BioNavigation – Selecting Resources to Evaluate Scientific Queries

Kaushal D. Parekh
CBS Internship Presentation

August 15th, 2005
The Internship

- Advisor – Dr. Zoé Lacroix
  - Scientific Data Management Lab, ASU
    - http://bioinformatics.eas.asu.edu
- Internship duration
  - Spring 2004 to Summer 2005
Introduction

Problems in Scientific Data Collection
Characteristics of Scientific Queries

• Navigational in nature
• Specified in terms of paths through resources
• Examples
  – *From a given gene sequence, return all of functional information available*
    • **BLAST** the sequence, follow the links to **Genbank** then get all functional annotations from there
  – *What genes are involved in a multi-genic neurological disorder?*
    • Search **OMIM** for the disorder and follow the links to other genes
  – *Get citations of articles related to a particular gene*
    • Go to **NCBI Gene** record of that gene and follow links to **PubMed**
Multiple paths match the same query

- gene → citation, has many solutions
  - OMIM → PubMed
  - NCBI Gene → PubMed
    • Two types of links
      - PubMed Links – Articles that involve this gene
      - GeneRIF Links – Annotations submitted by users providing citations that describe the gene function
  - Many other possible options
  - All paths don’t give the same set of results

- Which path is the most suitable?
Complexity of Resources

• Hundreds of Bioinformatics databases
  – Heterogeneous data formats and schemas
  – Curation, data quality and provenance
  – Frequent updates to both content and organization
  – Numerous capabilities provided by data sources – crossref. links, analysis tools, etc.

• Difficult for scientists to keep up with features of each new resource
  – Tend to using only familiar resources
  – Even if aware of a potential alternative
Existing Data Integration Systems

• **DB2 Information Integrator**
  – Allows querying heterogeneous resources through a single SQL query interface
  – Wrappers translate queries and data
  – Provides custom wrapper writing tools

• **SRS**
  – Access multiple bioinformatics resources and tools through single user interface
  – Results and data presented in uniform format
  – Maintains the links in the data to allow for navigational data collection

• **TAMBIS**
  – Queries do not need to specify resources to be used
  – Specify only higher level scientific concepts
  – Databases mapped to these concepts are queried transparently without user intervention
The BioNavigation Approach

Enabling the scientist
Query Formulation

• **Design** queries at a higher level
  – Scientific objects e.g. gene, protein, citation

• Without specifying the *Implementation*
  – e.g. OMIM or NCBI Gene for class ‘gene’

• Design the protocol independent of the characteristics of data sources
  – Not affected by the limitations of resources
  – Intended scientific meaning retained intact
Browsing the Resources

• Visualize the network of available data sources
• Obtain meta-information about each resource
  – e.g. the type of data contained, number of records, schema, url, etc.
• Identify other resources that offer similar capabilities
• Translate high level query to paths at resource level
  – Path = sequence of resources to be visited to evaluate the given query
• Obtain information about all possible alternative paths
• Identify the benefits of using one path over another
Data Collection

- Select a desired path from the list of alternatives
- View metadata information for resources on the path (if required)
- Execute actual queries on resources on the path using a mediator system
Design and Development of the BioNavigation System
Graph Representation

• Bi-Level Representation for resources
• Physical Level
  – Data sources as nodes
  – links as edges
  – Data collection at this level
• Logical or Conceptual level
  – Scientific objects as nodes
  – Relationships between these objects as edges
  – Queries expressed at this level
The BioMetaDatabase

• Provides a map of physical resources and their capabilities
  – e.g. the NCBI resource map

• Stores metadata about these resources to provide users with information
  – Sources: URL, Name, Schema, Identifier etc.
  – Links: Input, Output, URL, etc.
Cardinality Metrics

- In addition to above metadata
- For each data source
  - Cardinality – the total number of records
- For each directional link between two data sources
  - Link Cardinality – Total number of linked pairs
  - Link Image – Number of records having outgoing link(s)
  - Link Participation – Number of records having incoming link(s)
- These metrics will be used to provide an estimate about the paths generated

Cardinality: $S_1 = 4$, $S_2 = 3$
Link Cardinality: $S_1 \rightarrow S_2 = 5$, $S_2 \rightarrow S_1 = 1$
Link Image: $S_1 = 3$, $S_2 = 1$
Link Participation: $S_1 = 1$, $S_2 = 2$
Ontology to represent Conceptual Level

- **What is an Ontology?**
  - Model of important concepts and their relationships specified in an unambiguous language, machine and human readable

- **Applications**
  - AI - Knowledge Representation
  - Semantic Web - assigning meaning to web resources
  - Data Integration - mapping resources to common ontology
  - Controlled Vocabulary - e.g. Gene Ontology
BioNavigation Ontology

• Graph of the conceptual level
  – Maps data sources to classes and links to relationships
• An example,
Query Language

- Queries expressed using the Ontology
- A Navigational Query
  - Sequence of ontological classes and relationships
- Allow traversing unspecified intermediate nodes in the path
- Possible to specify particular resources to be included or excluded in the search
Example

- Get citations to articles that discuss a particular gene
- Get the protein sequence of a gene involved in a particular disease

Gene → Citation (Discussed_In)

Disease → Gene (Involves)

Gene → Protein (Encodes)

Any → Gene → Citation

Any → Protein
Regular Expression Language

- Queries defined by regular expression,
  - \( L(RE) = X (\varepsilon \mid Y X)^* \)
  - \( X = \varepsilon_c \mid c \mid c \ <\text{AnnotList}\>
  - \( Y = \varepsilon_a \mid a \mid a \ <\text{AnnotList}\>
  - \( \varepsilon = \varepsilon_c \varepsilon_a \)
  - where,
    - \( \varepsilon_c, \varepsilon_a = \) “any” or wildcard class or relation
    - \( c, a = \) set of ontological classes and relations respectively
    - \( \text{AnnotList} = \) list of physical resources to be filtered
ESearch Algorithm

- Developed by collaborators
  - Maria-Esther Vidal, Universidad Simon Bolivar, Venezuela
  - Louiqa Raschid, University of Maryland, College Park
- Input: regular expression query with resource annotations
- Process:
  - Breadth First Search (BFS) on the physical graph to identify matching resource paths
  - Search completes in polynomial time if there are no complex loops in the query
- Output: list of physical paths that can be used to evaluate the query
Ranking the Paths

• Different paths give different results
• Three semantic criteria to rank the paths
  – Path cardinality – number of instances of paths of the result
  – Target object cardinality – number of distinct objects retrieved from the final source
  – Evaluation cost – based on local processing cost, path length, remote network access delays, etc.
• These estimates are calculated based on cardinality metrics
• Help the user select a path that suits his needs
The BioNavigation Interface

A Demonstration
Features of the Interface

• **Visualize** the conceptual classes and the corresponding available physical sources
• **Query** integrated resources at the conceptual level
• Obtain a *ranked list* of paths that can be used to evaluate the query
Demonstration
Conclusion

And Future Work
BioNavigation achievements

• Design queries with an ontology independent of the Implementation
• Wildcards to allow users to identify alternate paths that may be exploited
• Physical source annotations to specify resources to be included or excluded
• ESearch algorithm to allow efficient search in the space of all possible evaluation paths
• Provide scientists a way to rank paths
Room for Improvements

• Better graph visualization (in progress)
• Highlighting the top ranked paths in the physical graph
• More meaningful ranking metrics, e.g.,
  – Data quality – curation
  – Trustworthiness – provenance
  – User preferences – favorites
• Ability to select a particular path and run the queries
Integration with SemanticBio

- SemanticBio project at the scientific data management lab
- Build data collection workflows and execute them using web services
- Path selected by a user in BioNavigation can be considered a workflow
- BioNavigation and SemanticBio together could act as a guided querying system
References


Acknowledgements

• This project funded in part by National Science Foundation, Division of Computer and Information Sciences and Engineering
  – Grant IIS-0223042 (Sep 03 – Aug 05)

• Committee members
  – Dr. Zoé Lacroix
  – Dr. Rosie Renaut
  – Dr. Michael Rosenberg