Exploring Heterogeneous Biological Data Sources

Elena Baralis 
elena.baralis@polito.it

Alessandro Fiori 
alessandro.fiori@polito.it

Politecnico di Torino, Dipartimento di Automatica e Informatica
Corso Duca degli Abruzzi 24, 10129, Torino

Abstract

Research activity in the life science area is becoming increasingly data intensive. Huge amounts of highly heterogeneous data, including high throughput experiment results, publication collections, and clinical records are generated at a fast pace by researchers all over the world. The capability of correlating heterogeneous information stored in separated data repositories is a compelling, yet currently unsatisfied, need for bioinformatics scientists.

Recently developed systems address this issue by building knowledge repositories for specific bioinformatics subdomains such as protein-protein interaction or array expression analysis. In this paper we present an overview of heterogeneous biological data sources and discuss the many difficulties faced by biological data querying and analysis.

1. Introduction

New advances in life sciences, for example in the domains of molecular biology and molecular medicine, biodiversity, or new drug discovery and development, increasingly depend on the management and analysis of a vast amount of heterogeneous data. Initially, on demand data integration for inferring new knowledge was manually performed by molecular biologists [9], but this approach has become unfeasible, due to the excessive complexity of the task. Currently, a huge amount of data is generated by high throughput experiments (e.g., complete genome sequences, DNA microarray chips) and contributed by a vast research community spread all over the world. Data is distributed across many sources on the web, is highly heterogeneous, both in structure and semantics, and is characterized by rather diverse levels of quality.

Many current research activities require the capability of jointly analyzing many heterogeneous data sources to extract knowledge not directly obtainable by autonomously exploring isolated data sources. For example, a cornerstone of genomic medicine is the ability to correlate genomic and proteomic experimental data with clinical data, related to the same disease or patient and collected daily in hospitals. This presents daunting informatics challenges such as devising data representations suitable for computational inference (knowledge representation) and capable of linking heterogeneous data sets (data integration) [13].

Recently, this problem has been tackled by proposing different approaches to provide an integrated view of (some) heterogeneous data sources [13, 6, 18]. However, this approach, frequently based on centralizing information in a single integrated repository [6, 18], is hardly able to closely follow the rapid growth and evolution of biological data sources. We believe that a more flexible approach is needed, due to the dynamism and strong heterogeneity which characterizes the life science research domain.

In this paper we present an overview of the main features of biological data sources, proposing a classification based on the available structural information and on the features of the query interface as main dimensions. Current integration efforts are introduced, together with some of their shortcomings. However, the answer to the question “How can we query concurrently all these heterogeneous information repositories, possibly without actually integrating the contained information?” has no definite answer yet and is still a challenging research issue.

2. Biological Data Sources

Life science data (e.g., microarray data) is currently generated at a fast pace by high throughput experiments. Data is stored in a large number of data sources, usually available on the web. Each source is characterized by its own data structure and query interface. Hence, providing an integrated view of (some of) the available information is considered a daunting goal [13], while the capability of seamlessly browsing or querying the available information is a challenging research task.
Many different issues may be considered when classifying biological data sources based on their characteristics. We consider two dimensions, the data structure, which observes the degree of structure associated to the stored data, and the query interface, which addresses the way data is released to external users.

## 2.1. Data Structure

The information characterizing life science processes may be stored in a variety of formats (e.g., XML, relational data), depending on its content. We focus here on the degree of structure provided by the data representation, ranging from unstructured data (i.e., textual data) to strongly structured database data.

**Textual data.** This is the form in which are represented medical abstracts and publications, the most relevant example being PubMed [2]. It is the U.S. National Institutes of Health (NIH) free digital archive of biomedical and life sciences journal literature. Publications in PubMed are typically characterized by a few structured fields (e.g., title, author) and the publication textual information, possibly organized in sections. This repository stores the scientific papers as semi-structured data in XML format. Since the stored information is rather unstructured, searching information in it is rather difficult. Hence, many recent research studies provide ontologies to infer knowledge or provide an infrastructure to browse it more effectively (e.g., iHOP [10]).

**Ontologies.** An ontology provides the representation of a set of concepts and of the relationships holding among the concepts for a particular domain, such as genomic or anatomy [19]. Ontologies specify object classes, relationships and functions [25]. They provide high level abstractions of data, which may facilitate both expressing complex queries over a single source, and querying several heterogeneous sources by exploiting a common set of concepts. The Gene Ontology (GO) project is one of the most popular ontologies in the bioinformatics domain [21]. It has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner [8]. A gene product might be associated with or located in one or more cellular components. It is active in one or more biological processes, during which it performs one or more molecular functions. Each entry in GO has a unique numerical identifier of the form GO: nnnnnn, and a term name (e.g., cell, fibroblast growth factor receptor binding, signal transduction). Each term is also assigned to one of the three ontologies, molecular function, cellular component or biological process.

MeSH (Medical Subject Headings) [14], published by the National Library of Medicine, mainly consists of the controlled vocabulary for bioinformatics research paper. It consists of sets of terms (descriptors) organized in a hierarchical structure that allows searching at various levels of specificity. MeSH descriptors are arranged in both an alphabetic and a hierarchical structure.

**Structured databases.** Other resources store the data in relational databases in order to store explicitly also the relations between the instances of classes. Relational databases provide a robust and query efficient technology to store data. However, because of its limited complex modeling capabilities, relational modeling may lack the flexibility needed to represent complex data types (e.g., nesting and hierarchies are not available). Relational-based repositories typically store a single information type and rely on the query engine of the DBMS for providing form-based data access. Examples of these type of repository are BioGRID [22], UniProt [3] and ArrayExpress [16].

BioGRID [22], Biological General Repository for Interaction Datasets database, was developed to house and distribute collections of protein and genetic interactions from major model organism species. The BioGRID search interface retrieves the results including description of gene/protein function, gene ontology process and molecular functions.

UniProt Knowledgebase consists of two main blocks [26]: (i) UniProtKB/Swiss-Prot which is a curated protein sequence database that provides a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and a high level of integration with other databases; (ii) UniProtKB/TrEMBL which is a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.

ArrayExpress is a public repository for microarray data, which is aimed at storing MIAME-compliant [7] data in accordance with MGED recommendations. The ArrayExpress data warehouse stores gene-indexed expression profiles from a curated subset of experiments in the repository. Public data are made available for browsing and querying on experiment properties, submitter, species and other properties of microarray experiments. All the data are stored in the data warehouse and can be queried on gene, sample, and experiment attributes. Results return graphed gene expression profiles, one graph per experiment.

## 2.2. Query Interface

The sources discussed in the previous section may provide the data through different types of interfaces. Typically, at least one of the following types of interface is provided.

**Keyword search.** This interface is normally provided on
text data collections. The search result is typically a set of documents indexed by the searched keyword(s), over which a very limited form of ranking may be performed. Examples of this interface type are PubMed [2] and BioGRID [22].

**Fixed format query.** Data may be accessed through a form interface, where form fields typically correspond to database attributes. The information provided in form fields is used to define query predicates processed by the underlying query engine. Examples of this interface type are ArrayExpress [16] and GOBASE [15].

**Information broker.** These tools provide a flexible indexing technique or query engine which allows posing more complex queries over the data. The search engine is typically strengthened by the availability of a rich semantic model (e.g., an ontology). The relevant data, which is normally a single data source, may be either stored locally, or referenced through a web link.

Many tools developed in recent years belong to this category [6, 10]. For example, iHOP [10] provides a concept model for networks of concurring genes and proteins, which extends through the scientific literature touching on phenotypes, pathologies and gene function. By exploiting genes and proteins as hyperlinks between sentences and abstracts, the information in PubMed is converted into one navigable resource. This approach provides a tool to enhance the effectiveness in searching research papers on a particular topic (e.g., a protein or a gene).

Finally, some sources, such as BioGRID [22] and ArrayExpress [16], also provide a download interface, by means of which data may be downloaded in an appropriate format for further delocalized data processing.

## 2.3. Source Integration

The issue of providing an integrated view of several heterogeneous sources has been addressed for many years in the database systems research community [20]. Some of the relevant dimensions (e.g., data structure) have already been discussed when dealing with single sources. An important integration issue is where the (integrated) data is stored [9]. The most relevant solutions are the following.

**Warehouse integration.** Data from different sources is consolidated into a local repository, the warehouse, characterized by a global schema integrating all the diverse sources. Queries are posed on the global schema and efficiently executed on the integrated data. In a warehouse, data is periodically updated to align it with source evolution. During this loading process, data usually undergoes some cleaning procedures, which increase the quality of the warehoused data. This strongly integrated architecture, while being an advantage for query performance and providing high quality data, hampers closely following source schema evolution and the addition of new sources. This issue may limit its applicability in very dynamic contexts.

**Mediator based integration.** Data is stored at the sources and integration is obtained through a mediator. Queries are posed on a global (mediated) schema and the mediator provides a seamless translation of the queries on the local schemata of the sources. Different degrees of integration may be achieved, depending on the tightness of the integration of local sources into the global schema, the two extremes being federated databases and peer data management systems [13]. This approach allows an easier integration of new data sources, and simplifies the management of source schema evolution. In this integration architecture, query performance may become an issue. Furthermore, data quality is typically lower, because cleaning procedures cannot be enforced on the fly during query execution.

A frequently adopted approach in bioinformatic data integration is the data warehouse [6, 18]. The approach in [18] integrates data on chip-based experiments, clinical data and gene annotations. Special mapping tables are exploited to relate clinical information with the corresponding chip-based experiments. Biozon [6] is a data warehousing project that stores heterogeneous biological sources. Several types of derived data are stored too, such as similarity relationship between objects (proteins, genes, sequences, etc.) and functional predictions. This derived data expands existing data types based on inference, refines existing objects, and generates new data types generated by processing existing and derived data. Biozon employs a vertical integration approach, where sources are not only incorporated into a single schema but are also integrated using a non-redundant object-centric model.

Otherwise, TAMBIS (Transparent Access to Multiple Bioinformatics Information Sources) [23, 24] is a mediator-based and ontology-driven integration system. Queries in TAMBIS are formulated through a graphical interface where a user browses through concepts defined in a global schema and selects the relevant ones for the particular query. The approach proposed by BioFuice [12] is based on establishing peer mappings. New instances can be easily integrated by mapping them to at least one integrated source (peer).

While current data integration techniques are not going to provide a definitive solution for enabling the seamless exploration of heterogeneous biological sources, the relevance of the data integration problem in the life science research area is demonstrated by the DILS workshop series [5, 4], which address exactly this issue.

## 3. Querying Biological Data Sources

We discuss by means of an example query the various issues arising when a single biological information source is not sufficient for answering a complex biological query.
The main issues are (i) the different notations used by each database, (ii) the consistency of the results returned by the query and (iii) the availability of the result for further post-processing steps. Consider the following (example) query:

Which protein interacts with the alpha subunit of TFIIA in Homo Sapiens?

Protein-protein interactions assemble the molecular machines of the cell and underlie the dynamics of virtually all cellular responses [17]. These interactions are important for many biological functions and their knowledge improves the understanding of diseases and may provide the basis for new therapeutic approaches.

Extracting relevant keywords from natural language. Typically a biological query expresses an information need arising from a biologist, not necessarily skilled in query languages or textual search. Hence, it may be expressed in natural language. This preprocessing step may be performed by an automatic tool or by an expert of the biological domain. The main issue is the detection of the symbol which uniquely represents the protein TFIIA or the gene which encodes it. A textual analysis or a summary of the information stored in the text returned by the query.

Querying a selected source. To identify the interactions with our target protein, a query may be performed on the BioGRID database [22], which stores protein interactions. As a first attempt, the “TFIIA alpha” symbol may be searched for. Unfortunately, query predicates with more than one term are not managed by the query engine. Hence, no result can be returned. When performing the query with the single symbol TFIIA, five results are returned, three of which for the “Drosophila melanogaster” and two for the “Homo sapiens” species. The first result for the Homo sapiens is selected, because it corresponds to the GTF2A1 symbol, which is the gene that encodes the target protein. When selecting this entry, BioGRID returns 17 protein interactions for our target protein. This simple query example highlights that even performing a query on a single database is not a trivial task. In particular, selecting the appropriate notation for the target protein is still an open issue and may strongly affect the query result.

Validating the result. To validate the information extracted from BioGRID, the same query may be performed on other public databases containing similar information. We considered the EMBL-EBI [1] federated database to retrieve all the available information about the target protein. The EMBL-EBI search engine returns the results categorized according to the database which stores the information. The first query is issued with the “TFIIA alpha” symbol. In this case, in the Molecular Interactions category, the query returned zero entries. By issuing the query with the TFIIA symbol, 27 interaction are obtained, while the GTF2A1 symbol returns 1 entry, both in the previous category. It is straightforward to notice that the returned results are rather inconsistent both in the appropriate symbol selection and cardinality of returned results with respect to the former query.

Correlating with textual information. A textual description of the interaction may complement the acquired knowledge. To this end, a search of publications on our target protein in scientific literature repositories may be performed. BioGRID returns some links to PubMed articles in which are described the interactions returned by the query. EMBL-EBI search engine returns also a list of references for the interactions. Unfortunately, the references returned by the previous databases are not consistent. In order to retrieve more references according to the target protein, we have performed the same query on iHOP [10] search engine. iHOP returns more references, but it shows the same notation problem of the previous queries performed on the other sources. Furthermore, iHOP does not provide any tool for a semantic analysis or a summary of the information stored in the text returned by the query.

Availability of results. If further, possibly more elaborated, processing of the query result is to be performed, a utility to download the query result is needed. Only BioGRID provides a download feature to extract query results, while the other queried data repositories not allow any extraction operation and the information is returned in an HTML page with different layouts. Furthermore, it is possible to perform queries only through the web interface. Bulk downloads are only allowed after an explicit authorization is granted.

This simple example shows that retrieving consistent information from different resources is a difficult task, due to the heterogeneity in notation, content of different data sources, and availability of results. Hence, the extraction of
patterns which consistently span over multiple data sources is an interesting challenge in the bioinformatics domain.

4. Conclusions

Research activity in the life science domain is characterized by huge amounts of highly heterogeneous data, dispersed in a large number of sources over the web. This information should be explored by queries and analyses spanning over a variety of biological data sources and allowing a biologist to gather relevant information, to formulate new hypotheses and, possibly, to validate them [9]. Current information system technology may provide some methods and tools to cope with this challenge. However, new and more effective techniques to explore a set of loosely integrated data sources have to be devised. This is the ambitious challenge that life science research poses to the database systems, information retrieval, and data mining communities.

References