

BioNavigation: Selecting Optimum Paths through Resources to Evaluate Scientific Queries

Zoé Lacroix¹, Kaushal Parekh¹, Maria-Esther Vidal²,
Marelis Cardenas², Natalia Marquez², Louiqa Raschid³

¹Arizona State University, Tempe, Arizona

²Universidad Simón Bolívar, Caracas, Venezuela

³University of Maryland, College Park, Maryland

A scientific data collection protocol is always specified in terms of scientific classes being studied and it need not specify the data sources from which to get the information about these classes. These protocols are also mostly navigational, i.e. scientists start with obtaining information about a particular scientific object then from there go to another using the provided links and so on, thus forming a path. Scientists tend to use only a set of resources which they are familiar with to express their protocols rather than selecting the best possible resource that matches their needs. Most of the times, they do not even know which is the best resource, or even if they are aware that such a source exists, they are not familiar with its features and query interface to effectively exploit it.

Scientists should be able to formulate their queries at the higher conceptual level of scientific classes and their relationships, without the concern of what source would be used underneath to collect the data. This is the ontology level. Classes in the ontology are mapped to the data sources which represent them, for e.g. the scientific class 'gene' is represented by many sources such as Entrez Gene, GeneCards, etc. Similarly the relationships in the ontology are mapped to the physical links between the data sources. These links could be in the form of navigational links, indices or applications that capture the semantics of the ontology level relationships.

The BioNavigation system provides the scientist with valuable guidance in selecting the most effective evaluation path through the physical resources for his ontological query. The main features of the system are:

1. Visualize the conceptual level ontology, the physical level graph of resources and the mappings between the two levels.
2. Browse the physical graph to obtain more information about the resources, e.g. their URL, data formats, schema, etc.
3. Build queries with the help of the ontology by selecting the desired classes connected by labeled relationships.
4. The ESearch algorithm traverses the space of possible physical paths and produces the ones that implement the query.

A query is represented as a regular expression made up of the sequence of scientific classes and relationships to be followed. The user can also specify a wildcard character within a regular expression to indicate that any possible resource can be used in its place. The ESearch algorithm performs an extensive breadth-first search on the physical graph to search for paths that match the users query expression. The algorithm uses metadata information about the data sources to estimate the relative ranks of these paths with respect to the ranking criteria selected by the user. For example the user can chose the path to return the maximum number of entries, and the list of paths will be sorted according to the target cardinality measure calculated by ESearch.

The BioNavigation interface and the ESearch algorithm are developed in Java and hence is platform independent. Although BioNavigation utilizes external packages for purposes like graph visualization, these are available through open source licenses and are included within the BioNavigation system itself and hence no separate installation is required. The system needs to have the Java Runtime Environment JRE v1.4.2 or greater to be pre-installed on the user's machine. The BioNavigation system is available freely for academic and research purposes and it can be obtained from our website <http://bioinformatics.eas.asu.edu/BioNavigation.html>. The system is easy to install and use and includes an installation guide and user manual.