Integration and Querying of Genomic and Proteomic Semantic Annotations for Biomedical Knowledge Extraction

Marco Masseroli, Arif Canakoglu, and Stefano Ceri

Abstract—Understanding complex biological phenomena involves answering complex biomedical questions on multiple biomolecular information simultaneously, which are expressed through multiple genomic and proteomic semantic annotations scattered in many distributed and heterogeneous data sources; such heterogeneity and dispersion hamper the biologists' ability of asking global queries and performing global evaluations. To overcome this problem, we developed a software architecture to create and maintain a Genomic and Proteomic Knowledge Base (GPKB), which integrates several of the most relevant sources of such dispersed information (including Entrez Gene, UniProt, IntAct, Expasy Enzyme, GO, GOA, BioCyc, KEGG, Reactome and OMIM). Our solution is general, as it uses a flexible, modular and multilevel global data schema based on abstraction and generalization of integrated data features, and a set of automatic procedures for easing data integration and maintenance, also when the integrated data sources evolve in data content, structure and number. These procedures also assure consistency, quality and provenance tracking of all integrated data, and perform the semantic closure of the hierarchical relationships of the integrated biomedical ontologies. At http://www.bioinformatics.deib.polimi.it/GPKB/, a Web interface allows graphical easy composition of queries, although complex, on the knowledge base, supporting also semantic query expansion and comprehensive explorative search of the integrated data to better sustain biomedical knowledge extraction.

Index Terms—Management and integration of heterogeneous and distributed biological data, Biological ontologies, Querying and retrieval of semantic biological annotations, Mining of semantically annotated biological data

1 Introduction

Increasingly large amounts of valuable, but heterogeneous and sparse, biomolecular data and information are characterizing life sciences [1]. In particular, semantic controlled annotations of biomolecular entities, i.e. the associations between biomolecular entities (mainly genes and their protein products) and controlled terms that describe the biomolecular entity features or functions, are of great value; they support scientists with several terminologies and ontologies describing structural, functional and phenotypic biological features of such entities (e.g. their sequence polymorphisms, expression in different tissues, or involvement in biological processes, biochemical pathways and genetic disorders).

These semantic annotations can effectively support the interpretation of genomics and proteomics test results and the extraction of biomolecular information, which can be used to formulate and validate biological hypotheses and possibly discover new biomedical knowledge. A comprehensive approach to such data integration, querying and analysis can help understanding complex biological processes and their pathological alterations, by answering related complex biomedical questions. Yet, the scattering of genomic and proteomic annotation data in many complementary but also overlapping sources is an important and not yet completely solved challenge. Specifically, data source heterogeneity in data representation and format, their fast evolution in number, data content and structure, the high variety of available data types, and also the great amount of data produced over time, are the facets of a very hard data integration problem [2]-[4].

Taking advantage of our previous experience with the GFINDer system [5], [6], we developed a software architecture to create and maintain an updated and publicly available integrative data warehouse of genomic and proteomic semantic annotations. It adopts a modular and multilevel global schema that we propose for integrated data management. This data schema supports integration of data sources, possibly overlapping, which are fast evolving in data content, structure and number, and assures provenance tracking of all the integrated data.

From an engineering perspective, our solution is based on several innovative principles: (i) modular and multilevel, domain-independent flexible schema for data integration, (ii) clear separation of source-specific data import from source-independent data integration, (iii) use of reflection in Java programming language to implement self-configuring parsers, (iv) usage of historical information to deal with changes in the data sources and with differences in the updating time of each data item, and (v) ontological processing, based upon semantic closure and search of lowest common ancestor between ontology terms, to support efficient semantic queries and evaluations. We also developed a user-friendly Web interface that supports the easy composition of complex multi-topic queries and their semantic expansion upon the integrated data; such interface fully enables users to compre-

M. Masseroli, A. Canakoglu and S. Ceri are with the Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milano, IT 20133 Italy. (phone: +39-02-2399-3553; E-mail: [masseroli\canakoglu\ceri]@elet.polimi.it ; fax: +39-02-2399-3411).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2453944, IEEE/ACM Transactions on Computational Biology and Bioinformatics

hensively select, extract and display all data of their interest that match, syntactically or semantically, the performed query and to take advantage of them for biomedical hypotheses formulation and knowledge discovery.

The outline of this paper is as follows. Section 2 discusses the related work in data integration, focusing on the biomedical domain. Section 3 describes our integrated data schema. Section 4 illustrates the developed software architecture for data integration, which ensures consistency, quality and provenance tracking of all the integrated data, eases their updating and extension and perform semantic closure of the integrated ontology hierarchical relationships. Section 5 presents the Genomic and Proteomic Knowledge Base (GPKB), which benefits from our integrated data schema and software architecture and provides Web interfaces to easily compose queries, although complex, on the integrated semantic data. Section 6 illustrates a relevant example of GPKB use for discovering common biological aspects in apparently unrelated genetic disorders. Section 7 discusses significant aspects of our work and concludes.

2 Related Work

Several approaches and systems have been proposed to integrate data from multiple heterogeneous data sources. They include *information linkage* (e.g. SRS [7], NCBI Entrez [8]), *multi-databases* (e.g. TAMBIS [9], BACIIS [10]), *federated databases* (e.g. BioKleisli [11], BioMart [12]), *mediator based* (e.g. BioDataServer [13], Biomediator [14]) or *workflow based* (e.g. Taverna [15], Galaxy [16]) *solutions* and *data warehousing* (e.g. BioWarehouse [17], Biozon [18]).

The last approach is well-known to have maintenance overhead, both in keeping the integrated data up-to-date with the original selected sources and in expanding the warehouse with additional data and data types from new sources [19]. Yet, the data warehousing approach is superior in supporting applications that require off-line data processing in order to efficiently organize the integrated data and comprehensively use them for knowledge discovery, which is our goal. Furthermore, it allows thoroughly checking data quality and consistency, within a single or across multiple data sources, in order to integrate and use only high quality consistent data. It also easily allows reconciling unsynchronized data, e.g. from distinct data sources with different updating times, by taking advantage of available historical evolution data during warehouse construction. For all these reasons, we adopted the data warehousing approach.

Other data integration solutions can easily offer updated data with a more limited maintenance overhead; among them, federated and mediator or workflow based approaches capitalize available Web services to directly access data in their original data sources. Yet, they too require maintenance, since available Web services may change their data structure overtime. Moreover, such approaches require performing all data transfer and processing online at the time of the user request; thus, they make it difficult and slow, if not impossible, performing thorough data quality and consistency checking and reconciliation of unsynchronized data.

On the other hand, warehousing drawbacks of maintenance overhead can be specifically tackled and drastically reduced by using automatic procedures to regularly update easily the data in the warehouse [19]. Such automatic procedures are particularly important when the data warehouse integrates data from biomolecular databases, since usually such databases are updated frequently.

Furthermore, difficulties in expanding the data warehouse with additional data sources mainly arise from the integrated data schema adopted. The data models proposed for biological data [5], [20], [21] are generally very expressive and complex, as they embody a lot of domain knowledge, but expressive descriptions carry a cost, making it difficult to face the integration challenges of evolving data[1]. Furthermore, the integrated biological data models proposed usually do not provide good support for data provenance and version tracking, as well as for integration of different and overlapping sources providing the same data type [2]-[4].

To overcome these issues, we developed and adopted a modular and multilevel feature-based integrated data schema, which is described in Section 3. It not only eases data warehousing updates and extensions, but also ensures provenance tracking of all the integrated data.

Data warehousing is a well-known approach also for data analytics though multidimensional data aggregation; however, as previous warehousing proposals, e.g. [17], in our approach we did not take advantage of this aspect, since the biomolecular semantic annotation data on which we focused rarely include additive attributes.

Recently, a lot of emphasis has been placed on the use of linked data for biological information [23]; yet, linked data provide only binary connections between pairs of source items and querying them still remains difficult due to the lack of uniformity in the representation of linked data datasets [24]. Furthermore, intuitive interfaces for querying biological linked data and using extracted results are still very limited. Conversely, although focused on some selected sources, our approach integrates and mediates data items extracted from multiple sources, with overall greater data quality and seamless querying and result usage support.

3 Integrated Data Schema

In this Section we illustrate and discuss the global data schema that we defined to integrate numerous, heterogeneous, controlled annotation data, i.e. data regarding different features or topics represented through multiple controlled vocabularies and ontologies, as well as their associations.

3.1 Feature Modules

Our integrated data schema is composed of multiple interconnected modules; each module represents a single feature, whose data are provided by one or more of the

[1] Yet, integration of additional data sources can improve integrated data coverage and quality, through identification of mismatching information by cross-verification of overlapping data [22].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2453944, IEEE/ACM Transactions on Computational Biology and Bioinformatics

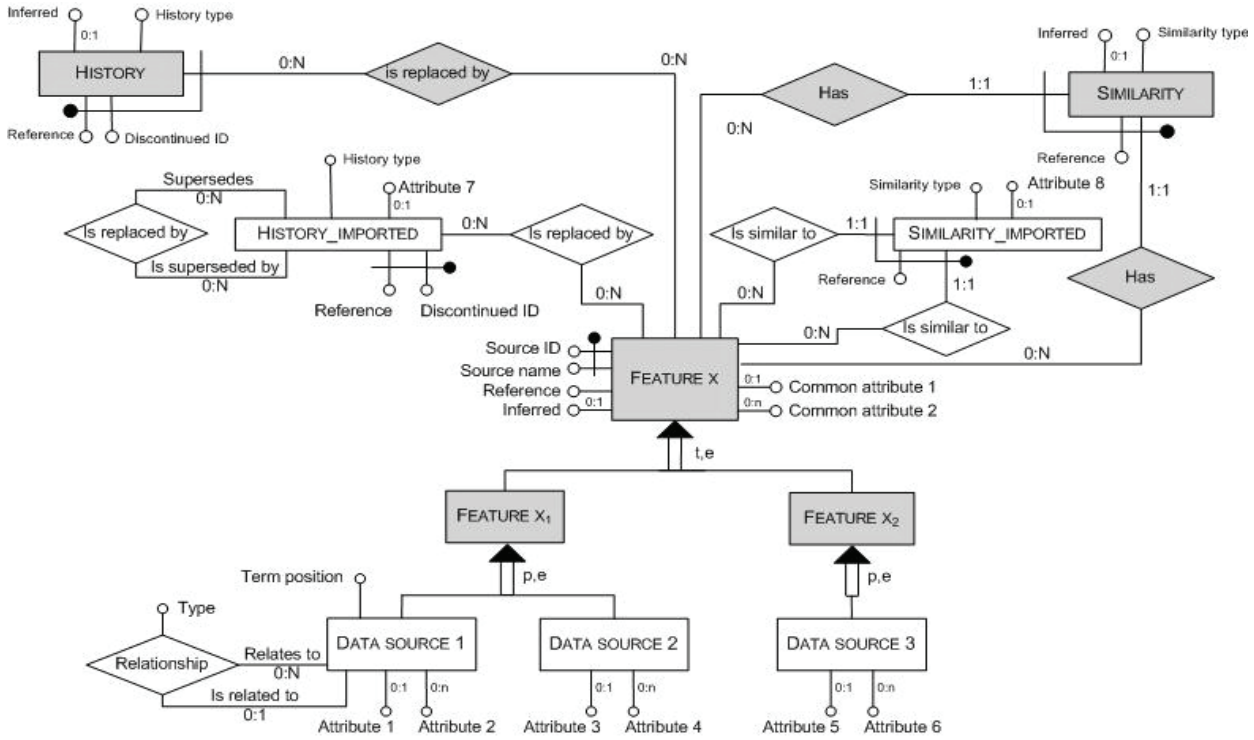

Fig. 1. Multilevel feature module of the defined integrated data schema. White shapes represent import tier data, while darker shapes represent aggregation tier data. FEATURE X_1 and FEATURE X_2 are different subtypes of FEATURE X; DATA SOURCE 1 and DATA SOURCE 2 are distinct sources that provide data of the same subtype of FEATURE X, the former provides ontology data and their relationships.

integrated data sources, and contains provenance information for each single feature instance entry (Figure 1). As we focus on controlled biomedical-molecular annotation data, a feature can be a biomolecular entity (i.e. DNA sequence, gene, transcript, protein) or a biomedical feature (e.g. pathway, genetic disorder, etc.), a feature instance can be, for example, a specific gene, protein, pathway or genetic disorder, and a feature entry is a specific representation of a feature instance (e.g. the data of a specific gene in a particular data source such as the Entrez Gene database). Each feature entry is identified by the value of its *Source ID* and *Source name* attributes (since each feature instance can have multiple IDs from different sources), and contains the *Reference* attribute, representing the source that provided the data, i.e. their provenance (which can be different from their ID source); for example, Table 1 shows two Gene feature entries, from distinct sources, that represent the Gene feature instance of the human BRCA1 gene.

TABLE 1
EXAMPLE OF GENE FEATURE ENTRIES AND MAIN ATTRIBUTES

Source ID	Source name	Reference	Symbol	Name	Taxonomy ID
672	Entrez Gene	Entrez Gene	BRCA1	Breast cancer 1, early onset	Homo sapiens
113705	OMIM	OMIM	BRCA1	Breast cancer 1 gene	Homo sapiens

More detailed additional provenance aspects are represented by specific attributes, depending on the information provided by the original data source. The *Inferred* attribute describes if and how a feature entry has been inferred from other data. The *Reference* and *Inferred* attributes, together with other specific attributes in each source entity, allow provenance transparent tracking of each data. This is a fundamental aspect to enable users to assess their confidence in the data. Furthermore, every biomolecular entity instance is characterized by its *Symbol* and the *Taxonomy ID* of the organism to which it belongs. Similarly, every biomedical feature instance is characterized by its *Name* and *Definition*.

Each feature module of our general schema can also include *History* and/or *Similarity* data. The former ones represent obsolete discontinued source IDs and, if they have been propagated, the current ID that replaced each of them. The latter ones describe equivalence of different feature entries (from the same or different sources) through their ID pairings, which link the different feature entries that the IDs identify (e.g. gene feature entries identified by different IDs from distinct sources, but representing the same gene; see Table 1 for an example). These similarity data can be imported from one or more sources, or inserted by expert curators, or inferred automatically by a computational process (e.g. based on Natural Language Processing of textual descriptions available within some data attributes, or based on semantic analysis of ontological data). Both history and similarity data are paramount to reconcile unsynchronized data and identify multiple feature instance entries, from single or multiple data sources, as representing the same feature instance

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2453944, IEEE/ACM Transactions on Computational Biology and Bioinformatics

(e.g. data regarding different gene IDs which actually represent the same single gene; for example, in the case of history data, this can occur when the gene ID changes and the gene data are provided by different sources, among the integrated ones, with different updating time, so that in some of them the gene ID has not yet been updated). Thus, history and similarity data directly enable the accurate mapping and integration of different feature data.

Finally, our data schema can represent controlled descriptions of feature instances expressed through either a flat terminology or an ontology (e.g. pathways described through the BioCyc Pathway controlled vocabulary, or biological function features described through Gene Ontology terms). In the latter case, ontological relationships among hierarchically related feature data from the same data source are represented in the schema by a *Relationship* auto-association of the data source entity (Figure 1), where the *Type* attribute represents the different types of semantic relationships that the source data describe.

In order to ease maintenance and extension of the integrated data schema defined, each feature module is internally organized in two levels: an *import tier* and an *aggregation tier*. The import tier allows structuring and locating together originally distributed data, while thoroughly checking their consistency and quality [22], as well as identifying the feature they refer to and their main attributes, and associating each feature entry with a unique OID. The import tier is composed of separated subschemas, each one for every single data source considered which provides data for that feature, individually structured as in the original data source, i.e. in a global-as-view (GAV) data integration fashion. This solution eases the maintenance and expansion of the global data schema. In fact, if data schema variations occur in the original data sources, they can be easily managed since they affect only the source-specific part of the global schema. Similarly, the integration of an additional data source only requires adding a sub-schema for the new source (according to its original data schema) in the module of the feature whose data are provided by the new source, without affecting other parts of the global schema. As an example, Figure 2 shows the main tables of the relational instantiation of the conceptual feature module in Figure 1 for the Gene feature, with the main attributes of the gene data form the Entrez Gene and OMIM sources integrated in GPKB.

The automatic aggregation of the main attribute data of each feature source occurs in the aggregation tier, where replicated entries are identified and merged (e.g. multiple feature entries regarding the same gene ID that contain data of distinct attributes of the gene). These operations, which are necessary to ensure correct integration of redundant data from different sources, are automatically performed by the software framework described in Section 4.

3.2 Feature Module Associations

Data feature modules are pairwise associated (through association/annotation data); also these associations are organized in an import and an aggregation tier. In the latter one, the association data, which are contained in the import tier as pairs of feature entry IDs, are automatically translated into pairs of unique OIDs and matched to the feature entry OIDs of the two associated features. By taking advantage of available ID history and similarity data, this translation also allows reconciling discontinued IDs to their current ones, and identifying as such different IDs that represent the same feature instance.

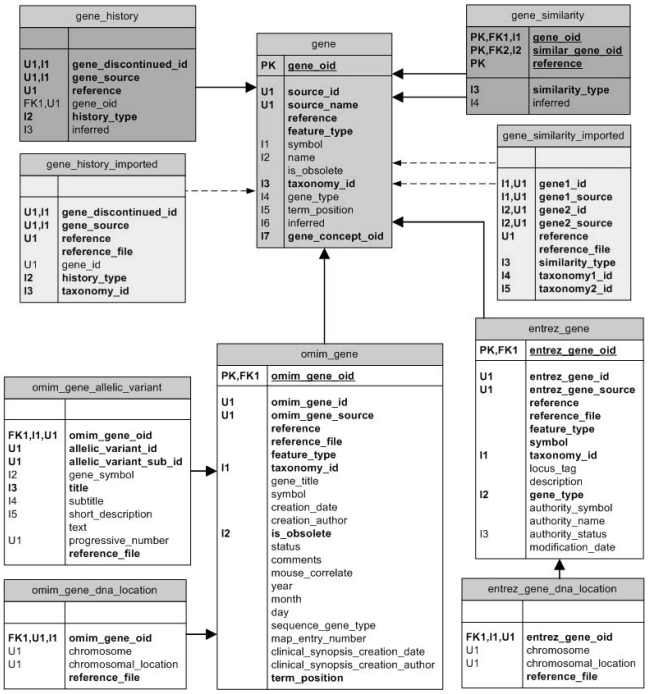

Fig. 2. Main tables of the relational instantiation of the feature module in Figure 1 for the Gene feature, which in GPKB is mapped to the data sources Entrez Gene and OMIM, showing their main attributes. PK: primary key; FK: foreign key; U: unique index; I: index; bold attribute names: required attribute value.

Finally, a third, higher and more general *integration tier* (not shown in Figure 1 and Figure 2) completes the integrated data schema by representing all the unique feature instances, or concepts, (e.g. all distinct genes, proteins, pathways, genetic disorders, etc.) and their associations described by the integrated data, regardless of the source(s) that provide(s) them (e.g. all the integrated distinct genes and their annotated features, regardless of the multiple IDs of each of them and their providing source). Yet, each of these unique concepts (e.g. each gene) is related to all its entries from distinct data sources in the lower schema levels (e.g. all the gene data associated with each gene ID from each integrated source that provides them), thus keeping all its provenance information.

3.3 Configuration for Automatic Creation

The specific implementation of our general integrated data schema depends on the particular features and concepts to be represented (in the *concept-integration* upper tier) and on the data sources being imported (in the *source-import* lower tier) and integrated. To automate its dynamic creation, we designed a XML file where register-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2453944, IEEE/ACM Transactions on Computational Biology and Bioinformatics

ing the data sources and their provided data to be imported and integrated. In this file, metadata are specified for each source, defining the location from where retrieving the source data, as well as their type (i.e. main, history, similarity, or association data) and the feature that each of them describes. General data import templates are associated with each data type, matched to the source provided data and, if needed, extended to include specific attribute data that the particular source may provide. This allows both standardizing the identification and import of the main data attributes, and their automatic aggregation in the data schema aggregation tier; on the other hand, it also lets full flexibility in the management of any source data, as well as eases and quickens the filling of the configuration XML file.

Therefore, our general data schema can be iteratively extended (both tier and feature wise) in a seamless, scalable and modular way (just by registering in the XML file new sources and their represented feature(s), if the latter one(s) is(are) not already present in the file), in order to include many different biomolecular entities, biomedical features and their annotations and associations from several different sources, virtually without limitations.

Thus, differently from previously proposed global schemas and architectures for biomolecular data (e.g. BioMart [12], or BioWarehouse [17]), this general schema *supports complete automation of the integration process*, as illustrated and discussed in Section 4. Furthermore, our proposed global data schema and architecture allow accessing the integrated data at different levels in order to efficiently perform different types of queries. It can be physically implemented with an object-oriented or relational style; since the latter one results most efficient with large data collections, we opted for a relational implementation.

4 Software Architecture for Semantic Data Integration

Benefitting from abstraction, modularity and configurability of our integrated data schema, in Java programming language we created a generalized and parametric software architecture, which supports the customized automated creation of a data warehouse adopting our data schema, and makes updating the data warehouse and extending it with new data sources easy. Our approach for the integration of distributed multi-source heterogeneous data is divided in two macro steps (Figure 3), performed according to the defined configuration metadata:

1) *Importing* data from their diverse sources in the *source-import* tier of our integrated data schema,
2) *Integrating* them in the *instance-aggregation* and *concept-integration* tiers of the data schema.

4.1 Data Import

The data import procedure is guided by an *import manager* that instantiates, configures and executes an *importer* for each considered data source. Each source specific importer coordinates a set of *loaders* (a loader for each data file, group of homogeneous data files, or data access API

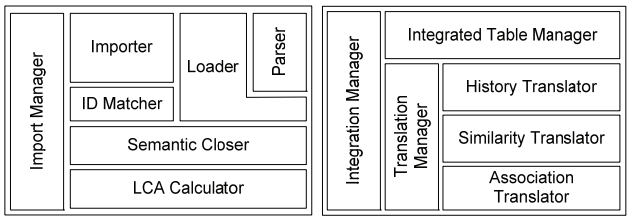

Fig. 3. Main components of the two parts (for data import and data integration, respectively) of our software architecture.

provided by the source) and a set of *parsers* (a parser for each data format). Each parser extracts the data from its associated input file(s) or API(s) and produces data tokens usable by a loader. Each loader is responsible for associating a semantic meaning to the tokens produced by the associated parser and inserting them into the warehouse.

When the imported data describe an ontology, the import manager also executes the semantic closure of such ontology along its IS_A relationships, i.e. an unfolding processing of the ontology hierarchical structure which inserts into the data warehouse an explicit relationship between each term (node) of the ontology and each of its ancestors related through IS_A hierarchical relationships. The aim of this semantic closure is to speed up subsequent semantic queries and computational analyses on such semantic data and their annotations. In addition, the import manager computes also the *Lowest Common Ancestor* (*LCA*) [25] for each pair of nodes in the ontology; this is a fundamental step for subsequent evaluations of the similarity between two terms of an ontology, and then between genes or proteins based on their semantic annotations to such terms, according to various metrics commonly used [26]-[29].

In the import process, each actor is independent: the import manager administers the importers via a standard interface based on Java reflection API. The parser is aware of data format, but agnostic of data semantics; the loader receives data in a standard format and inserts them in the proper data warehouse table(s). To guarantee flexibility and easy addition of new sources, the process is guided by several *configuration metadata*, describing the characteristics of all the registered data sources to be imported, all the features (biomolecular entities and biomedical features) represented by the data to be imported and their bindings. Such metadata are used to map each data source and feature to one or more data warehouse tables and their bindings are used to populate such tables.

The importing framework assigns to each imported "data record" an OID, which is unique across the data warehouse. It is used as the primary identification of the data entries, since there is no guarantee that the IDs provided by the different sources do not conflict with each other. In order to ensure correctness of imported data, a set of regular expressions has been defined to check and identify IDs [22]. They are used by the *ID matcher*, an additional component of our architecture that acts as a mediator between the loaders and the data warehouse. The main role of this mediator is to check ID syntactic correct-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2453944, IEEE/ACM Transactions on Computational Biology and Bioinformatics

ness and identify the semantic type of each ID, in order to insert the correct information in the appropriate data warehouse tables. During this process, each inserted tuple is also enriched with provenance data to track its source. Correct ID identification is paramount since data from multiple sources are then linked together thanks to association data provided by the integrated sources as pairs of IDs in different data sources.

Since input data may contain errors and their structure is subject to modifications in subsequent data versions, checking of input data is strictly enforced. Verification is done in three steps: during parsing, during data loading and at the end of the import process when index, unique, primary and foreign key integrity constraints are defined and enforced upon the data warehouse tables.

4.2 Data Integration

The data integration step consists of two automatic tasks: *aggregation* and *integration*. In the former task, data from the different sources, imported in the previous data import step, are gathered and normalized into a single representation in the *instance-aggregation* tier of our global data schema. In the latter task, data are organized into informative clusters in the *concept-integration* tier of the integrated data schema.

During the aggregation phase, based on the metadata included in the configuration file, tables of the features described by the imported data are automatically created and populated. Then, similar IDs (e.g. aliases of feature IDs) and historical IDs, which are sometimes provided by the data sources, are translated to our internal OIDs and respectively stored in the *similarity* and *history* tables of the feature to which they refer (see Figure 1). Unfolding of historical IDs is performed before OID translation, so as to associate repeatedly superseded and discontinued IDs with the translated OID of their latest ID. Entries derived from this processing are marked as *inferred through historical data*, in order to keep full track of their generation process. Both similarity and historical ID data are extremely valuable for subsequent data integration tasks. Translation tables for biomolecular entity and biomedical feature IDs are also created by using translated similarity data and unfolded historical ID data. These serve as main entry points to query and explore the data warehouse; they allow the conversion from a number of user-provided identifiers (also obsolete or alias of those in the warehouse feature tables) to a set of current OIDs, which are usable to navigate the warehouse.

Then, associations between pairs of feature entries are created by performing OID translation of the imported association (annotations) data expressed through the feature entry IDs. In doing so, association data are coupled with the related feature entries. Depending on the imported data sources and their mutual synchronization, association data may refer to feature entries, or even features, that have not yet been imported in the data warehouse. In this case, missing integrated feature entries are synthesized and marked as *inferred through synthesis from association data*. When a missing entry has an obsolete ID and its most current ID can be obtained through unfolded historical ID data, the association is transferred to the latest ID and marked as *inferred through historical data*. This association translation policy preserves, after integration, all the associations expressed by the imported association data from different sources. Thus, it allows subsequently using such associations for biomedical knowledge discovery (e.g. by transitive relationship inference [30], optionally involving also the synthesized entries).

During the final integration phase, through a "similarity analysis", it is checked whether single feature entries from different sources represent the same feature concept (e.g. different gene entries identified by different IDs of the same gene from distinct sources). In this case, they are associated with a new single concept OID (e.g. a gene concept OID). Furthermore, new entries can be inferred from the integrated data [30]. The *Inferred* attribute in the integrated tables is used to keep track of the inference method employed, if any, to derive an entry.

5 Genomic and Proteomic Knowledge Base

To demonstrate the relevance and effectiveness of the described general software architecture and the integrated data schema, which we implemented in a PostgreSQL RDBMS, we used them to create, maintain updated and progressively extend a multi-organism integrative Genomic and Proteomic Data Warehouse (GPDW). It constitutes a high quality and consistent integration of numerous biomolecular interaction and semantic annotation data describing several biomedical-molecular features of many biomolecular entities, particularly genes and proteins. Such data are imported from several distributed data sources, carefully selected for their renewed relevance, which include Entrez Gene, UniProt, IntAct, Expasy Enzyme, GO, GOA, BioCyc, KEGG, Reactome and OMIM. At the time of writing, the GPDW contains more than 2.36 billion data tuples; in the multi-level data schema used, they amount to a total of about 737 GB of disk space (including their indexing). They include about 17,535,404 genes of 14,995 different organisms, 19,544,576 proteins of 23,368 species and a total of 16,772,399 gene annotations and 30,440,619 protein annotations expressed though 10 biomedical controlled terminologies or ontologies. The latter ones included 41,829 Gene Ontology terms and their 42,057,775 semantic annotations, 359,511 biochemical pathways from BioCyc (321,832), KEGG (469) or Reactome (37,210) and 1,733,389 pathway annotations, as well as 7,936 human genetic disorders from OMIM and their 12,473 gene annotations, together with 34,177 phenotypes (signs and symptoms). We extracted phenotypes from the OMIM clinical synopsis semi-structured descriptions as illustrated in [5]; at the time of writing, to our knowledge, they are not included in structured, easy queryable form in any other integrative database publicly available. Furthermore, the GPDW integrates 626,516 valuable interaction data from IntAct between different biomolecular entities, including 609,864 protein-protein interactions. In addition, the GPDW contains 3,616,108 semantic annotations of 988,899 genes which we recently detected by transitive relationship from the integrated

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2453944, IEEE/ACM Transactions on Computational Biology and Bioinformatics

annotations of the proteins that these genes encode; the semantic processing that we used to this purpose is thoroughly illustrated and discussed in [30].

The GPDW constitutes the backend of a Genomic and Proteomic Knowledge Base (GPKB) publicly available at http://www.bioinformatics.deib.polimi.it/GPKB/. A Web interface provides different functionalities that enable the scientific community to access and comprehensively query all the data integrated in the GPDW and take full advantage of them.

The *Basic search* functionality is available for searches aimed at retrieving all information directly associated with a single feature instance, either imported from external sources or inferred based on the integrated data; for example, all annotations and interactions of a specific gene or protein (e.g. the human *insulin-like growth factor 2 (somatomedin A)* (*IGF2*) gene, Entrez Gene ID 3481), or all genes and proteins annotated to a particular biomedical feature instance, such as a specific pathway or genetic disorder (e.g. the *Alzheimer disease*, OMIM ID 104300). Full provenance information for each retrieved association (annotation or interaction) is provided. When annotations are inferred by transitive relationships [30], an additional window provides the available annotations on which each inferred annotation is based. All extracted data can be downloaded in text format for their easy further use.

We also implemented an enhanced functionality and graphical interface for multi-feature search, named *Easy search*. It supports the simple graphical composition of complex queries on multiple features just by orderly selecting the required features, e.g. gene, pathway, enzyme, biological function feature, genetic disorder, clinical synopsis, etc. (Figure 4); if needed, display and filtering constrains can be defined for any attribute of each selected feature just by specifying them in the feature window (Figure 5). For example, let us suppose the GPKB user wants to search for genes whose variants are known to be associated with genetic diseases and find the clinical aspects of such diseases and all the biological functions in which those genes are known to be involved, in order to check if common gene functions and clinical aspects exist in different but related pathologies (e.g. in *Muscular dystrophy, Duchenne type* and in *Amyotrophic lateral sclerosis 1*). Using the *Easy search* functionality, the user can orderly selects the *gene* feature, then the gene associated *biological function feature* and *genetic disorder* features, and then the genetic disorder associated *clinical synopsis* feature; finally, before submitting the query, if the user wants to

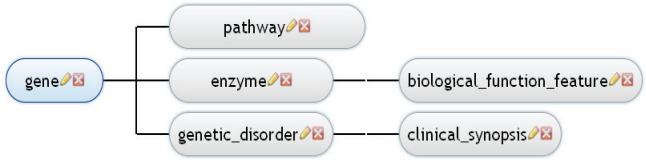

Fig. 4. Easy graphical composition of queries on multiple integrated features through the GPKB *Easy search* user interface. A query involving *genes*, *pathways*, *enzymes* and their *biological function features* (e.g. cellular components), *genetic disorders* and their *clinical synopses* is being composed. Clicking on the pencil icon of a feature opens the feature window (Figure 5).

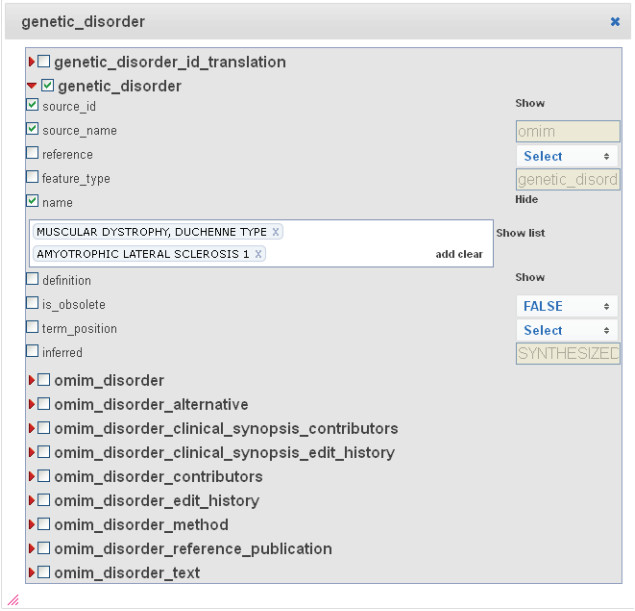

Fig. 5. Feature window of the *genetic disorder* feature where display and filtering constrains can be defined for any attribute of the feature.

investigate only some related pathologies, he/she can specify them as value of the *name* attribute in the *genetic disorder* feature window (Figure 5). To our knowledge, these complex multi-topic queries cannot be performed in such an easy way in any other available system. Furthermore, despite the high amount of data contained, the GPKB also provides good query performances: the response time of the described complex query is 7 seconds.

Query results are shown in a table view, whose columns can be freely composed by the user for best and easy exploration of the results (Figure 6). Additionally, the user can select all or a subset of the query results (and of their attributes) and expand the initial performed query in order to refine or augment them, according to the *liquid query* principle [31]; this supports very useful explorative searches of the numerous and heterogeneous data integrated in the GPDW and eases biomedical knowledge extraction, particularly when the search query full specification and data filtering values are not known *a priori*, but can be determined by observing partial query results. For example, a GPKB user can start searching for the genes known to be involved in a given pathway (e.g. *Apoptosis*, or *RNA transport*). Then, he/she can refine the initial query results by selecting only some of the found genes (e.g. the ones he/she is more interested in), and by extracting only those of them that encode for enzymes (by selecting the *gene* associated *enzyme* feature and adding it in the query). Finally, the user can expand the obtained results to find in which cellular components, if any, each of such enzymes is known to be expressed (see upper and central part of Figure 4).

Furthermore, by taking advantage of the biomolecular ontologies integrated in the GPDW and their semantic closure performed by the GPDW software architecture, the GPKB also supports semantic expansion of the user query; this functionality is not available through the Web

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2453944, IEEE/ACM Transactions on Computational Biology and Bioinformatics

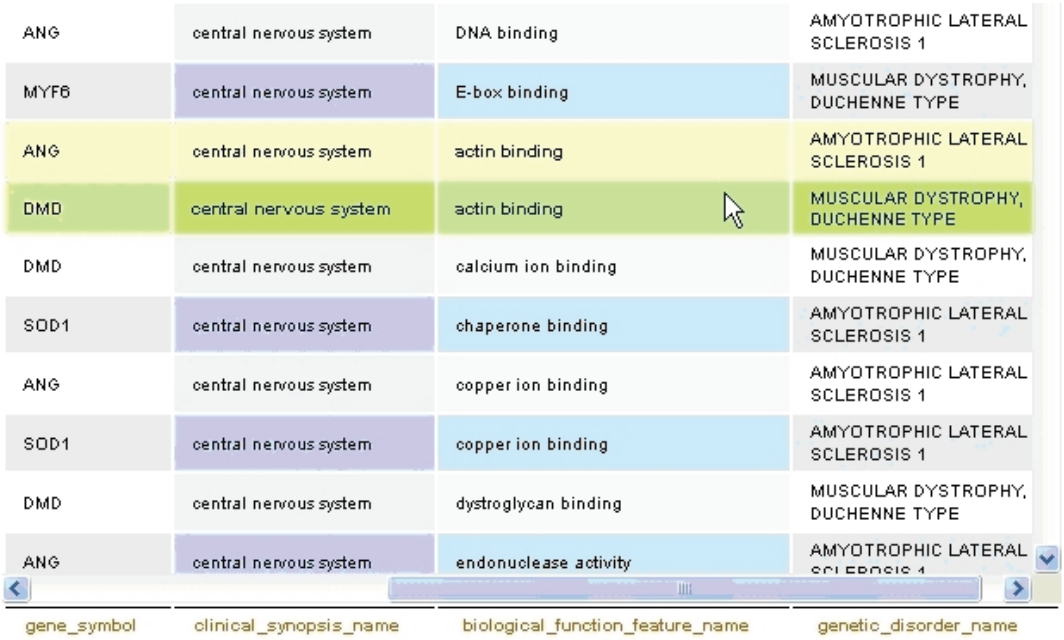

Fig. 6. Gene biological functions (e.g. *actin binding*) and clinical aspects (e.g. *central nervous system*) found with the GPKB in common in *Muscular dystrophy, Duchenne type* and in *Amyotrophic lateral sclerosis 1* genetic disorders; pointing the mouse on an attribute value, highlights all query results with that attribute value.

interfaces of the original sources from where the semantic annotations integrated in the GPDW have been taken. All biomolecular databases store only the most specific ontological annotations of a biomolecular entity, whereas all the many less specific, indirect annotations are left implicit for space reasons. This semantic expansion allows retrieving also all such indirect, less specific ontological annotations of a biomolecular entity in the GPDW, e.g. of a gene, easing the understanding of all the biomolecular entity characteristics and the use also of such annotations, e.g., for gene enrichment analysis [6], [32] or semantic annotation prediction [33]-[35].

6 Example Use Case

In order to show potential and relevance of the GPKB, in this Section we illustrate and discuss a complete example of its use in order to answer significant biological questions about complex disorders and unveil common biological aspects in apparently unrelated genetic disorders.

Complex disease mechanisms can be explained by molecular pathways, involving subgroups of genes and their interactions, where phenotypic complexity results from interactions among genomic variants in different loci. In particular, polygenic traits are due to the synergetic activity of multiple genes, which in turn can be pleiotropic genes, i.e. involved in more pathological phenotypes. Thus, the complex relationships among gene variants and diseases can be better understood by pointing out the genes and metabolic pathways involved in pathological phenomena. Comprehensive search of integrated biomedical information and knowledge from multiple sources can help clarifying which genes and pathways (processes among genes) may be the causal candidates of a disease. In fact, some genes may be hubs connecting different complex disease modules (e.g. different cancer types) and so they may have a key role in the disease development (e.g. in carcinogenesis). This approach can also be the basis for some new discovery, or at least clarification, of broader phenomena.

As an example, suppose we want to highlight possible molecular mechanisms at the basis of distinct types of cancer, e.g. *Breast cancer* and *Prostate cancer* which are very different and complex diseases involving different genders. To do so, we want to detect all those genes that are involved in both cancer forms, and the pathways in which those genes are known to be involved. In the GPKB, just by simply selecting the *gene* and *genetic disorder* features and specifying *Breast cancer* and *Prostate cancer* as values of the *name* attribute of the *genetic disorder* feature, we can discover that six genes (i.e. *AR*, *ARL11*, *BRCA2*, *CASP8*, *CDH1*, and *CHEK2*) are involved in both diseases. Then, we can augment such initial query result and expand the *gene* feature, just by selecting its associated *pathway* feature, to search also for pathways in which those genes are known to be involved. By doing so, we found the 19 pathways shown in Table 2.

Such findings may help uncovering common molecular causes of the diseases, despite their phenotypic variability. For example, the found *Homologous recombination* and *Meiosis* pathways specifically point out that possible disease causing variations may occur during cell division of the germ line, where meiosis is at the base of variability and inheritance through genetic recombination. Furthermore, using the GPKB we found that variations of one of the genes found involved in both breast and prostate cancers, the *cadherin 1, type 1, E-cadherin (epithelial)* (*CDH1*) human gene, are known to be involved also in other cancer forms, i.e. *Colorectal cancer*, *Endometrial cancer*, *Gastric*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2453944, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE 2
KNOWN PATHWAYS OF GENES INVOLVED IN BOTH *BREAST CANCER* (BC) AND *PROSTATE CANCER* (PC) DISEASES

Gene ID	Gene symbol	Genetic disease	Pathway	Pathway ID	Pathway source
999	CDH1	BC PC	Adherens junction	4520	KEGG
999	CDH1	BC PC	Apoptosis	578	Reactome
999	CDH1	BC PC	Bacterial invasion of epithelial cells	5100	KEGG
999	CDH1	BC PC	Bladder cancer	5219	KEGG
999	CDH1	BC PC	Cell adhesion molecules (CAMs)	4514	KEGG
11200	CHEK2	BC PC	Cell cycle	4110 115566	KEGG Reactome
999	CDH1	BC PC	Cell-Cell communication	111155	Reactome
675	BRCA2	BC PC	DNA repair	216	Reactome
999	CDH1	BC PC	Endometrial cancer	5213	KEGG
675	BRCA2	BC PC	Homologous recombination	3440	KEGG
999	CDH1	BC PC	Immune System	6900	Reactome
675	BRCA2	BC PC	Meiosis	111183	Reactome
999	CDH1	BC PC	Melanoma	5218	KEGG
11200	CHEK2	BC PC	p53 signaling pathway	4115	KEGG
675	BRCA2	BC PC	Pancreatic cancer	5212	KEGG
999	CDH1	BC PC	Pathogenic Escherichia coli infection	5130	KEGG
675 999	BRCA2 CDH1	BC PC BC PC	Pathways in cancer	5220	KEGG
999	CDH1	BC PC	Signal transduction	111102	Reactome
999	CDH1	BC PC	Thyroid cancer	5216	KEGG

cancer, and *Ovarian cancer*. This shows that this gene may play a central role in regulating different biological processes, and so different cancer disease modules.

7 DISCUSSION AND CONCLUSIONS

Relevant progresses in biotechnology and system biology are creating a remarkable amount of biomolecular data and semantic annotations; they increase in number and quality, but are dispersed and only partially connected. Integration and mining of these distributed and evolving data and information have the high potential of discovering hidden biomedical knowledge useful in understanding complex biological phenomena, normal or pathological, and ultimately of enhancing diagnosis, prognosis and treatment; but such integration poses huge challenges. Our work has tackled them by developing a novel and generalized way to define and easily maintain updated and extend an integration of many evolving and heterogeneous data sources; our approach proved useful to extract biomedical knowledge about complex biological processes and diseases.

The multilevel integrated data schema developed allows data warehouse integrations based on the features described by the imported data sources. Differently from previous warehousing work (e.g. BioWarehouse [17], or Biozon [18]), our data schema and software architecture are generic; thus, they have the potential for overcoming the maintenance and extension issues posed by the warehousing technique. Our approach is particularly suited for life science data, which often evolve both in content and, although less frequently, in structure and number; we guarantee data consistency, reconcile unsynchronized data, and identify equivalence of data from different sources. Furthermore, we take advantage of and process the integrated semantic data to make them ready for efficient semantic querying and further evaluations. These remarkable assets, which increase data integration and quality significantly and enable complex biological queries, to the best of our knowledge are not all together present in other biomedical integrative systems.

Our proposed data schema and architecture enable us to easily create, keep updated and progressively extend the GPKB, a publicly available collection of numerous, semantic biomolecular annotation data expressed through multiple ontologies and originally available separately in many different sources. Our system supports comprehensive semantic queries with adequate performance, even if running on big data [36]. Furthermore, by recording data provenance and modeling associations among the integrated data, GPKB supports comprehensive reliable data analyses and mining. This helps answering complex multi-topic biomedical questions, which cannot be equally managed by other available systems such as BioMart [12]. As our example use case shown, the GPKB is mostly useful to unveil hidden biomolecular and biomedical associations; these can be crucial to better understand the molecular mechanisms of complex diseases and foster relevant biomedical knowledge discoveries. The developed GPKB Web interface well enables users to easily compose queries, although complex, on multiple features/topics and to extract valuable biological insights, including hidden associations, which may answer significant biomedical questions.

Relevance of the GPKB is also proved by the about 47,000 accesses received by more than 1,280 visitors since when it opened on the Web about 30 months ago, and by its use within multiple projects, including our frameworks for detection and prediction of semantic biomolecular annotations [30], [35], [37], and the Bio-SeCo system [38] for the ranking aware composition of search results in support of distributed bio-data explorative search to answer complex biomedical questions.

In the near future, we plan to open to the public also a programmatic access to the GPDW through a collection of Web services that we are developing for this purpose. Public availability of a service interface will enable the access to the GPDW data by other computational systems, with possible inclusion in scientific workflows and new foreseen applications, e.g. in drug repurposing [39]; we are carrying on a project on this topic in collaboration with the US National Library of Medicine.

ACKNOWLEDGMENT

The authors thank Fernando Palluzzi, who suggested the here presented example use case of queries on the biomolecular semantic annotations integrated in the GPKB

and who assessed the relevance of the obtained results. We also thank Stefano Gennaro for the implementation of the *Easy search* Web interface.

This work was supported by the "Data-Driven Genomic Computing (GenData 2020)" PRIN project (2013-2015), funded by the Italian Ministry of the University and Research (MIUR).

REFERENCES

[1] M.Y. Galperin, D.J. Rigden, and X.M. Fernández-Suárez. "The 2015 Nucleic Acids Research Database Issue and Molecular Biology Database Collection", *Nucleic Acids Res*., vol. 43, D1, pp. D1-D5, 2015.

[2] W. Sujansky, "Heterogeneous database integration in biomedicine", *J. Biomed. Inform*., vol. 34, 4, pp. 285-298, 2001.

[3] B. Louie, P. Mork, F. Martin-Sanchez, A. Halevy, and P. Tarczy-Hornoch, "Data integration and genomic medicine", *J. Biomed. Inform*., vol. 40, 1. pp. 5-16, 2007.

[4] C. Goble, and R. Stevens, "State of the nation in data integration for bioinformatics", *J. Biomed. Inform*., vol. 41, pp. 687-693, 2008.

[5] M. Masseroli, O. Galati, and F. Pinciroli, "GFINDer: Genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists", *Nucleic Acids Res*., vol. 33, pp. W717-W723, 2005.

[6] M. Masseroli, "Management and analysis of genomic functional and phenotypic controlled annotations to support biomedical investigation and practice", *IEEE Trans. Inf. Technol. Biomed*., vol. 11, 4, pp. 376-385, 2007.

[7] T. Etzold, A. Ulyanov, and P. Argos, "SRS: Information Retrieval System for molecular biology data banks", *Methods Enzymol*., vol. 266, pp. 114-128, 1996.

[8] T.A. Tatusova, I. Karsch-Mizrachi, and J.A. Ostell, "Complete genomes in WWW Entrez: data representation and analysis", *Bioinformatics*, vol. 15, pp. 536-543, 1999.

[9] R. Stevens, P. Baker, S. Bechhofer, G. Ng, A. Jacoby, N.W. Paton, C.A. Goble, and A. Brass, "TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources", *Bioinformatics*, vol. 16, pp. 184-185, 2000.

[10] Z. Ben Miled, N. Li, G.M. Kellett, B. Sipes, and O. Bukhres, "Complex life science multidatabase queries", *Proc. IEEE*, vol. 90, pp. 1754-1763, 2002.

[11] S.B. Davidson, C. Overton, V. Tanen, and L. Wong, "BioKleisli: a digital library for biomedical researchers", *Int. J. Digit. Libr*. , vol. 1, pp. 36-53, 1997.

[12] D. Smedley, S. Haider, B. Ballester, R. Holland, D. London, G. Thorisson, A. and Kasprzyk, "BioMart - Biological queries made easy", *BMC Genomics*, vol. 10, 22, pp- 1-12, 2009.

[13] A. Freier, R. Hofestädt, M. Lange, U. Scholz, and A. Stephanik, "BioDataServer: a SQL-based service for the online integration of life science data", *In Silico Biol*., vol. 2, no. 2, pp. 37-57, 2002.

[14] E. Cadag, B. Louie, P.J. Myler, and P. Tarczy-Hornoch, "Biomediator data integration and inference for functional annotation of anonymous sequences", *Pac. Symp. Biocomput*., pp. 343-354, 2007.

[15] K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, et al., "The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud", *Nucleic Acids Res*., vol. 41, no. Web Server issue, pp. W557-W561, 2013.

[16] D. Blankenberg, N. Coraor, G. Von Kuster, J. Taylor, A. Nekrutenko; and Galaxy Team, "Integrating diverse databases into an unified analysis framework: a Galaxy approach", *Database*, vol. 29; 2011:bar011, pp. 1-9, 2011.

[17] T.J. Lee, Y. Pouliot, V. Wagner, P. Gupta, D.W. Stringer-Calvert, J.D. Tenenbaum, and P.D. Karp, "BioWarehouse: a bioinformatics database warehouse toolkit", *BMC Bioinformatics*, vol. 7, 170, pp. 1-14, 2006.

[18] A. Birkland, and G. Yona, "BIOZON: a system for unification, management and analysis of heterogeneous biological data", *BMC Bioinformatics*, vol. 7, 70, pp. 1-24, 2006.

[19] S.B. Davidson, C. Overton, and P. Buneman, "Challenges in integrating biological data sources", *J. Comput. Biol*., vol. 2, no. 4, pp. 557-572, 1995.

[20] N.W. Paton, S.A. Khan, A. Hayes, F. Moussouni, A. Brass, K. Eilbeck, et al., "Conceptual modeling of genomic information", *Bioinformatics*, vol. 16, no. 6, pp. 548-557, 2000.

[21] E. Bornberg-Bauer, and N.W. Paton, "Conceptual data modelling for bioinformatics", *Brief. Bioinform*., vol. 3, no. 2, pp. 166-180, 2002.

[22] G. Ghisalberti, M. Masseroli, and L. Tettamanti, "Quality controls in integrative approaches to detect errors and inconsistencies in biological databases", *J. Integr. Bioinform*. vol. 7, 119, pp. 1-13, 2010.

[23] F. Belleau, M.A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2RDF: towards a mashup to build bioinformatics knowledge systems", *J. Biomed. Inform*., vol. 41, no. 5, pp. 706-716, 2008.

[24] A. Callahan, J. Cruz-Toledo, and M. Dumontier, "Ontology-based querying with Bio2RDF's linked open data", *J. Biomed. Semantics*, vol. 4, no. Suppl 1, S1, pp. 1-13, 2013.

[25] M.A. Bender, M. Farach-Colton, G. Pemmasani, S. Skiena, and P. Sumazin, "Lowest Common Ancestors in trees and directed acyclic graphs", *J. Algorithms*, vol. 57, no. 2, pp. 75-94, 2005.

[26] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy", *Proc. IJCAI'95*, vol. 1, pp. 448-453, 1995.

[27] J.J. Jiang, and D.W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy", *Proc. ROCLING97*, ACLCLP, pp. 19-33, 1997.

[28] D. Lin, "An information-theoretic definition of similarity", *Proc. ICML '98*, vol. 98, pp. 296-304, 1998.

[29] A. Schlicker, F.S. Domingues, J. Rahnenfoehrer, and T. Lengauer, "A new measure for functional similarity of gene products based on Gene Ontology", *BMC Bioinformatics*, vol. 7, 302, pp. 1-16, 2006.

[30] M. Masseroli, A. Canakoglu, and M. Quigliatti, "Detection of gene annotations and protein-protein interaction associated disorders through transitive relationships between integrated annotations", *BMC Genomics*, vol. 16, no. Suppl. 6, S5, pp. 1-16, 2015.

[31] A. Bozzon, M. Brambilla, S. Ceri, and P. Fraternali, "Liquid query: multi-domain exploratory search on the web". *Proc. WWW '10*, pp. 161-170, 2010.

[32] D.W. Huang, B.T. Sherman, and R.A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists". *Nucleic Acids Res*., vol. 37, no. 1, pp. 1–13, 2009.

[33] O.D. King, R.E. Foulger, S.S. Dwight, J.V. White, and F.P. Roth, "Predicting gene function from patterns of annotation". *Genome Res.*, vol. 13, no. 5, pp. 896–904, 2003.

[34] P. Khatri, B. Done, A. Rao, A. Done, and S. Draghici, "A semantic analysis of the annotations of the human genome". *Bioinformatics*, vol. 21, no. 16, pp. 3416–3421, 2005.

[35] P. Pinoli, D. Chicco, and M. Masseroli, "Enhanced probabilistic latent semantic analysis with weighting schemes to predict genomic annotations". *Proc. BIBE 2013*, IEEE, 92, pp. 1–4, 2013.

[36] F. Pessina, M. Masseroli, and A. Canakoglu, "Visual composition of complex queries on an integrative genomic and proteomic data warehouse", *Engineering*, vol. 5, no. 10B, pp. 94-98, 2013.

[37] D. Chicco, and M. Masseroli, "Software suite for gene and protein annotation prediction and similarity search", *IEEE/ACM Trans. Comput. Biol. Bioinform.*, preprint, 2015, doi: 10.1109/TCBB.2014.2382127.

[38] M. Masseroli, M. Picozzi, G. Ghisalberti, and S. Ceri, "Explorative search of distributed bio-data to answer complex biomedical questions", *BMC Bioinformatics*, vol. 15, no. Suppl. 1, S3, pp. 1-14,, 2014.

[39] T. Cohen1, D. Widdows, R.W. Schvaneveldt, P. Davies, and T.C. Rindflesch, "Discovering discovery patterns with Predication-based Semantic Indexing", *J. Biomed. Inform.*, vol. 45, no. 6, pp. 1049-1065, 2012.

Marco Masseroli All received the Laurea Degree in electronic engineering in 1990 from Politecnico di Milano, Italy, and a PhD in biomedical engineering in 1996, from Universidad de Granada, Spain. He is Professor at the Dipartimento di Elettronica, Informazione e Bioingegneria of Politecnico di Milano, and lecturer of Bioinformatics and BioMedical Informatics. He carried out research activity in the application of Information Technology to the medical and biological sciences in several Italian and international research centers. He has also been Visiting Professor at the Departamento de Anatomía Patológica, Facultad de Medicina of the Universidad de Granada - Spain, and Visiting Faculty at the Cognitive Science Branch of the National Library of Medicine, National Institute of Health, Bethesda - US. His research interests are in the area of bioinformatics and biomedical informatics, focused on distributed Internet technologies, biomolecular databanks, controlled biomedical terminologies and bio-ontologies to effectively retrieve, manage, analyze, and semantically integrate genomic information with patient clinical and high-throughout genetic data. He is the author of more than 170 scientific articles, which have appeared in international journals, books and conference proceedings.

Arif Canakoglu received the Master Degree in computer engineering in 2010 from Politecnico di Milano and now he is a PhD Student at the Politecnico di Milano. His research interests are in the areas of bioinformatics data integration and distributed access to heterogeneous data through Service Oriented Architectures and Web Service composition, in order to support answering complex biomedical questions requiring comprehensive analysis of multiple heterogeneous data.

Stefano Ceri is Professor of Database Systems at the Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB) of Politecnico di Milano. He obtained his Dr. Eng. Degree from Politecnico di Milano in July 1978. He was visiting professor at the Computer Science Department of Stanford University (1983-1990), Chairman of the Computer Science Section of DEI (1992-2004), Director of Alta Scuola Politecnica (ASP) of Politecnico di Milano and Politecnico di Torino (2010-2013). In 2008 he has been awarded an advanced ERC Grant on Search Computing (2008-2013). He is co-founder (2001) of WebRatio (http://www.webratio.com/). His research work covers about four decades (1976-2014) and has been generally concerned with extending database technology in order to incorporate new features: distribution, object-orientation, rules, streaming data, crowd-based and genomic computing. He is currently leading the PRIN project GenData 2020, focused on building query and data analysis systems for genomic data as produced by fast DNA sequencing technology. He is the recipient of the ACM-SIGMOD "Edward T. Codd Innovation Award" (2013), and an ACM Fellow and member of the Academia Europaea.