Information Integration from Semantically Heterogeneous Biological Data Sources

Doina Caragea[1,4], Jie Bao[1,4], Jyotishman Pathak[1,4], Adrian Silvescu[1,4],
Carson Andorf[1,3,4], Drena Dobbs[2,3,4] and Vasant Honavar[1,2,3,4]

[1]AI Research Laboratory, Department of Computer Science, 226 Atanasoff Hall
[2]Department of Genetics, Development and Cell Biology, 1210 Molecular Biology
[3]Bioinformatics and Computational Biology Program, 2014 Molecular Biology
[4]Computational Intelligence, Learning and Discovery Program, 214 Atanasoff Hall
Iowa State University, Ames, IA 50011
E-mail: honavar@cs.iastate.edu

Abstract

We present the first prototype of INDUS (Intelligent Data Understanding System), a federated, query-centric system for information integration and knowledge acquisition from distributed, semantically heterogeneous data sources that can be viewed (conceptually) as tables. INDUS employs ontologies and inter-ontology mappings, to enable a user to view a collection of such data sources (regardless of location, internal structure and query interfaces) as though they were a collection of tables structured according to an ontology supplied by the user. This allows INDUS to answer user queries against distributed, semantically heterogeneous data sources without the need for a centralized data warehouse or a common global ontology.

1 Introduction

Ongoing transformation of biology from a data-poor science into an increasingly data-rich science, with the attendant increase in the number, size, and diversity of sources of data (e.g., protein sequences, structures, expression patterns, interactions) offer unprecedented, and as yet, largely unrealized opportunities for large-scale collaborative discovery in a number of areas including characterization of macromolecular sequence-structure-function relationships, discovery of complex genetic regulatory networks, etc.

Biological data sources developed by autonomous individuals or groups differ with respect to their ontological commitments, that is, assumptions concerning the *objects* that exist in the *world*, the *properties* or *attributes* of the objects, *relationships* between objects, the possible *values* of attributes, and their *intended meaning*, as well as the *granularity* or *level of abstraction* at which objects and their properties are described [12, 11]. Therefore, *semantic differences* among autonomous data sources are simply unavoidable. Effective use of multiple sources of data in a given context requires reconciliation of such semantic differences, which in fact involves solving a data integration problem.

Driven by the semantic Web vision [2], there have been significant community-wide efforts aimed at the construction of ontologies in life sciences. Examples include the Gene Ontology (www.geneontology.org) in biology and Unified Medical Language System (www.nlm.nih.gov/research/umls) in heath informatics. However, because data sources that are created for use in one context often find use in other contexts or applications (e.g., in collaborative scientific discovery applications involving data-driven construction of classifiers from semantically disparate data sources [4]), and because users often need to analyze data in different contexts from different perspectives, there is no single privileged ontology that can serve all users, or for that matter, even a single user, in every context. Effective use of multiple sources of data in a given context requires flexible approaches to reconciling such semantic differences from the user's point of view.

Against this background, we have investigated a federated, query-centric approach to information integration and knowledge acquisition from distributed, semantically heterogeneous data sources, from a user's perspective.

The choice of the federated, query-centric approach was influenced by the large number and diversity of loosely linked, autonomously maintained data repositories involved and the context and user-specific nature of integration tasks that need to be performed. Our work has led to INDUS, a

Proceedings of the 16th International Workshop on Database and Expert Systems Applications (DEXA'05)
1529-4188/05 $20.00 © 2005 **IEEE**

system for information integration and knowledge acquisition.

We associate ontologies with data sources and users and show how to define mappings between them. We exploit the ontologies and the mappings to develop sound methods for flexibly querying (from a user perspective) multiple semantically heterogeneous distributed data sources in a setting where each data source can be viewed (conceptually) as a single table [5, 4].

The rest of the paper is organized as follows: Section 2 introduces the problem that we are addressing more precisely through an example from biology. Section 3 describes the first prototype of INDUS. We end with conclusions, discussion of related work and directions for future work in Section 4.

2 Motivating Example

The problem that we address is best illustrated by an example. Consider two biological laboratories that independently collect information about protein functions based on the protein sequences. The data collected by the first laboratory contains information about human proteins and their functions (see the entry corresponding to D_1 in Table 1), whereas the data collected by the second laboratory contains information about yeast proteins and their functions (see the entry corresponding to D_2 in Table 1). Suppose that a biologist (user) U wants to assemble a data set based on the two data sources of interest D_1 and D_2 from his or her own perspective. The representative attributes from the user's perspective are *ID, AA composition* (i.e., the number of occurrences of each amino acid in the amino acid sequence corresponding to the protein), and *GO Function* (see the entry corresponding to D_U in Table 1).

However, we observe that the attributes in the data sources D_1 and D_2 are different from the user attributes. In order to reconcile these differences, the user must observe that the attributes *Protein ID* in D_1 and *Accession Number* in D_2 are similar to the user attribute *ID* in D_U; the attributes *Protein Sequence* in D_1 and *AA Sequence* in D_2 are also similar, and they can be used to derive the attribute *AA Composition* in D_U; furthermore, the attributes *EC Number*[1] in D_1 and *MIPS Funcat* [2] in D_2 are similar to the user attribute *GO Function*.

To establish the correspondence between values that two similar attributes can take, we need to associate types with attributes and map the domain of the type of an attribute to the domain of the type of the corresponding attribute (e.g., *AA Sequence* to *AA Composition* or *EC Number* to *GO Function*). We assume that the type of an attribute can be a standard type such as a collection of values (e.g., amino acids, Prosite motifs, etc.), or it can be given by a simple hierarchical ontology (e.g., species taxonomy). Figure 1 shows examples of (simplified) attribute value hierarchies for the attributes *EC Number* in D_1 and *GO Function* in D_U. Examples of semantic correspondences in this case could be: *EC 2.7.1.126* in D_1 is equivalent to *GO 0047696* in D_U, *EC 2.7.1.126* in D_1 is lower than (i.e., hierarchically below) *GO 0004672* in D_U, or for that matter *EC 1.14* is higher than *GO 0004597*, etc.

In general, a biologist might want to answer queries (e.g., *proteins that are involved in catalytic activity* or *the number of human proteins that are involved in kinase activity*) from the integrated data. INDUS, the system that we develop in our lab, can be used to answer such queries against distributed, semantically heterogeneous data sources without the need for a centralized data warehouse or a common global ontology. We will describe INDUS in more detail in the next section.

3 Prototype Implementation of INDUS

The current prototype of INDUS enables a biologist with some familiarity with the relevant data sources to integrate and analyze relevant data sources by specifying a user ontology, simple mappings between data source specific ontologies, and executing queries - all without having to write code. The current implementation of INDUS includes support for:

- Import, adaptation and reuse of selected fragments of existing ontologies (e.g., Gene Ontology GO), editing of ontologies, specification of semantic relationships between ontologies using inter-ontology mappings [1].
- Specification of semantic correspondences between a user ontology and data source ontologies [4, 3]. Semantic correspondences between ontologies can be defined at two levels: schema level (between attributes that define data source schemas) and attribute level (between values of attributes). INDUS allows the following types of semantic correspondences at both schema and attribute level: semantic equality (e.g., $AASequence : O_1 \equiv ProteinSequence : O_U$), semantic subsumption (e.g., $MIPS : 16.19.01 : O_1 \leq GO : 0017076 : O_U$), and procedural mappings (e.g., from $AASequence : O_1$ to $AAComposition : O_U$ at attribute level).
- Registration of a new data source (using a data-source editor for defining the schema of the data source by specifying the names of the attributes and their corresponding ontological types, location, type of the data source and access procedures that can be used to interact with a data source as though it were a table structured according to its schema and the ontology. In the

[1] Enzyme Commission Number, http://www.chem.qmul.ac.uk/iubmb/enzyme/
[2] Munich Information Center for Protein Sequences, http://mips.gsf.de/

Table 1. Data sets D_1, D_2 and user data D_U.

Data	*Protein ID*	*Protein Name*	*Protein Sequence*	*Prosite Motifs*	*EC Number*
D1	P35626	Beta-adrenergic receptor kinase 2	MADLEAVLAD VSYLMAMEKS ...	RGS $PROT\_KIN\_DOM$ PH_DOMAIN	2.7.1.126 Beta-adrenergic receptor kinase
	Q12797	Aspartyl/asparaginyl beta-hydroxylase	MAQRKNAKSS GNSSSSGSGS ...	TPR TPR_REGION TRP	1.14.11.16 Peptide-aspartate beta-dioxygenase

Data	*Acc. Num.*	*Gene ID*	*AA Sequence*	*Length*	*Pfam Domains*	*MIPS Funcat*
D_2	P32589	SSE1	STPFGLDLGN NNSVLAVARN ...	692	$HSP70$	16.01 protein binding
	P07278	BCY1	VSSLPKESQA ELQLFQNEIN ...	415	$cNMP\_binding$ *RIIa*	16.19.01 cyclic nucleotide binding (cAMP, cGMP, etc.)

User	*ID*	*AA composition*	*GO Function*
D_U	P35626	7 3 9 14 ...	0047696:beta-adrenergic-receptor kinase activity
	Q12797	5 1 7 12 ...	0004597: peptide-aspartate beta-dioxygenase activity
	P07278	10 8 6 15 ...	0009408: Biological process- response to heat

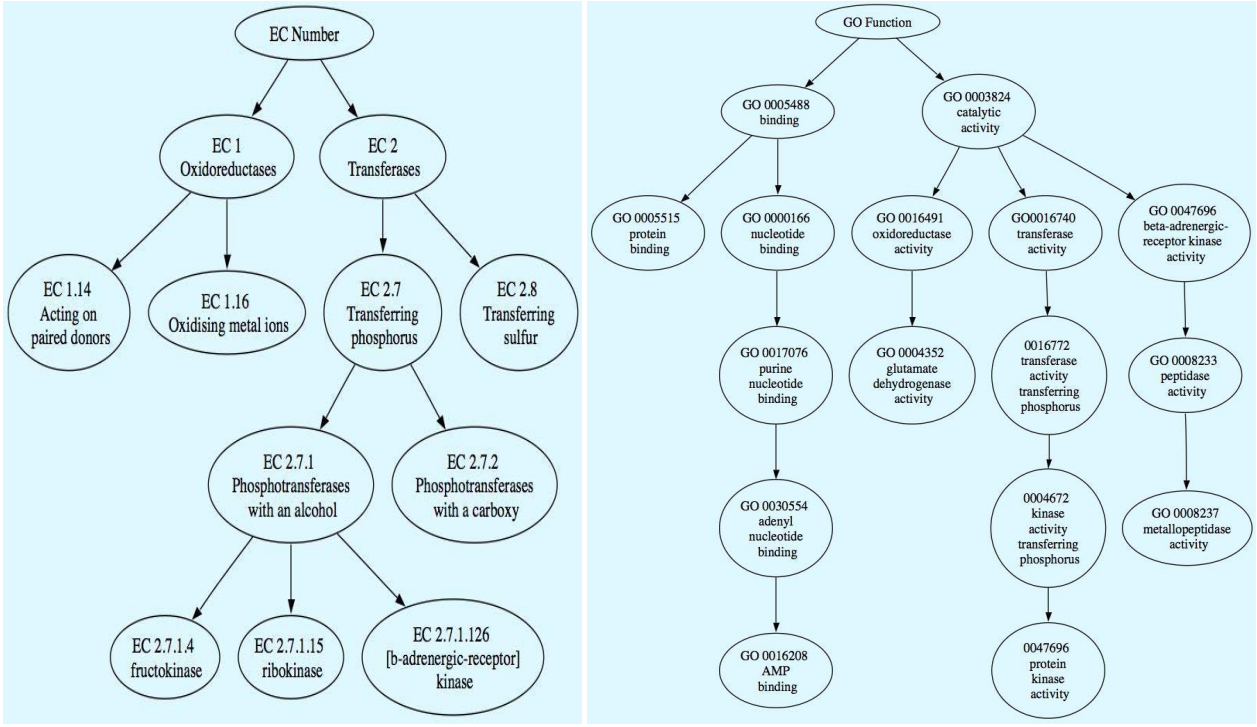

Figure 1. Hierarchies associated with the attributes *EC number* in D_1 and *Go Function* in D_U.

current implementation several types of data sources can be defined including multiple relational databases (Oracle, MySQL, PostgreSQL), and files (e.g., ARFF files used in WEKA, a widely used open source machine learning software package). Work in progress is aimed at the design and implementation of extensions that allow definition of complex views that allow execution of complex statistical queries against sequence, structure, expression, and interaction databases based on multiple ontologies as well as inter-ontology mappings.

- Specification and execution of queries across multiple large, semantically heterogeneous data sources with different interfaces, functionalities and access restrictions. Each user may choose data sources of interest to him/her from a list of data sources that have been previously registered with the system, specify a user ontology (by selecting an ontology from a list of available ontologies or by invoking the ontology editor and defining a new ontology). Once the ontology-extended data sources and the user ontology have been specified, the user can select mappings between data source ontologies and user ontology from the available set of existent mappings (or invoke the mappings editor to define a new set of mappings). After all the mappings are specified, the system can be used to answer queries posed by the user. The data needed for answering a query is specified by selecting (and possibly restricting) attributes from the user ontology, through a friendly interface. Queries posed by the user are sent to a query-answering engine (QAE) that decomposes a user query into sub-queries that can be answered by the individual data sources (using predefined or user-supplied mappings between the respective ontologies). The results of the partial queries answered by the distributed data sources are sent back to the QAE which composes them to generate the answer to the user query (expressed in terms of user ontology) and presents it to the user.

Note that in the current release of the INDUS software, we have assembled two relational databases which contain a subsets of the information gathered from SWISSPROT and MIPS to demonstrate how the user can query the two databases flexibly using user-supplied mappings.

4 Summary and Discussion

Summary: We present the first prototype of INDUS, a federated, query-centric approach to answering user queries from distributed, semantically heterogeneous data sources. INDUS assumes a clear separation between data and the semantics of the data (ontologies) and allows users to specify ontologies and mappings between data source ontologies and user ontology. These mappings are stored in a mappings repository to ensure their re-usability and are made available to a query answering engine. The task of the query answering engine is to decompose a query posed by a user into subqueries according to the distributed data sources and compose the results into a final result to the intial user query. An initial version of INDUS software and documentation are available at: `http://www.cild.iastate.edu/software/indus.html`.

Discussion: There is a large body of literature on information integration and systems for information integration. Davidson et al. [7] and Eckman [8] survey alternative approaches to data integration. Hull [14] summarizes theoretical work on data integration. Several systems have been designed specifically for the integration of biological data sources. It is worth mentioning SRS [10], K2 [19], Kleisli [6], IBM's DiscoveryLink [13], TAMBIS [18], OPM [15], BioMediator [17], among others.

Systems such as SRS and Kleisli do not assume any data model (or schema). It is the user's responsability to specify the integration details and the data source locations, when posing queries. Discovery Link and OMP rely on schema mappings and the definition of views to perform the integration task. TAMBIS and BioMediator make a clear distinction between data and the semantics of the data (i.e., ontologies) and take into account semantic correspondences between ontologies (both at schema level and attribute level) in the process of data integration.

Most of the above mentioned systems assume a predefined global schema (e.g., Discovery Link, OMP) or ontology (e.g., TAMBIS), with the notable exception of BioMediator, where users can easily tailor the integrating ontology to their own needs. This is highly desirable in a scientific discovery setting where users need the flexibility to specify their own ontologies.

While some of these systems can answer very complex queries (e.g., BioMediator), others have limited query capabilities (e.g, SRS which is mainly an information retrieval system). Furthermore, for some systems it is very easy to add new data sources to the system (e.g., SRS or Kleisli, where new data source wrappers can be easily developed), while this is not easy for other biological integration systems (e.g., Discovery Link or OMP, where the global schema needs to be reconstructed).

On a different note, there has been a great deal of work on ontology development environments. Before developing INDUS editor, off-the-shelf alternatives such as IBM's Clio [9] or Protege [16] were considered, but they proved insufficient for our needs. Clio provides support only for schema mapping, but not for hierarchical ontology mapping. Pro-

tege is a purely knowledge base constructing tool (including ontology mappings). It does not provide support for the association of ontologies with data, data management or queries over the data. Furthermore, neither of these systems allow procedural mappings (a.k.a., conversion functions), which are essential for data integration.

Work in progress is aimed at:

- Integrating some machine learning algorithms with INDUS. This would enable collaborative construction of predictive models or classifiers without having to first construct a data set [5, 4, 3].
- Development of data-source specific data retrieval procedures (iterators) for several commonly used data sources in bioinformatics.
- Development of support for handling modular ontologies including support for integration and reuse of ontologies, some basic inference procedures.
- Development of a conceptual framework for exploiting self-describing web services (much along the lines of ontology-extended data sources).
- Implementation of techniques for optimization of query execution across multiple data sources to minimize data transfer and computational overhead subject to constraints imposed by the individual data sources.
- Support for sharing of data, analysis results, and programs securely between users, groups of users, or the world, across the Internet.
- Documentation and dissemination of INDUS software, along with sample ontologies, inter-ontology mappings, data source descriptions, to the user community and further refinement of the software based on user feedback.

Acknowledgements: This work was funded in part by grants from the National Science Foundation (IIS 0219699) and the National Institutes of Health (GM 066387).

References

[1] J. Bao and V. Honavar. Collaborative ontology building with wiki@nt - a multi-agent based ontology building environment. In *Proceedings of the Third International Workshop on Evaluation of Ontology based Tools, at the Third International Semantic Web Conference ISWC*, Japan, 2004.

[2] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, May 2001.

[3] D. Caragea, J. Pathak, J. Bao, A. Silvescu, C. Andorf, D. Dobbs, and V. Honavar. Information integration and knowledge acquisition from semantically heterogeneous biological data sources. In *2nd International Workshop on Data Integration in the Life Sciences*, 2005.

[4] D. Caragea, J. Pathak, and V. Honavar. Learning classifiers from semantically heterogeneous data. In *Proceedings of the International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems*, 2004.

[5] D. Caragea, A. Silvescu, and V. Honavar. A framework for learning from distributed data using sufficient statistics and its application to learning decision trees. *International Journal of Hybrid Intelligent Systems*, 1(2), 2004.

[6] J. Chen, S. Chung, and L. Wong. The Kleisli query system as a backbone for bioinformatics data integration and analisis. *Bioinformatics*, pages 147–188, 2003.

[7] S. Davidson, J. Crabtree, B. Brunk, J. Schug, V. Tannen, G. Overton, and C. Stoeckert. K2/Kleisli and GUS: experiments in integrated access to genomic data sources. *IBM Journal*, 40(2), 2001.

[8] B. Eckman. A practitioner's guide to data management and data integration in bioinformatics. *Bioinformatics*, pages 3–74, 2003.

[9] B. Eckman, M. Hernndez, H. Ho, F. Naumann, and L. Popa. Schema mapping and data integration with clio (demo and poster). In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology (ISMB 2002)*, Edmonton, Canada, 2002.

[10] T. Etzold, H. Harris, and S. Beulah. SRS: An integration platform for databanks and analysis tools in bioinformatics. *Bioinformatics Managing Scientific Data*, 2003.

[11] R. Fikes, A. Farquhar, and J. Rice. Tools for assembling modular ontologies. In *The Fourteenth National Conference on Artificial Intelligence*, 1997.

[12] T. Gruber. Ontolingua: A mechanism to support portable ontologies. Technical Report KSL91-66, Stanford University, Knowledge Systems Laboratory, 1991.

[13] L. Haas, P. Schwarz, P. Kodali, E. Kotlar, J. Rice, and W. Swope. DiscoveryLink: a system for integrated access to life sciences data sources. *IBM System Journal*, 40(2), 2001.

[14] R. Hull. Managing semantic heterogeneity in databases: A theoretical perspective. In *PODS*, Tucson, Arizona, 1997.

[15] A. Kosky, I. Chen, V. Markowitz, , and E. Szeto. Exploring heterogeneous biological databases: Tools and applications. In *Proceedings of the 6th International Conference on Extending Database Technology (EDBT98), Lecture Notes in Computer Science Vol. 1377, Springer-Verlag*, 1998.

[16] N. F. Noy, R. W. Fergerson, and M. A. Musen. The knowledge model of protege-2000: Combining interoperability and flexibility. In *Second International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000)*, Juan-les-Pins, France, 2000.

[17] 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 *Proceedings of the Workshop on Information Integration on the Web (held in conjunction with VLDB 2004)*, Toronto, ON, 2004.

[18] R. Stevens, C. Goble, N. Paton, S. Becchofer, G. Ng, P. Baker, and A. Bass. Complex query formulation over diverse sources in tambis. *Bioinformatics*, 2003.

[19] V. Tannen, S. Davidson, and S. Harker. The information integration in K2. *Bioinformatics*, pages 225–248, 2003.