

A Semantic Map to select Structural Bioinformatics services

Zoé Lacroix* and Hervé Ménager
Scientific Data Management Laboratory
Arizona State University
Tempe, AZ 85287-5706, USA
zoe.lacroix@asu.edu
and herve.menager@asu.edu

Pierre Tufféry
Equipe de Bioinformatique Génomique et Moléculaire
INSERM U726, Université Denis Diderot (Paris 7)
Tour 53-54, 1er étage, case 7113
2, place Jussieu, 75251 Paris Cedex 05, France
pierre.tuffery@ebgm.jussieu.fr

Structural Bioinformatics covers the prediction and analysis *in-silico* of biological molecular structures with the goal to understanding functional mechanisms at a molecular level. Due to the significant effort of the scientific community, this field has dramatically evolved over the recent years, in particular for proteins. The techniques available to predict and analyze protein structures are continuously improving both in their focus and their performances while new algorithms are developed. In such a context, the scientists face the increasing difficulty of identifying and accessing existing tools and understanding how the results should be interpreted and contribute to scientific discovery.

BioNavigation is a path-based system that guides scientists facing difficulties to identifying resources suitable to express their scientific protocols. It exploits a description of bioinformatics resources, including applications and data repositories, organized with respect to their scientific aim in an semantic map. Scientists express their scientific protocols in terms of scientific concepts and relationships selected from an *ontology*, the system returns the succession of scientific resources suitable to express their protocols.

In this paper, we present a semantic map of services for structural bioinformatics, applied to proteins including: (a) A semantic description of each service that captures an abstraction of the service rather than a low level syntactic description. This description is expressed in terms of an ontology of the services, linking items of the structural bioinformatics concepts ontology; (b) Service identification, detailing for each their purpose, the type of the data on which they are effective, and the type of result they provide; and (c) Exploration of available services, navigating through the graph composed of the possible interconnections between the services. The semantic map for structural bioinformatics services and its consultation of resources via BioNavigation is made available freely at <http://bioserv.rpbs.jussieu.fr/SemanticMap/>.

1. Introduction

Scientists face increasing difficulties identifying and locating reliable resources to implement the data retrieval and analysis tasks of their protocols. This situation is similar for the systems, robots, and applications that are accessing these resources, developing and updating data warehouses, or wrapping and querying data from public data repositories. Interoperability between biological resources persists as a problem, despite efforts to facilitate the exchange between applications and data repositories by developing specialized formats (such as FASTA for sequences) or wrapping packages (such as BioPerl¹⁴),

and attempts to develop unified languages to describe resources (such as BioMoby¹⁶ or myGrid¹⁵ for Web services¹⁰). Standards for data representation and a semantic layer that specifies the characteristics of the resource will undoubtedly ease the integration of scientific resources. Nevertheless, the problem of understanding, exploiting, and representing semantic information such as data coverage, data quality, and reliability of resources has not been fully solved, and the selection of the resources to be considered at each step of a protocol remains a critical task.

*Path-based guiding systems*³ are designed to help scientists to identify the resources best meeting their

*Corresponding author.

needs. BioNavigation^{7, 9, 8} uses a semantic map of bioinformatics resources composed of an ontology of the domain and a repository of available resources described in terms of the ontology. Users express their protocols as *conceptual paths* (i.e., successions of concepts of the ontology linked by relationships) and the system returns *physical paths* specifying a selection of resources to implement the protocol.

The *semantic map for structural bioinformatics* Web site located at <http://bioserv.rpbs.jussieu.fr/SemanticMap/> aims at exploiting BioNavigation within the *Ressource Parisienne en Bioinformatique Structurale (RPBS)*², a Web portal designed to provide a wide selection of structural bioinformatics resources to the community.^a Most portals devoted to a particular scientific domain typically list available resources. In contrast, our approach using BioNavigation goes a step beyond by allowing scientists to navigate through the resources via an ontology that captures their scientific meaning. Indeed each data source may be mapped to the concept(s) that characterize(s) its entries in the ontology. Similarly, each application may be mapped to a relationship linking two concepts in the ontology. By exploiting this mapping, BioNavigation allows the scientist to express scientific protocols against the ontology to identify the resources that may capture the selected scientific meaningful path, thus implementing the protocol. The use of BioNavigation in a domain portal enhances significantly the ability to identify resources suitable for a particular protocol. The approach monitored by an expert committee and allowing users to interact and suggest modifications (adding resources, transforming of the ontology itself, or the mapping) offers a valuable service to the community.

The first section of this paper describes the issues raised by the development process of the semantic map, and the need for a collaborative design. Section 3 briefly depicts the architecture of the system. Finally, Section 4 illustrates it with a user scenario.

2. Semantic Map development process - Collaborative construction of the map

The development of the semantic map has been initiated by a short seminar and a set of interviews with a set of domain specialists. These early steps of the development process led us to identify two essential properties:

- (a) The complexity of the domain, that includes an important number of concepts and properties.
- (b) The variability in the perceptions of the domain itself, as the experts can identify distinct concepts and properties, diverge on the importance they assign to each of them, and even structure their perception of the domain differently.

The first property of the map, its complexity, raises the issue of usability. Such a system, if too complex, is not a valuable help to its users, who seek to reduce the effort necessary to develop digitalized protocols.

The latest leads us to design a system oriented toward a collaborative development. The two requirements are:

- (a) to let scientists enrich easily the map, on both on the conceptual level (ontology) and the service level.
- (b) to maintain its consistency in spite of its flexibility.

We plan to publish a CGI interface to our system to let scientists enter new services, that can be easily validated by a committee. On the other hand, any modification to the ontology will be requested by mail, due to its greater complexity and potential impact on the registered services.

3. System architecture

The system we are designing to meet the previously cited issues and requirements can be divided in three parts:

- (a) A repository that stores both the ontology and the service descriptions.

^aThe Ressource Parisienne en Bioinformatique Structurale (RPBS) is available at <http://bioserv.rpbs.jussieu.fr/RPBS/>.

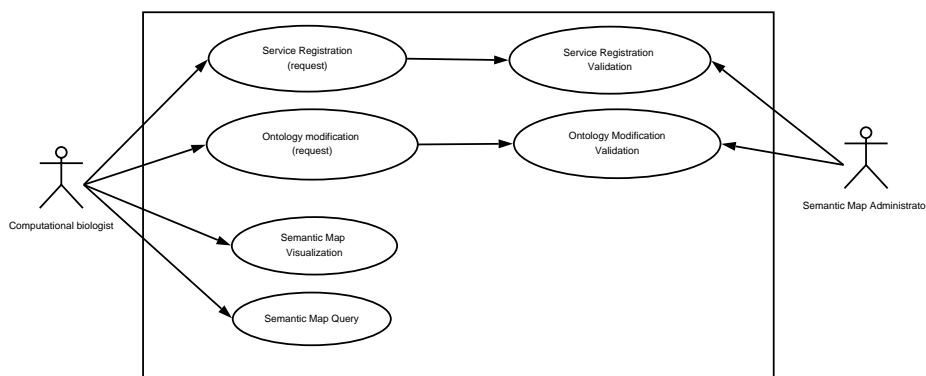


Figure 1. Use Case Diagram

- (b) A collaborative interface that lets users submit new service descriptions.
- (c) A Java™ Applet that permits the visualization and the queries to the map.

3.1. Repositories

The ontology is developed and maintained using the Protégé Ontology Editor⁶, and stored as an OWL file. The service descriptions are stored as XML files, where the mappings to the ontology are represented by references to the ontology. The ontology is a complex graph composed of concepts (nodes) and multi-labeled directed edges (two concepts may be related by several different relationships). The complexity of the graph motivates the use of a sophisticated visualization tool as explained in Section 3.3.

3.2. CGI interface

These descriptions are generated using a CGI form. The modifications and additions to the list are moderated, as well as any modification to the ontology: this validation process is intended to avoid any potential pollution of the system, but more importantly to maintain its global consistency.

3.3. Visualization applet

The overall size and complexity of the managed data calls for the use of a highly powerful and customizable interface. It is fundamental that the graphic representation is simple enough to be meaningful to the user. We use the ZVTM API¹³ to display the map in our interface. Its advanced capabilities to assist navigation, such as zooms, radar views or smooth

spatial motions, improve the perception of the context. Our interface will also let users differentiate various types of relationships and concepts, as well as the resources that “manipulate” them, using distinct display formats and specifying whether they are displayed or not in the graph. The layout of the map is generated on the server side by the GraphViz library.⁵

Figure 2 represents the technical processing of the information

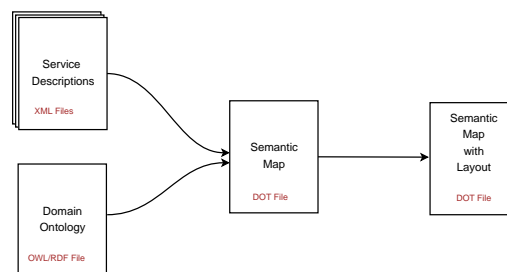


Figure 2. Information Flow in Semantic Map Architecture

4. User Scenario - Exploration of Solvent Accessible Surface related services

To identify the parts of a protein that can potentially interact with a solvent, scientists can use Solvent Accessible Surface Methods. A first approach will be to request what services generate such results. A second one can be to query the graph, to test what kind of protocol (pipeline), composed by the succession of one or more services, can produce the solvent accessible surface if fed with a given amino-acid sequence.

Figure 3 shows a screenshot of the applet. The

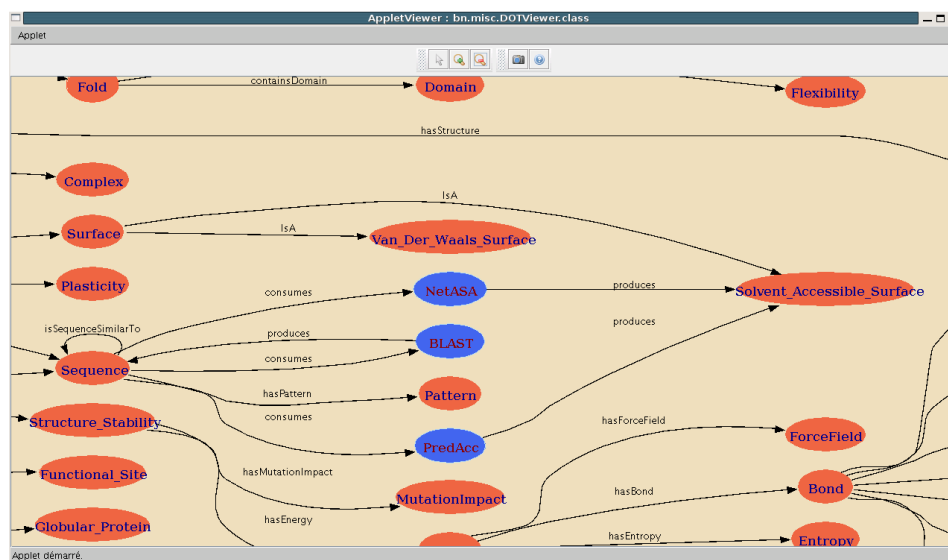


Figure 3. Semantic Map Visualization and Query Applet

concepts are represented with nodes (orange in the interface, lighter gray labeled in black in Figure 3), whose properties can be displayed by clicking on it. For example, **Surface**, **Plasticity**, and **Sequence** are concepts. A right-click informs of all the services (represented as blue nodes in the interface, darker gray labeled in gray in Figure 3) that produce solvent-accessible surface data. By specifying the input data type, an amino acid sequence, the scientist may also query the graph for existing services or pipelines that produce such outputs. In our case, the query results will include:

- services that directly predict the solvent accessible surface from the sequence data, such as NetASA¹ or PredAcc,¹²
- pipelines of services that have the same output, for instance predicting the structure from sequence then computing solvent accessible surface from structure.

When the user knows another method that predicts solvent accessibility from a sequence, he can describe it by going to the service registration page and entering the characteristics of the method that takes as input one sequence, one bank to search for homologous proteins, and produces the result (categorization of residues buried/exposed to solvent). When receiving this new submission, the moderator checks that the method is effective, and if so validates

it. This new service is then available to the next user that wants to know about solvent accessible surface methods.

If the user wants to explore services about the conserved residues at the surface of a protein structure, the present version of the ontology does not include the concept of "conservation". In such case, the user can explain this to the moderator and propose to include the concept. In order to assess this query, the moderator transmits this request to a committee that accepts or rejects the proposal. If accepted, the concept is introduced in the next release of the ontology. This implies checking the consistency of the new ontology with the description of all the services.

5. Conclusion - Future work

The system presented in this short paper is currently under development. Its aim is to enhance the Web portals that provide lists of resources for a particular scientific domain by mapping each resource to its scientific meaning expressed in an ontology. We demonstrate its use in the context of a complex field, structural bioinformatics, by using elements of semantic integration to assist services classification and composition. Future works include the integration of the system with an execution engine such as the SemanticBio workflow engine,¹¹ to provide scientists the

ability to not only select the resources best meeting their protocol needs, but also to execute them, thus computing answers to a scientific problem from a conceptual query.

Acknowledgment

This research was partially supported by the National Science Foundation (grants IIS 0223042 and IIS 0222847) and the Conservatoire National des Arts et Métiers. The authors wish to thank many people for useful discussions about this project, in particular J. Chomilier, J. Pothier, B. Villoutreix, J.-F. Zagury and the members of Equipe de Bioinformatique Génomique et Moléculaire (EBGM), Paris, France. The BioNavigation system is developed in collaboration with L. Raschid, University of Maryland at College Park, and Maria-Esther Vidal, Universidad Simón Bolívar, Caracas, Venezuela.

References

1. S. Ahmad and M. M. Gromiha. NETASA: neural network based prediction of solvent accessibility. *Bioinformatics*, 18(6):819–824, June 2002.
2. C. Alland, F. Moreews, D. Boens, M. Carpentier, S. Chiusa, M. Lonquety, N. Renault, Y. Wong, H. Cantalloube, J. Chomilier, J. Hochez, J. Pothier, B. O. Villoutreix, J.-F. Zagury, and P. Tuffery. RPBS: a web resource for structural bioinformatics. *Nucl. Acids Res.*, 33(suppl.2):W44–49, 2005.
3. S. Cohen-Boulakia, S. Davidson, C. Froidevaux, Z. Lacroix, and M.-E. Vidal. Path-based systems to guide scientists in the maze of biological data sources. (submitted).
4. E. Rahm, editor. *First International Workshop on Data Integration in the Life Sciences (DILS), Proceedings*, volume 2994 of *Lecture Notes in Computer Science, Subseries: Lecture Notes in Bioinformatics*. Springer, 2004.
5. E. R. Gansner and S. C. North. An open graph visualization system and its applications to software engineering. *Software — Practice and Experience*, 30(11):1203–1233, 2000.
6. H. Knublauch. An AI tool for the real world - Knowledge modeling with Protégé, June 2003.
7. Z. Lacroix, T. Morris, K. Parekh, L. Raschid, and M.-E. Vidal. Exploiting Multiple Paths to Express Scientific Queries. In *16th International Conference on Scientific and Statistical Database Management (SSDBM)*, pages 357–360. IEEE Computer Society, June 2004.
8. Z. Lacroix, H. Murthy, F. Naumann, and L. Raschid. Links and Paths Through Life Science Data Sources. In E. Rahm⁴, pages 203–211.
9. Z. Lacroix, L. Raschid, and M.-E. Vidal. Efficient Techniques to Explore and Rank Paths in Life Science Data Sources. In E. Rahm⁴, pages 187–202.
10. P. Lord, S. Bechhofer, M. Wilkinson, G. Schiltz, D. Gessler, D. Hull, C. Goble, and L. Stein. Applying semantic web services to Bioinformatics: Experiences gained, lessons learnt. In *ISWC*, pages 350–364. Springer-Verlag, 2004.
11. H. Ménager and Z. Lacroix. A workflow engine for the execution of scientific protocols. In *ICDE Workshops*, 2006. Accepted for the IEEE Workshop on Workflow and Data Flow for Scientific Applications (SciFlow 2006).
12. M. H. Mucchielli-Giorgi, S. A. Hazout, and P. Tufféry. PredAcc: prediction of solvent accessibility. *Bioinformatics*, 15(2):176–177, 1999.
13. E. Pietriga. A toolkit for addressing hci issues in visual language environments. In *IEEE Symposium on Visual Languages and Human-Centric Computing*, pages 145–152, Sept. 2005.
14. J. Stajich, D. Block, K. Boulez, S. Brenner, S. Chervitz, C. Dagdigian, G. Fuellen, J. Gilbert, I. Korf, H. Lapp, H. Lehvaslaiho, C. Matsalla, C. Mungall, B. Osborne, M. Pocock, P. Schattner, M. Senger, L. Stein, E. Stupka, M. Wilkinson, and E. Birney. The bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, 12(10), Oct 2002.
15. R. D. Stevens, A. J. Robinson, and C. A. Goble. myGrid: personalised bioinformatics on the information grid. *Bioinformatics*, 19(Suppl.1):i302–i304, 2003.
16. M. D. Wilkinson, D. Gessler, A. Farmer, and L. Stein. The BioMOBY Project Explores Open-Source, Simple, Extensible Protocols for Enabling Biological Database Interoperability. In *Virt Conf Genom and Bioinf*, pages 16–26, 2003.