

Algorithms and Software for Information Extraction, Integration, and Data-Driven Knowledge Acquisition From Heterogeneous, Distributed, Autonomous Biological Information Sources

Honavar, Vasant G.^{*1,2,3,4}, **Dobbs, Drena L.**^{3,4,5}, **Jernigan R.L.**^{3,4,6}, **Caragea, D.**^{1,2,3}, **Reinoso-Castillo, J.**^{1,2,3}, **Silvescu, A.**^{1,2,3}, **Pathak, J.**^{1,2,3}, **Andorf, C.**^{1,2,3}, **Yan, C.**^{1,2,3,4}, **Zhang, J.**^{1,2,3}

¹Artificial Intelligence Research Laboratory; ²Department of Computer Science; ³Bioinformatics and Computational Biology Graduate Program; ⁴Laurence H. Baker Center for Bioinformatics and Biological Statistics; ⁵Department of Molecular, Cell, and Developmental Biology; ⁶Department of Biochemistry, Biophysics, and Molecular Biology; Iowa State University, Ames, Iowa, USA

The effective use of increasing amounts of data from disparate information sources to explore specific scientific questions (e.g., characterization of macromolecular sequence-structure-function relationships) specific presents several challenges in bioinformatics:

- a. Data repositories of interest in computational molecular biology are large in size, dynamic, and physically distributed. Consequently, it is neither desirable nor feasible to gather all of the data in a centralized location for analysis. Hence, there is a need for algorithms that can efficiently extract the relevant information (e.g., statistics needed by data mining algorithms in the context of specific data-driven knowledge acquisition tasks e.g., exploration of macromolecular sequence-structure-function relationships) from disparate sources.
- b. Data sources of interest are autonomously owned and operated. Consequently, the range of operations that can be performed on the data source (e.g., the types of queries allowed), and the precise mode of allowed interactions can be quite diverse. Hence, strategies for obtaining the necessary information (e.g., statistics needed by data mining algorithms) within the operational constraints imposed by the data source are needed.
- c. Data sources are heterogeneous in structure (e.g., relational databases, flat files) and content. Each data source implicitly or explicitly uses its own ontology (concepts, attributes and relations among attributes) to represent data. For example, the gene ontology (www.geneontology.org) project is aimed at the development of several ontologies and their XML encodings for use in Bioinformatics. Thus, approaches to effective integration of information from different sources bridging the syntactic and semantic mismatches among the data sources are needed.
- d. In scientific discovery, because users often need to examine data in *different contexts from different perspectives*, there is no single universal ontology that can serve all users, or for that matter, even a single user, in every context. Hence, methods for context-dependent dynamic information extraction and integration from distributed data based on user-specified ontologies are needed to support knowledge acquisition from heterogeneous distributed data.

INDUS (Intelligent Data Understanding Environment) (<http://www.cs.iastate.edu/~honavar/indus.html>), being developed in our laboratory includes some of the key elements of principled approaches to addressing some of these challenges or user, task, and context-specific information extraction and knowledge acquisition (using machine learning). INDUS implements an ontology-driven, query-centric, approach to data integration that allows users to impose their own semantics (or points of view) on disparate data sources to efficiently acquire knowledge (e.g., in the form of classifiers) from distributed data. Representative knowledge acquisition tasks that arise in exploration of macro-molecular sequence-structure-function relationships are being used to evaluate and refine the current prototype of INDUS.

This work is supported in part by grants (0219699, 09972653) from the National Science Foundation, a Biological Information Science and Technology Initiative (BISTI) award (GM066387) from the National Institutes of Health, and graduate fellowships from Pioneer Hi-Bred and IBM.