An Approach for Semantic Data Integration in Cancer Studies

Iliyan Mihaylov, Maria Nisheva-Pavlova(✉), and Dimitar Vassilev

Faculty of Mathematics and Informatics, Sofia University St. Kliment Ohridski,
5 James Bourchier Blvd., 1164 Sofia, Bulgaria
{mihaylov,marian,dimitar.vassilev}@fmi.uni-sofia.bg

Abstract. Contemporary development in personalized medicine based both on extended clinical records and implementation of different high-throughput "omics" technologies has generated large amounts of data. To make use of these data, new approaches need to be developed for their search, storage, analysis, integration and processing. In this paper we suggest an approach for integration of data from diverse domains and various information sources enabling extraction of novel knowledge in cancer studies. Its application can contribute to the early detection and diagnosis of cancer as well as to its proper personalized treatment.

The data used in our research consist of clinical records from two particular cancer studies with different factors and different origin, and also include gene expression datasets from different high-throughput technologies – microarray and next generation sequencing. An especially developed workflow, able to deal effectively with the heterogeneity of data and the enormous number of relations between patients and proteins, is used to automate the data integration process. During this process, our software tool performs advanced search for additional expressed protein relationships in a set of available knowledge sources and generates semantic links to them. As a result, a set of hidden common expressed protein mutations and their subsequent relations with patients is generated in the form of new knowledge about the studied cancer cases.

Keywords: Data integration · Ontology · Linked data · Knowledge extraction · Cancer studies

1 Introduction

Data integration is one of the challenges of contemporary data science with great impact on different practical and research domains. Data generated in medicine from both clinical and omics high-throughput sources contribute to changes in data storage and analytics and to a number of related bioinformatics approaches aiming at better diagnostics, therapy and implementation of personalized medicine [1, 2].

These integrative efforts for research and therapy generate a huge amount of raw data that could be used to discover new knowledge in the studied domains. The extraction of new knowledge is a complex and labor-intensive task with many components – different data sources may have the same attributes but with different

J. M. F. Rodrigues et al. (Eds.): ICCS 2019, LNCS 11538, pp. 60–73, 2019.
https://doi.org/10.1007/978-3-030-22744-9_5

semantics. The used data sources are also heterogeneous: each of them has its own structure and its own data format.

In this paper we present a study of neuroblastoma and breast cancer. These cancers are a great threat for children and women respectively. Breast cancer concerns approximately one in eight women over their lifetime [3], whilst neuroblastoma is the most common cancer in less than one year old children. It accounts for about 6% of all cancers in children [4]. Both cancer datasets vary strongly for each particular case study.

It is essential to discover relations between these two cancers, based on an effective mapping of common mutated proteins. Such an approach can present classes of proteins related to both diseases, which can serve as a basis for more accurate and well annotated discovery of potential molecular markers in cancer studies. From a data science point of view this set of problems can be tackled by an approach based on semantic data integration [5, 6].

Data integration is understood as a mean to combining data from different sources, creating a unified view and improving their accessibility to a potential user [3, 4]. Data integration and biomedical analysis are separate disciplines and have evolved in relative isolation. There is a general agreement that uniting both these disciplines in order to develop more sustainable methods for analysis is necessary [7, 8]. Data integration fundamentally involves querying across different data sources. These data sources could be, but are not limited to, separate relational databases or semi-structured data sources distributed across a network. Data integration facilitates dividing the whole data space into two major dimensions, referring to where data or metadata or knowledge reside and to the representation of data and data models. Biomedical experiments [9] take advantage of a vast number of different analytical methods that facilitate mining relevant data from the dispersed information. Some of the most frequent experiments are related to gene expression profiling, clinical data analytics [10], rational drug design [6], which attempt to use all available biological and clinical knowledge to make informed development decisions. Moreover, machine learning-based approaches for finding and highlighting the useful knowledge in the vast space of abundant and heterogeneous data are applied for improving these analytics. Metadata, in particular, is gaining importance, being captured explicitly or inferred with the aid of machine learning models. Some examples include the use of machine learning methods in the inference of data structure, data distribution, and common value patterns.

The heterogeneity of data makes any integrative analysis highly challenging. Data generated with different technologies include different sets of attributes. Where data is highly heterogeneous and weakly related, two interconnected integrative approaches are applied: horizontal and vertical integration (Fig. 1). The horizontal data integration unites information of the same type, but from different sources and in different formats. It helps to unite heterogeneous data from many different sources in one data model. The vertical integration has a potential to relate different kinds of information, helping for example to manage links between the patient, gene expression, clinical information, chemical knowledge and existing ontologies, e.g., via web technologies [11–13]. Most existing approaches for data integration are focused on one type of data or one disease and cannot facilitate cross-type or disease integration [1, 2].

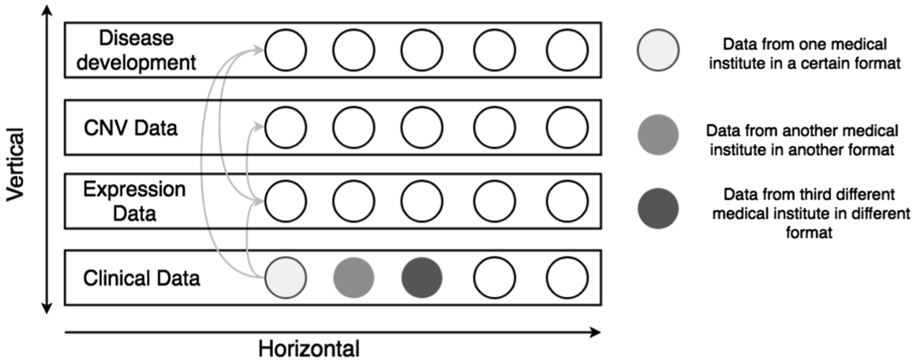
Vertical
Disease development
CNV Data
Expression Data
Clinical Data
Horizontal
Data from one medical institute in a certain format
Data from another medical institute in another format
Data from third different medical institute in different format

Fig. 1. Horizontal and vertical data integration. Grey arrows show the relations between the used types of data (clinical, expression, CNV and disease development). The horizontal arrow shows a flow of integration of all provided data sources, like medical institutes. The vertical arrow shows a potential to link all existing different types of data.

The main objective of this paper is to present a novel efficient data integration model for the studied cancer cases by data mining and knowledge extraction approaches which can find relationships between certain data patterns, related to mutated proteins, expression and copy number variation (CNV). Our approach utilizes NoSQL databases, where we combine clinical and expression profile data, using both raw data records and external knowledge sources.

2 Problem Description

Intelligent exploitation of large amounts of data from different sources, with different formats, and different semantics is among the most important challenges especially in the biomedical area [14]. Life sciences and in particular medicine and medical research generate a lot of such data due to recent developments of high-throughput molecular technologies and large clinical studies [2]. The major challenge here is to integrate, analyze and interpret this data in the scope of contemporary personalized medicine. Personalized medicine (PM) has the potential to tailor therapy and to ensure adequate patient care. By enabling each patient to receive early diagnoses, risk assessments, and optimal treatments, PM holds promise for improving health care while also lowering costs. The information background of the PM is a key element for its successful implementation. This information background is based on semantically reach, accurate and precisely analyzed bio-medical data [15].

The circle of problems in our study can be described in the scope of a successful analysis and integration of the massive datasets, now prevalent in the medical sciences – more precisely, integration of these datasets with linked data sources to find additional relations between proteins and clinical attributes for the two studied diseases. The challenge is to provide a method based on Linked Data and open source technologies [16, 17] to combine knowledge from many existing open sources for efficient integration of raw libratory data. Raw libratory data that can eventually be integrated for the comprehensive elucidation of complex phenotypes include functional gene annotations, gene expression profiles, proteomic profiles, DNA polymorphisms, DNA copy number variations, epigenetic modifications, etc. [18].

The general challenges in our study are based on the use of different data: clinical data, RNA-Seq and microarray gene expression data, and CNV from comparative genome hybridization (aCGH) [19]. The solution of these challenges is to demonstrate and examine the power of semantic data integration and knowledge extraction for the purposes of real world clinical settings in breast cancer (BC) and neuroblastoma (NB).

A specific challenge in the study is to integrate and analyze sets of unbalanced and non-structured data. The molecular data is in raw format with all fields and attributes generated from the sequencing or microarray technology. Before starting the integration process, it is necessary to perform some preprocessing operations on the raw formats and to generate an appropriate new data structure. We are working with datasets, rich in relations, and it is essential to be able to find many annotations for the existing relations which will help one to enhance the set of relations by proper resources from the available knowledge bases.

Solutions based on semantic integration of data have already been successfully applied to cancer datasets to find driver proteins and pathways [20]. We chose an approach based on semantic integration because most features of our data have different semantics for each patient, which is an essential background for personalized medicine. The expected results of such type of approach include identification of hidden protein subtypes distinguished by common patterns of network alteration and a predictive model for cancer development based on the knowledge about joined proteins.

3 Data Description

The raw data in each studied dataset are in a specific format and have specific semantics. A field (an attribute) in each dataset has different meanings due to the technologies and the subsequent recording. The provided data by itself also contain information for mutated proteins, expression and CNV.

The initial point for transformation, grouping and integration are the patients/ sample files. The generated record for each particular patient contains attributes like age, gender, nationality, etc. Two datasets – neuroblastoma (NB) and breast cancer (BC), are used in this study. The neuroblastoma set contains RNA-Seq gene expression profiles of 498 patients as well as Agilent microarray expression and aCGH data for a matched subset of 145 patients and corresponding clinical information. The breast cancer set contains profiles for microarray and copy number data, and clinical information (survival time, multiple prognostic markers, therapy data) for about 2,000 patients.

Each patient record contains many files with expression, mutation and CNV data shown on Fig. 2. For example, each record from expression files refers to another file with detailed information. It includes 100–200 sample records for one expression and contains also information about the mutation type, expressed chromosome, expression start position in the DNA, expression end position. All properties were generated by the used sequencing technology. A mutation file contains proteins, their attributes and reference to the expression file with detailed information. The relationships patient – protein expression and patient – protein mutation are fundamentally different. A patient who has an expressed protein may not have the same protein mutated.

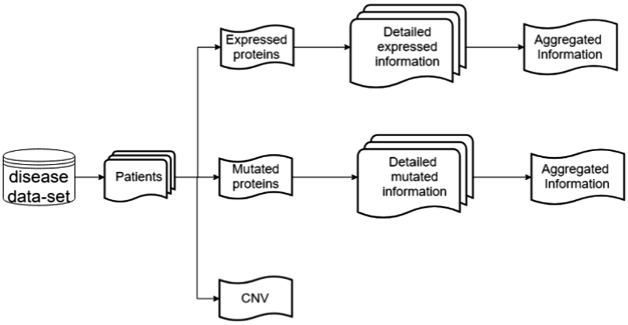

Fig. 2. General structure of raw data.

Both datasets contain some additional files for medical patient stability, meta clinical information, meta expression/mutation information, etc.

4 Related Work

Data integration is a real challenge for querying across multiple autonomous and heterogeneous data sources [5, 9]. It is crucial in the medical and health sector that use multiple sources for producing datasets and is of great importance for their subsequent analysis for study, diagnostics, and therapy purposes [11–13].

A major objective of data integration is to enable the use of data with implicit knowledge and to aid the use of the integrated sources to answer queries [6, 21].

In medical studies and in particular in cancer studies, there are several quite different and heterogeneous kinds of data that need to be integrated: clinical data, medical check data, various types of molecular information [2]. All these massively emerging amounts of data, accompanied with the new requirements for the purposes of preventive and personalized medicine, emphasize the great importance of data integration in cancer studies [10].

The semantic side of data integration in cancer studies gives a large horizon for using related data with different meaning and structure and makes possible to extract unknown knowledge about various aspects of the studied cancer case(s) [15]. The semantic integration of data from different types of cancer studies challenges the problem how to use all provided data aside with the opportunity to make proper associations of clinically controlled parameters and information based on “omics” data profiles [19, 22].

Technically the integration of cancer studies data can be supported by development of workflows and systems, capable for fast and accurate integration of disparate cancer data and enhanced models for cancer data analytics. Such systems allow constructing and executing queries at a conceptual level and in a way utilizing the available rich in semantics cancer information. Some existing tools for semantic data integration in cancer studies are based on utilization of external domain knowledge sources like Gene Ontology (GO). There are several popular RDF-based data integration platforms: Reactome [23], BioModels [24], BioSamples [25], Expression Atlas [5], ChEMBL [9], UniProt RDF [11]. These platforms can be used to search across datasets. For example, a query for gene expression data will integrate results from Expression Atlas with relevant pathway information from Reactome and compound-target information from ChEMBL. The structured data are available for download or can be queried directly.

This long list of platforms and resources can be used as good examples of linking various sources of domain knowledge. Our methodology goes much further aiming at the extraction of new knowledge from the results of integration of raw, unstructured and heterogeneous data.

5 Main Characteristics and Novelty of the Proposed Approach

The approach to data integration we suggest in this paper was originally oriented to a particular application – to unite data from real studies and treatments of neuroblastoma and breast cancer – but its design characteristics make it sufficiently generic and applicable in a wide range of subject areas. As a result of its application, different datasets are joined and the semantic integrity of the data is kept and enriched. In our particular case, through combining data from multiple sources (Fig. 3) we create a new network of data where entities, like proteins, clinical features and expression features, are linked with each other [26]. In this network, nodes represent patients and edges represent similarities between the patients' profiles, consisting of clinical data, expression profiles and CNV data. Such a network can be used to group patients and to associate these groups with distinct features. The main challenges here are: (1) building an appropriate linked data network, discovering a data model semi-structure [14] and mapping assertions by the applied model for data integration [27]; and (2) data cleaning, combined into a formal workflow for data integration.

We focus on two aspects of data integration: horizontal and vertical. As explained, horizontal data integration means combining data from different sources for the same entity. Entities can be identified in our particular datasets as clinical data, patients, expression profiles and CNV. Each kind of data is measured by a specific technology and available in various data formats. Groups of entities are semantically similar. Vertical data integration, on the other hand, is applied to creating relations between all horizontally integrated objects. This vertical data integration provides a connection between all different types of entities. This connection, in our case, covers relations between patients, expression profiles, clinical data and CNV data. Based on these relations we can easily detect all patients closely related to each other by protein mutations, diagnosis and therapy. We used different databases for horizontal and for vertical data integration. These different databases are required because horizontal and vertical data integration address different aspects of the integration problem. Data for the horizontal data integration are unstructured and heterogeneous. Thus, we use a document-oriented database, which can handle different data types and formats. For vertical data integration a graph database is used, as it is suitable for representing relations – crucial in this case. In this study, all relations are established between existing records for each entity, and represented by a semi-structure.

An integration model over a NoSQL database can potentially unite medical studies data, alternatively to the most frequently used statistical/machine learning methods. Most NoSQL database systems share common characteristics, supporting scalability, availability, flexibility and ensuring fast access times for storage, data retrieval and analysis [9, 11]. Very often when applying cluster analysis methods for grouping or joining data issues, small classes occur – mainly with outliers and mostly with data dynamically changing their relatedness. Applying our integration model we do believe that all these problems can be overcome. Moreover, we can extend the potential of the model by using multiple datasets, regardless of the level of heterogeneity, particular formats, types of data, etc. – all very specific for cancer studies.

6 Suggested Methodology

Our methodology for integration of unstructured data from the studied samples is based on the proper use of schema-less databases and domain ontologies like the Gene Ontology (GO) [28]. The data we are manipulating contain hidden relationships between the proteins provided from different patients within studies of both diseases (BC and NB). We use all available information about already built relationships in our data sources and try to find additional information in some third party sources to attain semantic integration of the data. Thus, step-by-step, we develop a network, which combines protein relations between patients and diseases. The challenge here is to store all relationships with their cycle dependencies. The latter are possible because one patient has relationships with mutated protein(s), mutated protein(s) has/have reference (s) to expression and other patients have references to the same protein(s).

In each disease (BC and NB) every patient has a different set of mutated or expressed proteins. Only small sets of mutated proteins are equal and exist in each patient. All proteins belong to families, which contain many related proteins. By application of semantic annotation and search techniques we aim to find and combine all proteins which are semantically related to the studied diseases. So, we can aggregate and discover all needed information for an enhanced number of related proteins.

The role of GO in our methodology is to provide a controlled vocabulary for annotating homologous gene and protein sequences in the studied cancers. GO classifies genes and gene products on the base of three hierarchical structures that describe a given entry's biological processes, cellular components and molecular functions, and organizes them into a parent-child relationship [22]. For the purposes of our study, an essential key are protein families and relationships between them. We associate all this information with the inferred relationships between patients and proteins to provide a complete schema of mutated and related proteins for each studied patient. In this way we demonstrate the possibilities of using appropriate subject knowledge bases (like GO) for the purpose of semantic integration of data from different sources. As a result we build a base of semantically linked data from biomedical research and create a framework supporting the practical extraction of new knowledge in life sciences and other significant areas (Fig. 3).

We developed a workflow for data integration in order to overcome the heterogeneity of the data and the enormous number of relations between the studied patients and proteins. First of all we are aiming to integrate datasets from the two cancer studies. In this line we are trying to find some relations within the given data as well as to generalize some information about commonly related proteins. This process invokes the semantic integration of the data, which is a key part of our study.

During the analysis of raw data we create a sort of "semi-structure" of the data – a structure containing only attributes, existing in each record. In semi-structured data, the entities belonging to the same class (protein mutated, expressed and CNV) may have different attributes even though they are grouped together, and the attributes' order is not important. Semi-structured data is becoming more and more prevalent, e.g. in structured documents and when performing simple integration of data from multiple sources. Traditional data models and query languages are inappropriate, since semi-structured data often is irregular: some data is missing, similar concepts are represented using different types, heterogeneous sets are present, or object structure is not entirely known [13].

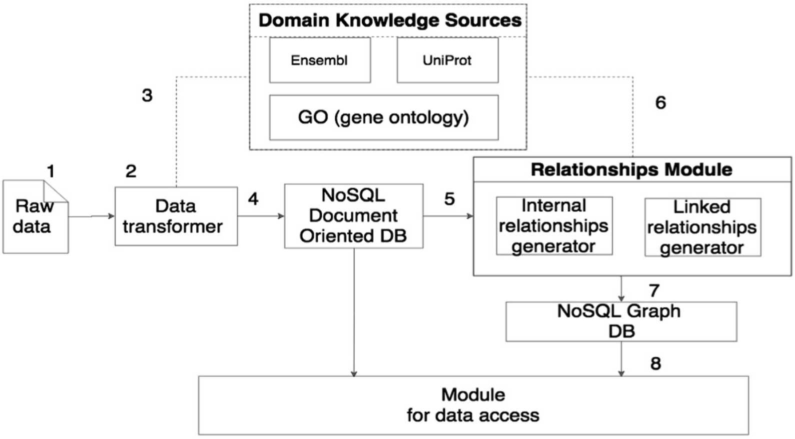

Fig. 3. Workflow of semantic data integration.

Our workflow covers eight stages (Fig. 3). The first two stages correspond to data generated from different high-throughput technologies – microarray and RNA-Seq data (1), preprocessed (2) for databases generation. In this step we aim to complete the missing data (3), to define data for a patient, to generate the "semi-structured data" and to store it. Raw and "semi-structured" data are stored in a document-oriented NoSQL database (MongoDB) with all attributes (4). At the next stage (5) we try to find relationships between proteins, patients and diseases. For the enrichment of our dataset we try to find additional protein relationships (6) in a set of external domain knowledge sources and to generate semantic links to them. These enriched protein relations are traced with other proteins and the discovered relations are explored. We use in this case the Neo4J database. In this graph database we store only relationships between entities (7). Each relation contains a "semi-structure" with ID from MongoDB. Module 8 is intended for user access to data.

Starting point (1) in our workflow on Fig. 3 is the raw data entry, read in different formats (CSV, TXT, XML, etc.). We analyze each record and create a general data structure ("semi-structure"). The meaning of this semi- structure plays a key role in the semantic integration of data. The process of generating the semi-structure is dynamically changed after reading of each record. This parallelized process finishes by reading of all records. Data from different sequencing technologies has different attributes for proteins. In order to unify these attributes, we use several external domain knowledge sources (EDKS): Ensembl [21], UniProt [29], and GO, which provide additional knowledge about the existing annotated proteins. We search for proteins and their annotations in these EDKS by generating a request by URL. This URL is stored in the protein records in the document-oriented NoSQL database, which is schema-less and can store both structured and unstructured data of the same record. Each record contains different number of attributes.

The next step is to find relationships and to connect our data to other linked data from the explored EDKS in order to generate more relationships between the proteins. According to [16], "linked data is a set of design principles for sharing machine-readable data on the WWW for use by public administrations, business and citizens". Linked data is a methodology for representing structured data so that it can be interlinked and become more useful through semantic queries. It allows data from different sources to be combined and used together when executing queries for semantic search and information retrieval.

For the purposes of providing tools for flexible formulation and adequate execution of semantic search queries we provided two types of relationships: ones, extracted from the studied raw datasets, and others, extracted from the domain knowledge sources like GO. All relationships between all entities (proteins) in the semi-structured data are more than 1,000,000,000. The goal here is to store all relationships with their cycle dependencies. Cycle dependencies are possible and expected because one patient has relationships with mutated protein(s), the mutated proteins have reference to expression and the same patient has relation to this expression.

Our approach provides a mechanism to create relationships between all patients on the basis of the mutations, expressions and CNV. We use a graph database (Neo4j) to manage and store all known relationships. In Neo4j we insert only the "semi-structure" for each related entity. This "semi-structure" contains a 128-bit identifier, and by using it we create and manage the relations between MongoDB and Neo4j. This is necessary also because full data attributes are stored in MongoDB which cannot provide such relationship management.

After relationships are generated, we generate the new relations via Ensembl and UniProt using GO, because GO by itself does not provide suitable programmable access. These sources contain information about proteins, protein families and relationships between them. We take this information and store the links of similar proteins as relationships in our database (Neo4j). This approach provides a method for creation of co-relationships between proteins. The relations generated by dint of the used external sources are assigned as unreliable. They are considered at this stage as unreliable because in deep search all proteins by one or another way are related. In the working process, if one unreliable relationship is "accessed" by the same proteins multiple (more than 10) times, such relationship is transformed as a normal (or trusted) one [30]. Therefore, we dynamically create reliable relationships between proteins in the workflow. For example, we use the "similar proteins" section of UniProt to extract the related proteins. This section provides links to proteins that are similar to the protein sequence(s) described in the search query at different levels of sequence identity thresholds (100%, 90% and 50%) based on their membership in UniProt Reference Clusters.

At the final step, we create a workflow module for sending queries to the involved databases. So the workflow can produce pathways, predict relationships between proteins, patients and diseases and thus discover new knowledge about the protein relationships based on the semantic integration of data and external domain knowledge sources. Our approach enables users to formulate and execute a wide spectrum of dynamically constructed semantic search queries.

7 Results and Discussion

An essential part of our methodology is based on the use of the validation mechanisms provided by MongoDB. The latter are applied to create a specific "validation filter" for all attributes in the semi-structure. In our case these filters are set with minimal level which guarantees that there will be a value for each attribute and thus our semi-structure cannot contain empty attribute fields. On the other hand, our approach is consistent with the fact that document-oriented databases are not appropriate for storing such related data. For storage and management of all relationships we use a graph database – Neo4j. Graph databases (GDB) provide a suitable environment for developing and managing relations between the entities (proteins, patients). GDB have native solutions for management of complex cycling dependencies (relations). As already mentioned, our data imply many cycling references between patients and proteins and it is necessary to have a trusted path for each relation between patient and protein. Methodologically this problem is solved in graph theory by trusted relational trees [31], and we use this solution with Neo4j.

It is possible for each relation in Neo4j to insert additional information, containing the relation type, IDs etc. We use this to connect our semi-structured data to the linked data via dynamically generated URLs, which refer to external domain knowledge sources for more detailed annotations (protein, protein family).

The mentioned EDKS are accessed via specific APIs based on HTTP/S protocol realized by RESTFul methodology – a software architectural style that defines a set of constraints to be used for creating Web services. For example, for RERE protein (a protein which is related to apoptosis triggering) we generate the following URL [12]: https://www.ebi.ac.uk/proteins/api/proteins?offset=0&size=100&gene=RERE.

Based on the answer from the EDKS we dynamically create new relationships between all references of the searched protein. This enhances the network where indirectly all proteins are related. Such type of indirectly generated relationships has a small score in the query answer. Initially they are accounted as untrusted. Their score is generated dynamically and depends on the number of requests. Based on this score we rank the respond to the user who has posed the request. Automatically these relationships become trusted after a certain level of score value (Fig. 4).

Obviously the used EDKS have some limitations. GO provides no interface for programming access to it and the only way to use directly the ontology is to download it. That is why we use Ensembl to access GO as an internal resource. Initially, we create a request to EBI. The returned information from EBI contains reference IDs to other knowledge sources. We use the particular reference ID for GO and create a request to Ensembl for getting ontology-based relationships for the requested protein. For tracing the relationships, we build the following request:
https://rest.ensembl.org/ontology/ancestors/GO:0003677?content-type=application/json.

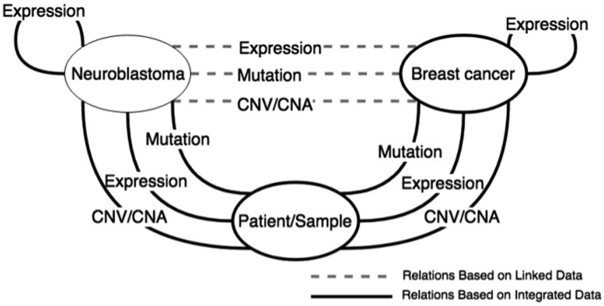

Fig. 4. Relationships between the proteins.

Here GO:0003677 is the reference ID from the first request to EBI. If information about this protein does not exist in GO, the same procedure is repeated to find it in another database (InterPro), etc. As a consequence of such intelligent integration of the two studied cancer datasets, we discover new knowledge about the mutated proteins, related to those two types of cancer. We suppose that following this approach, based on semantic data integration, it is possible to discover a unique set of proteins and their functional annotations for a particular cancer.

As an example for the provided opportunities, we can demonstrate how to find relations between the proteins RERE, EMP2 and KLF12. These three proteins are related to neuroblastoma and breast cancer. In both diseases they are mutated. To find other related proteins we create a request to our system using GraphQL – a query language for APIs, and a server-side runtime for managing and executing typified queries. The result is illustrated on Fig. 5 where proteins are shown as circles (graph nodes). The size of each node is based on its resulting score. Proteins are connected by different types of relationships shown on the figure in different shades of grey and different style.

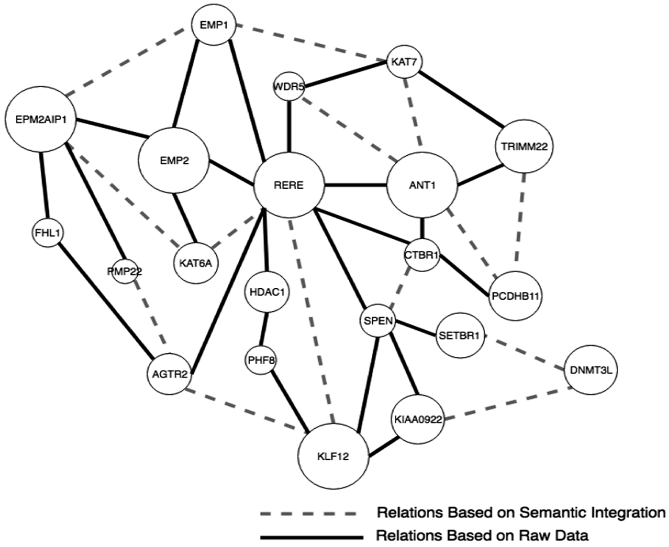

Fig. 5. Semantic data integration by use of linked data URLs.

Found proteins have two types of relations. If the relation is generated from linked data (as URLs), it is classified as non-reliable. Many relations are generated from linked data in external sources and marked by a score. This score is generated dynamically and depends on the number of requests to our system. Based on this score we rank the response to the user who posed the request. If a found relation is built on the raw dataset then it has a higher score than the ones built from linked data. The score of such relation is increasing and after a number of requests it can be classified as a trusted one. In such manner our workflow enables automated enrichment of protein relationships and extraction of new knowledge related to both cancers.

8 Conclusions

In this paper we discuss an original methodology for extraction of hidden relations in an integrated dataset by combining data from disparate sources and consolidating them into a meaningful and valuable information pool by the use of semantic technologies. The use of linked and overplayed NoSQL database technologies allowed us to aggregate the non-structured, heterogeneous cancer data with their various relationships. The applied semantic integration of different cancer datasets has an obvious merit concerning the enrichment of the studied data by discovery of mutual internal relations and relations with external domain knowledge sources.

A novel approach for automated semantic data integration has been proposed and analyzed. It provides means, supporting augmented and precise enough search for hidden common (protein) relations. The discovery of these hidden common proteins and joining their functionality is in fact an extraction of new knowledge about the studied cancer cases. All these methodological procedures are built as a workflow, based on NoSQL databases and exploring external domain knowledge sources for the purposes of efficient integration of data from cancer studies.

This study shows also that using customized analysis workflows is a necessary step towards novel discoveries and potential generalization in complex fields like personalized therapy.

Acknowledgements. This research is supported by the National Scientific Program eHealth in Bulgaria.

References

1. Stein, L.: Creating a bioinformatics nation. Nature **417**, 119–120 (2002). https://doi.org/10.1038/417119a
2. Kashyap, V., Hongsermeier, T.: Can semantic web technologies enable translational medicine? In: Baker, C.J.O., Cheung, K.H. (eds.) Semantic Web, pp. 249–279. Springer, Boston (2007). https://doi.org/10.1007/978-0-387-48438-9_13
3. DeSantis, C., Ma, J., Goding, S., Newman, L., Jemal, A.: Breast cancer statistics, 2017, racial disparity in mortality by state. Cancer J. Clin. **67**(6), 439–448 (2017)
4. American Cancer Society: Cancer Statistics Center. http://cancerstatisticscenter.cancer.org. Accessed 23 Mar 2019
5. Kapushesky, M., et al.: Gene expression atlas update—a value-added database of microarray and sequencing-based functional genomics experiments. Nucleic Acids Res. **40**(D1), D1077–D1081 (2012)
6. Lenzerini, M.: Data integration: a theoretical perspective. In: PODS 2002, pp. 233–246 (2002)
7. Sioutos, N., et al.: A semantic model integrating cancer-related clinical and molecular information. J. Biomed. Inform. **40**(1), 30–43 (2007)
8. Louie, B., et al.: Data integration and genomic medicine. J. Biomed. Inform. **40**(1), 5–16 (2007)
9. Gaulton, A., et al.: ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. **40**(D1), D1100–D1107 (2012)
10. Ruttenberg, A., et al.: Advancing translational research with the semantic web. BMC Bioinform. **8**(Suppl. 3), S2 (2007)
11. Nicole, R.: UniProt in RDF: Tackling Data Integration and Distributed Annotation with the Semantic Web. Nature Precedings (2009). http://dx.doi.org/10.1038/npre.2009.3193.1
12. Jupp, S., et al.: The EBI RDF platform: linked open data for the life sciences. Bioinformatics **30**(9), 1338–1339 (2014)
13. Beeri, C., Milo, T.: Schemas for integration and translation of structured and semi-structured data. In: Beeri, C., Buneman, P. (eds.) ICDT 1999. LNCS, vol. 1540, pp. 296–313. Springer, Heidelberg (1999). https://doi.org/10.1007/3-540-49257-7_19
14. Zhang, H., Guo, Y., et al.: Data integration through ontology-based data access to support integrative data analysis: a case study of cancer survival. In: Proceedings of IEEE International Conference on Bioinformatics Biomed, pp. 1300–1303 (2017)
15. Pittman, J., et al.: Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. PNAS **10**(22), 8431–8436 (2004)
16. Berners-Lee, T.: Linked Data (2006). http://www.w3.org/DesignIssues/LinkedData.html. Accessed 23 Mar 2019

17. Oren, E., et al.: Sindice.com: a document-oriented lookup index for open linked data. J. Metadata Semant. Ontol. **3**(1), 37–52 (2008)
18. Famili, F., Phan, S., Fauteux, F., Liu, Z., Pan, Y.: Data integration and knowledge discovery in life sciences. In: García-Pedrajas, N., Herrera, F., Fyfe, C., Benítez, J.M., Ali, M. (eds.) IEA/AIE 2010. LNCS (LNAI), vol. 6098, pp. 102–111. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13033-5_11
19. Holford, M., et al.: A semantic web framework to integrate cancer omics data with biological knowledge. BMC Bioinform. **13**(Suppl. 1), S10 (2012)
20. Glaab, E., et al.: Extending pathways and processes using molecular interaction networks to analyse cancer genome data. BMC Bioinform. **11**, 579 (2010)
21. Yates, A., et al.: Ensembl 2016. Nucleic Acids Res. **44**(Database issue), D710–D716 (2016)
22. Zhang, H., et al.: Data integration through ontology-based data access to support integrative data analysis: a case study of cancer survival. In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, pp. 1300–1303 (2017)
23. Fabregat, A., et al.: The reactome pathway knowledgebase. Nucleic Acids Res. **46**(Database issue), D649–D655 (2018)
24. Chen, L., et al.: BioModels database: an enhanced, curated and annotated resource for published quantitative kinetic models. BMC Syst. Biol. (2010). https://doi.org/10.1186/1752-0509-4-92
25. Gostev, M., et al.: The BioSample Database (BioSD) at the European bioinformatics institute. Nucleic Acids Res. **40**(D1), D64–D70 (2012)
26. Semantic Web Health Care and Life Sciences (HCLS) Interest Group. https://www.w3.org/blog/hcls/. Accessed 23 Mar 2019
27. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool (2011)
28. The gene ontology consortium: gene ontology consortium: going forward. Nucleic Acids Res. **43**(D1), D1049–D1056 (2015)
29. The UniProt Consortium, Bateman, A., et al.: UniProt: the universal protein knowledgebase. Nucleic Acids Res. **45**(Database issue), D158–D169 (2017)
30. Golbeck, J., Parsia, B.: Trust network-based filtering of aggregated claims. Int. J. Metadata Semant. Ontol. **1**(1), 58–65 (2006)
31. He, P.: A new approach of trust relationship measurement based on graph theory. Int. J. Adv. Comput. Sci. Appl. **3**(2), 19–22 (2012)