catalog of resources for biological data sources and for bioinformatics tasks[44] would be highly useful, and could be structured according to preferences and biological entities relative to the resources in a "resource ontology". However, for a particular use of the system it may be more appropriate to have a specialized conceptual level tailored for the domain area, i.e. equipped with partial ontologies that model the current knowledge about the biological entities needed or the available resources.

As an illustration, BioNavigation provides a semantic map of services for structural biology[47].[i] BioGuide was first used manually for oncology-related queries and is currently being used on top of the SRS integration system (BioGuideSRS[7,8]) for genomic queries.

Path-based systems must be based on top of an integration platform for the implementation of queries[9]. Ideally, they should also be placed in the context of a scientific workflow system to help with the specification and management of bioinformatic tasks, flow of data between tasks, and to store and reuse workflow components. For example, BioGuide was initially tested on top of the HKIS-platform, a workflow system for managing scientific workflows[12]; on-going work on BioNavigation is placing it within SemanticBio[24], a workflow design and execution system, enhancing it by allowing the exploration of execution paths prior to the execution of the scientific protocol.

[i]In that context, a new version of BioNavigation is being developed.[27,30]

In the context of workflow systems, the tool-links offered should include all possible ways of connecting data. Consequently, analysis tools (e.g., statistical tools to analyze micro-array data) should be included in such solutions. Another challenging issue is to find automatic ways to extract metadata from sources. For example, grid infrastructures[45] typically exploit statistics such as usage frequency and response time.

It will also be extremely important to consider data provenance, i.e., record where data came from and how it was used in a chain of derivations leading to some result[6]. Thus a large part of future work for path-based systems will be to answer questions like *"Where did this data come from?*, *"What sequence of bioinformatics tasks led to this result?"* or *"Where was this data used in a bioinformatics task?"*

Acknowledgements

Development of BioNavigation was partially supported by the NSF grants IIS0431174 and IIS0223042, and by the grant 1 R03 LM008046 from the NIH National Library of Medicine, as well as USB-DID grants.

Work on BioGuide was supported in part by the European Project HKIS IST-2001-38153, the Fulbright Program and a Hitachi Chair at INRIA. This material is also based upon work supported by the National Science Foundation under Grants No. 0415810 and 0513778.[j]

References

1. J. C. Bartlett and E. G. Toms. Developing a protocol for bioinformatics analysis: An integrated information behaviors and task analysis approach. *Journal of the American Society for Information Science and Technology*, 56(5):469–482, 2005.
2. A. Birkland and G. Yona. Biozon: a system for unification, management and analysis of heterogeneous biological data. *BMC Bioinformatics*, 2006.
3. S. Bowers and B. Ludäscher. Actor-oriented design of scientific workflows. In *International Conference on Conceptual Modeling (ER)*, pages 369–384, 2005.
4. P. Buneman, S. Khanna, and W. Tan. Why and where: A characterization of data provenance. In *International Conference on Database Theory (ICDT)*, pages 316–330, 2002.
5. C. Burks. Molecular biology database list. *Nucleic Acids Research*, 27(1):1–9, 1999. http://nar.oupjournals.org/cgi/content/full/27/1/1.
6. S. Cohen, S. Cohen-Boulakia, and S. Davidson. Towards a model of provenance and user views in scientific workflows. In *Data Integration in the Life Science*, volume 4075, pages 264–279. LNBI Springer-Verlag, 2006.
7. S. Cohen-Boulakia. *Integrating Biological data: Selection of Sources following a User-centric Approach*. PhD thesis, PhD in Computer science, University of Paris-Sud, 2005.
8. S. Cohen-Boulakia, O. Biton, and C. Froidevaux. Bioguidesrs: Querying multiple

[j]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.

sources with a user perspective. Technical report, University of Paris-Sud, #1436, 2006.
9. S. Cohen-Boulakia, Olivier Biton, S. Cohen, Z. Ives, V. Tannen, and S. Davidson. Sharq guide: Finding relevant biological data and queries in a peer data management system. 2006.
10. S. Cohen-Boulakia, S. Davidson, and C. Froidevaux. A user-centric framework for accessing biological sources and tools. In *Data Integration in the Life Sciences*, volume 3615, pages 3–18. LNBI Springer-Verlag, 2005.
11. S. Cohen-Boulakia, C. Froidevaux, and E. Pietriga. Selecting biological data sources and tools with xpr, a path language for rdf. In *Pacific Symposium on Biocomputing (PSB)*, 2006.
12. S. Cohen-Boulakia, S. Lair, N. Stransky, S. Graziani, F. Radvanyi, E. Barillot, and C. Froidevaux. Selecting biomedical data sources according to user preferences. *Bioinformatics*, 20:i86–i93, 2004.
13. J.W. Ely, J.A. Osheroff, P.N. Gorman, M.H. Ebell, M.L. Chambliss, E.A. Pifer, and P.Z. Stavri. A taxonomy of generic clinical questions: classification study. *British Medical Journal BMJ*, 321 (7258):429–432, 2000.
14. T. Etzold, H. Harris, and S. Beaulah. *SRS - An Integration Platform for Databanks and Analysis Tools*, chapter 5, pages 109–145. Morgan Kaufmann Publishing, 2003.
15. J.C. Flanagan. The critical incident technique. *Psychological Bulletin*, 5(4):327–358, 1954.
16. I. Foster, J. Voeckler, M. Wilde, and Y. Zhao. The virtual data grid: A new model and architecture for data-intensive collaboration. In *Conference on Innovative Data System Research (CIDR)*, 2003.
17. F. Fowler. *Survey Research Methods.* London:Sage, 1984.
18. M. Y. Galepin. The molecular biology database collection: 2006 update. *Nucleic Acids Research*, 34:D3–D5, 2006.
19. C. Goble. Position statement: Musings on provenance, workflow and (semantic web) annotations for bioinformatics. In *Workshop on Data Derivation and Provenance*, 2002.
20. U. Greiner, R. Muller, E. Rahm, J. Ramsch, B. Heller, and M. Loffler. Adaptflow: Protocol-based medical treatment using adaptive workflows. In *Methods of Information in Medicine*, pages 80–88, 2005.
21. G. Hoinville, R. Jowell, and Associates. *Survey Research Practice.* London:Heinemann, 1978.
22. Z. Lacroix. Public data sources and applications used by scientists. Technical report, 2003.
23. Z. Lacroix. Evaluating similar implementations of a scientific protocol on ncbi resources. Technical report, 2005.
24. Z. Lacroix and H. Ménager. Semanticbio: Building conceptual scientific workflows over web services. In *Data Integration in the Life Sciences*, volume 3615. LNBI Springer-Verlag, 2005.
25. Z. Lacroix, T. Morris, K. Parekh, L. R., and M.-E. Vidal. Exploiting multiple paths to express scientific queries. In *Scientific and Statistical Database Management (SSDBM)*, pages 357–360. IEEE Computer Society, 2004.
26. Z. Lacroix, H. Murthy, F. Naumann, and L. Raschid. Links and paths through life science data sources. In *Data Integration in the Life Science*, volume 2994, pages 203–211. LNBI Springer-Verlag, 2004.
27. Z. Lacroix, K. Parekh, M.-E. Vidal, M. Cardenas, and N. Marquez. BioNavigation: Selecting Optimum Paths through Biological Resources to Evaluate Ontological Navi-

gational Queries. In *Data Integration in the Life Science*, volume 3615, pages 275–283. LNBI Springer-Verlag, 2005.
28. Z. Lacroix, L. Raschid, and B. Eckman. Techniques for optimization of queries on integrated biological resources. *Journal of Bioinformatics and Computational Biology*, 2(2):375–411, 2004.
29. Z. Lacroix, L. Raschid, and M.-E. Vidal. Efficient techniques to explore and rank paths in life science data sources. In *Data Integration in the Life Science*, volume 2994, pages 187–202. LNBI Springer-Verlag, 2004.
30. Z. Lacroix, L. Raschid, and M.-E. Vidal. Semantic Model to Integrate Biological Resources. In *Semantic Web and Databases Workshop*, 2006.
31. A. Levy. Combining artificial intelligence and databases for data integration. *Artificial Intelligence Today: Recent Trends and Developments*, 1600:249–268, 1999.
32. B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger-Frank, M. Jones, E. Lee, J. Tao, and Y. Zhao. Scientific workflow management and the kepler system. *Concurrency and Computation: Practice & Experience, Special Issue on Scientific Workflows*, 2005.
33. M. Mecella, M. Scannapieco, A. Virgillito, R. Baldoni, T. Catarci, and C. Batini. Managing data quality in cooperative information systems. *Journal of Data Semantics, Volume I*, 2800:208–232, 2003.
34. P. Mork, A. Halevy, and P. Tarczy-Hornoch. A model for data integration systems of biomedical data applied to online genetic databases. In *American Medical Informatics Association (AMIA) Annual Symposium, American Medical Informatics Association, ePublication*, 2001.
35. H. Muller and F. Naumann. Data quality in genome databases. In *International Conference on Information Quality*, pages 269–284, 2003.
36. F. Naumann, U. Leser, and J.C. Freytag. Quality-driven integration of heterogenous information systems. In *International Conference of Very Large DataBases (VLDB)*, pages 447–458, 1999.
37. T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, and P. Li. Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17):3045–3054, 2004.
38. M. Ould. *Strategies for Software Engineering : The Management of Risk and Quality.* Chichester: Wiley. (Wiley series in software engineering practice.), 1990.
39. S. Presser, J. M. Rothgeb, M. P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. *In Methods for Testing and Evaluating Survey Questionnaires.* New York: Wiley, 2004.
40. A. Rowe, D. Kalaitzopoulos, M. Osmond M., Ghanem, and Y. Guo. The discovery net system for high throughput bioinformatics. In *Bioinformatics*, pages i225–i231, 2004.
41. R. Shaker, P. Mork, J. S. Brockenbrough, L. Donelson, and P. Tarczy-Hornoch. The biomediator system as a tool for integrating biologic databases on the web. In *Workshop on Information Integration on the Web (held in conjunction with VLDB), ePublication*, 2004.
42. A. Slominski and G. von Laszewski, 2006. http://www.extreme.indiana.edu/swf-survey/.
43. R. Stevens. http://www.cs.man.ac.uk/ stevensr/ontology.html.
44. R. Stevens, C. A. Goble, P.G. Baker, and Andy Brass. A classification of tasks in bioinformatics. *Bioinformatics*, 17(1):180–188, 2001.
45. R. D. Stevens, A. J. Robinson, and C. A. Goble. mygrid: personalised bioinformatics on the information grid. *Bioinformatics*, 19(90001):302i–304, 2003.

46. R. Targino, M.C. Cavalcanti, and M. Mattoso. An environment to define and execute in-silico workflows using web services. In *Data Integration in the Life Sciences*, volume 3615, pages 288–291. LNBI Springer-Verlag, 2005.
47. P. Tufféry, Z. Lacroix, and H. Ménager. Semantic map of services for structural bioinformatics. In *International Conference on Scientific and Statistical Database Management (SSDBM)*, 2006.
48. M. Vidal, L. Raschid, N. Marquez, M. Cardenas, and Y. Wu. Query rewriting in the semantic web. In *Workshop on Database Interoperability (InterDB06). Held in conjunction with ICDE*, 2006.
49. http://www.w3.org/TR/xpath.
50. J. Zhao, C. Wroe, C. Goble, R. Stevens, D. Quan, and M. Greenwood. Using semantic web technologies for representing e-science provenance. In *Semantic Web Conference (ISWC2004)*, pages 92–106, 2004.

Sarah Cohen Boulakia is a post-doctoral researcher at the University of Pennsylvania where she works with Prof. Susan Davidson. She defended her PhD in Computer Science in 2005, under the supervision of Prof. Ch. Froidevaux at the Laboratoire de Recherche en Informatique, University of Paris-Sud 11, France.

Dr. Cohen Boulakia's research interests are in the design and application of integration systems dedicated to biological and biomedical domain. She is best known for her work on BioGuide and techniques for supporting biologists navigate the maze of biological resources available over the web. In this work, she collaborates closely with biologists, physicians, and computer scientists.

Susan B. Davidson received the B.A. degree in Mathematics from Cornell University, Ithaca, NY, in 1978, and the M.A. and Ph.D. degrees in Electrical Engineering and Computer Science from Princeton University, Princeton NJ, in 1980 and 1982. Dr. Davidson joined the University of Pennsylvania in 1982, and is now the Weiss Professor of Computer and Information Science and Deputy Dean of the School of Engineering and Applied Science. She is an ACM Fellow, a Fulbright scholar, and recently stepped down as founding co-Director of the Center for Bioinformatics at UPenn (PCBI).

Dr. Davidson's research interests include database systems, database modeling, distributed systems, bioinformatics and real-time systems. Within bioinformatics she is best known for her work with the Kleisli data integration system, and more recently with XML as a data update and integration strategy.

With a background in Mathematics (École Normale Supérieure Fontenay, agregation of Mathematics), **Christine Froidevaux** has defended her PhD in 1983 in computational linguistics at the University of Paris 7. She is full Professor of Computer Science since 1993 at the University of Paris-Sud 11, France. She is head of the Bioinformatics research group at the Laboratoire de Recherche en Informatique since 2001. She works in close collaboration with microbiologists and oncologists.

Her main research topics include integration of heterogeneous genomic data, bio-ontology design, representation and analysis of genomic data, and inference in biological databases dedicated to functional annotation.

Zoé Lacroix received her Ph.D. in Computer Science in 1996 from the University of Paris XI - Orsay (France). She has been a researcher at the French Institut National de la Recherche en Informatique et Automatique (INRIA), at the Institute for Research in Cognitive Science (IRCS) at the University of Pennsylvania (USA), and at two biotech companies Gene Logic and at SurroMed, where her research focused on bioinformatics. She is currently Associate Professor of Research at Arizona State University where she is directing various projects on scientific data management, database integration, optimization, semantics of Web queries, and Semantic Web.

She co-edited the first textbook on scientific data management, entitled "Bioinformatics: Managing scientific Data" and published by Morgan Kaufmann in July 2003, and is the Principal Investigator of research projects funded by the National Science Foundation and National Institutes of Health: National Institute on Aging and National Library of Medicine.

María Esther Vidal received a Bachelor on Computer Engineering on 1987, a Master on Computer Science on 1991, and PhD on Computer Science on 2000 from Universidad Simón Bolívar, Caracas Venezuela. Dr. Vidal is a Full Professor of the Computer Science department at Universidad Simón Bolívar. Dr. Vidal has been Assistant Researcher at the Institute of Advanced Computer Studies at the University of Maryland (UMIACS) (1995-1999), and Visitor Professor at UMIACS (2000-2005) and at Universidad Politecnica de Catalunya (2003).

Dr. Vidal has reported her research on AAAI, IJCAI, SIGMOD, CoopIs, WIDM, WebDB, ICDE, DILS, DEXA, OTM, and SIGMOD RECORDS. Her current research interests are query rewriting and optimization in emerging infrastructures.