Integration of distributed heterogeneous biomolecular data to support biological discovery

Masseroli M, Ceri S, Tettamanti L, Campi A, Sormani S

Dipartimento di Elettronica e Informazione, Politecnico di Milano,
Piazza Leonardo da Vinci 32, 20133 Milano, Italy

Motivation

Publicly accessible molecular biology databanks are continuously increasing in number (more than 1150 in January 2009) and in coverage of the included biomolecular entities (e.g., genes, transcripts, proteins), as well as of their described structural and functional biomedical features (e.g., expression in different tissues; involvement in biological processes and biochemical pathways). Such databanks provide extremely valuable information to support the interpretation of experimental results (e.g., high-throughput lists of candidate relevant genes or proteins) and infer new biomedical knowledge. Yet, the information about a given biomolecular entity is often scattered across many different databanks. As many questions in bioinformatics can be addressed only by comprehensively evaluating different types of data, combining information from multiple databanks is paramount. Furthermore, since different biomolecular databanks often contain redundant or overlapping information, integration of data from different databanks allows cross-validation of the provided information in order to identify redundant and mismatching information. Different approaches, including mediator-based solutions, have been proposed to integrate data from multiple sources. Yet, data warehousing seams to be the most adequate when the data to be integrated are very numerous and off-line processing is required to efficiently and comprehensively mine the integrated data. This approach requires that information from the distributed databanks to be integrated are automatically retrieved and processed in order to create and easily maintain updated an integrated, consistent and easily to extend data collection able to effectively support high-throughput bioinformatics analyses. Existing examples of genomic data warehouse design and implementation present, as a major limitation, a complex data warehouse schema difficult to be maintained and extended with additional data types. We designed and implemented a generalized genomic and proteomic data warehouse schema and a software to easily create, extend and automatically update a data warehouse that integrates annotations from different biomolecular databanks, guarantees the quality of the integrated annotations and structures them in a suitable way to be used for high-throughput data driven biological discoveries.

Methods

In order to design our generalized data warehouse schema, we analyzed the types of data provided by several different of the most relevant biomolecular databanks publicly accessible, and abstracted a general global schema that can incorporate the schema that the data to be integrated in the data warehouse have in their source databank. The abstracted schema is composed of two distinct tiers: 1) a data source tier that includes source specific data type schemas, where data from each source databank are imported; 2) an integrated data tier that, for each considered biomolecular entity (i.e., genomic DNA, gene, transcript, protein) and biomedical feature (e.g., biochemical pathway, biological process, molecular function, cellular component), holds references to the data imported in the data source tier from the source databanks and their common most relevant information. The software to easily create, extend and automatically update the data warehouse was designed by abstracting and generalizing the processing steps required to import and integrate data in different formats from different biomolecular databanks in a data warehouse, and it has been implemented in Java programming language.

Results

A prototypical genomic and proteomic data warehouse, which implements our generalized schema by using a PostgreSQL relational DBMS, has been created and is maintained automatically updated monthly by our implemented software. It currently integrates data from 9 databanks (Entrez Gene, Homologene, UniProt/Swiss-Prot, eVOC, Gene Ontology, GOA, BioCyc, KEGG, Reactome), including 4,451,160 genes (4,184,605 protein-coding) of 5,354 different organisms (including 25,152 human protein-coding genes), 305,905 proteins of 1,038 species (including 46,873 human proteins), 26,166 Gene Ontology (GO) terms, 808 different pathways, and several eVOC ontology terms describing the expression of genes (517 anatomical system,

192 cellular type, 157 developmental stage, and 199 pathology terms). By querying the integrated data, unexpected information patterns possibly leading to data driven biological discoveries, as well as inconsistencies in information provided by different databanks, can be unveiled. As an example, by checking cross references existing between Gene Ontology, UniProt and Entrez Gene databanks, we found that 321 (1.83%) GO annotations (regarding 197 different GO terms) of 81 human proteins provided by GOA were not comprised in the GO annotations of the codifying genes provided by Entrez Gene databank, including also 143 (44.56%) annotations with evidence stronger than that Inferred from Electronic Annotation.

Contact : masseroli@elet.polimi.it