

Unknown, Uncertain, Untrue: Challenges in Inference Using Semantic Life Science Data

Bence Bruncsics, Andras Gezsi, Peter Antal

Budapest University of Technology and Economics

Department of Measurement and Information Systems

Budapest, Hungary

Email: {brunscics, gezsi, antal}@mit.bme.hu

Abstract—The rapidly accumulating electronically available data and knowledge in life sciences currently cannot be integrated into a general, unified knowledge base with adequate inference. A unified knowledge representation is still missing, which could include large scale uncertain expert knowledge and *in silico* predictions from machine learning. Thus, the available information is fragmented and differently processed. In order to overcome this barrier and provide a common base for automatic data processing many databases became available in a semantic form. However, the current query languages for semantic linked open data have many limitations, therefore they are not sufficient for creating an integrated knowledge representation and reasoning. Overcoming this obstacle there are different solutions for knowledge representation but currently no such tool has the power to predict or find solutions for the most important biological questions. Combination of semantic web technologies and probabilistic techniques can result a novel powerful tool for life sciences.

Index Terms—graph databases, intelligent systems, probabilistic knowledge representation, semantic Web, uncertain databases

I. INTRODUCTION

Besides to regular free-text publication an emerging trend is to make accessible data and knowledge in representations more suitable for automated processing. Amongst databases and information sources semantic technologies have the widest range covering most of the life science entities and properties with additional information about their relationships.

A. Semantic Solutions

Providing an efficient foundation for constructing, storing and interpreting life science data is challenging due to its size, complexity and the excessive need for an easily accessible way to process these data. A candidate solution originates from World Wide Web Consortium in form of semantic technologies providing such scalable, efficient, accessible and simple standards as the RDF (Resource Description Framework) [1] and the OWL (Web Ontology Language) [2].

In practice the different biological entities like diseases, genes or compounds are handled in different databases and cross-domain integration is ensured by unique identifiers, URIs (Uniform Resource Identifier) of these entities. The URIs form a consistent system, because they are usually URLs pointing to locations maintained by the databases creators or by projects dedicated to RDFize multiple databases, technically they are

formed by proper database specific prefixes and IDs for each entity.

There are numerous properties for most entities in the databases and in RDF these are connected to the entities using OWL based terms as a form of knowledge representation e.g. genes have a location property describing a position in the DNA where the gene can be found. In addition to properties from the databases, there is information about the relationship between entities from different databases like a gene is associated with a disease. The associations can have further properties like which mutation of a gene is associated with a disease, resulting in complex relationships between databases.

B. Linked Open Data and Semantic Databases

Data sharing became a central topic in life sciences, balancing aspects of privacy, proprietary, scientific advance and repeatability. To contribute to the semantic Web, the data needed to be RDFized. It is usually done by database providers such as PubChem [3], but more commonly separate organizations perform this conversation, such as the European Bioinformatics Institute (EMBL-EBI) [4], or collaborative projects such as Ontobee [5]. Semantic data has multiple locations, different origins and can even have duplicates, therefore its proper management is essential for its applicability. The Linked Open Data approach (Fig. 1) is a large scale community project aimed to set RDF links between the open RDF datasets and to encourage the community to RDFize the separate databases expanding the semantic Web [6].

Currently the most relevant semantic life science databases contain thousands ($10^3 - 10^4$) of diseases, over a million genes from different species, and millions ($10^6 - 10^8$) of compounds and many more databases and descriptions with billions of links between entities [7].

II. BACKGROUND KNOWLEDGE OF LIFE SCIENCES

In life sciences, many revolutionary measurement technologies have emerged in the last three decades, producing unprecedented amount of data and many databases were formed to store and organize these data. Further databases and ontologies were made to organize the already existing current information resulting in a large variety of information sources. Most of the databases are formed along various aspects of

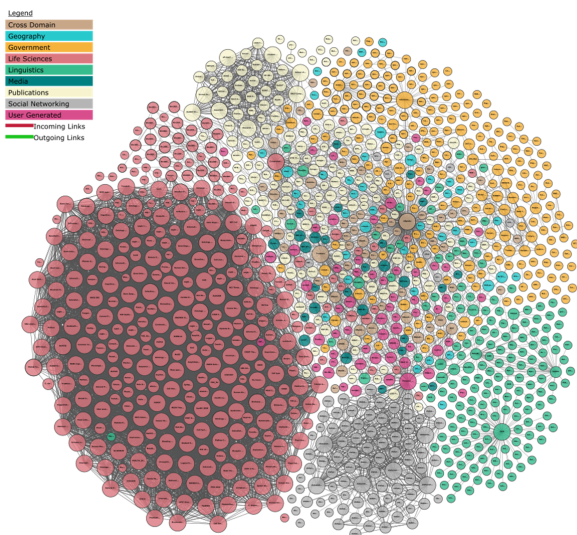


Fig. 1. Linked Open Data (LOD): Life sciences contribute the most data for LOD [8]

drugs, genes, and diseases, such as genetic, proteomic, phenotypic or chemical information. The most important databases are focusing on listing all the entities at a given level (such as all the genes) and usually contain additional descriptions of their entities via common properties. Other databases describe ontologies and refer to other entities affected by the ontologies; and there are databases focusing on linking databases based on specific techniques like gene expression profiles in different diseases (based on microarray technology).

However, inherent challenges of representing scientific information is still not addressed by the semantic framework, such as the (1) representation of large scale implicit, negative information (cf. closed world assumption), (2) representation of uncertain information (3) and representation of inconsistent information, partly as a consequence of untrue scientific reports from misconducted or fraudulent studies.

A. Biases for known and unknown

Life sciences always had biases due to scientific paradigms, trends, measurement constraints, budget limitations or publication and grant policies. For example, the publication bias for positive findings, is a serious challenge for text-mining, machine learning and automated discovery systems.

The need for publication and gathering citations creates trends in science resulting popular fields but also popular entities such as popular genes like CD4 gene which plays a role in HIV infections but less important for the general understanding of biology [9]. These trends result in disparate distributions and characteristics for the different entities and the popularity of a field or entity is not necessarily a consequence of the biological relevance.

The effect of the uneven mapping stands out more when only a small portion of field is discovered which is the case in most part of the life sciences. For an integrating model in drug

discovery the chemical data should contain information about the targets (usually proteins) of the compounds. Knowing this information, a matrix can be created containing the compound target interactions, but in practice the coverage of these matrices can be as low as 0.1% (a standard bioactivity database ChEMBL contains 14 million values including duplicates for the 2 million x 10 thousand interactions of its compounds and targets, [10]). Knowing that the data in these sparse matrices are unevenly distributed rises further challenges for processing this information.

B. Uncertain scientific knowledge

Statistical data analysis provides the base for our empirical science and evidence-based medicine, however the limits and conditions of statistical test driven scientific publication policies are often overlooked. Even if we assume honesty for all researchers, it is easy to make mistakes (like excluding seemingly wrong parts from data) without a statistical mindset, not knowing the consequences of certain actions. And unfortunately, experiments with positive results will have higher chance to be accepted than those where the methods were rigorously followed, but had negative results.

Further problem is the mystified requirement and unconditional acceptance of p-values as ultimate proofs. In the field of biology, a result showing a p-value lower than 0.05 is accepted as significant effect, even without considering its power and the number of the corresponding experiments. Often the number of trials or the parallelly tested theories are not included in the statistics or sometimes these are not even published resulting in an overwhelmingly large number of false positive results. In summary, currently the representation of certainty and the context of a statistical evidence is severely restricted.

C. Untrue, Non-repeatable Scientific Reports

It is a suspected, but still astonishing finding that at least 50% of published studies of early-stage venture capital firms cannot be repeated with the same conclusions by an industrial lab [11]. Furthermore, a large-scale study by Bayer in 2011 showed that only 20-25% of their in-house findings were completely in line with the original publications [12]. Therefore, it is clear that the databases based on publications with overwhelming amount of false positive findings will have the same biases.

III. REASONING USING LIFE SCIENCE DATA

The need for automation of reasoning using life science data is clear, because it could overcome many questions that the current state of science wouldn't be able to answer otherwise, but cooperative schemes with scientific discovery support systems can also accelerate research and increase efficiency.

A. Challenges in Reasoning

Despite rapid advancements in deep linguistic analysis, using solely free-text scientific publications is not a viable option currently for reasoning in life sciences, therefore the

manually curated, computationally processable semantic data remains a default option. However, the usage of semantic data and technologies is hindered by many obstacles:

1) *Vulnerable distributed inference*: Life science databases are usually not integrated into a common datastore; therefore, distributed queries are strongly affected by the accessibility of the sources; even one missing endpoint could fail the query.

2) *Intractable inference*: The computational complexity of unrestricted queries is so high, that usual inference is intractable even in a problem-specific unified database using dedicated servers.

3) *Expert queries*: Query languages provide a wide range of logical and calculation methods, but these can be rather complicated, especially using negation for filtering.

4) *Uncertain reasoning*: There is no simple automated way to construct uncertain evidence over different entities and combine uncertain evidences, especially using further quantitative measures as well, such as similarity measures, which are popular in biomedicine.

5) *Transparency*: Checking the inference paths and calculation in case of complex queries requires deep understanding of the underlying techniques and it is time consuming. Hence, user-level interfaces are essential to translate the complex inference processes for a domain-expert scientist to be able to recognize the significance of a given inference.

There are numerous bioinformatics tools for the large-scale fusion of heterogeneous data and knowledge, but these tools either support the general, non-quantitative (logical) inference over semantic resources or focus on a dedicated task, such as gene prioritizers [13], [14]. Further problem in the current approaches is the lack of support for high-dimensional, partial, noisy evidences and deep control for the inference process.

IV. TOWARDS QUANTITATIVE SEMANTIC FUSION

To cope with these deficiencies, the Quantitative Semantic Fusion (QSF) Framework was developed at the ComBineLab, Department of Measurement and Information Systems. Epistemological aspects of linked open life science data can be approached in QSF as follows:

1) *Predicting the unknown*: There are techniques to complete sparse matrices based on the available data creating a better foundation for reasoning [15]. Furthermore, the used similarities and properties could be used to approximate probabilities for the filled or even for the original data.

2) *Probabilistic representation of uncertainty*: Semantic life science databases often formed around techniques like chemical screening or microarrays providing access to the sources. The statistical parameters of the data are characteristic to the techniques, therefore it is possible to calculate or approximate Bayesian probabilities for these entities.

3) *Representing trust*: There are dedicated solutions for knowledge representation such as Evidence Code Ontology (ECO) [16] in bio-medicine, providing information about the data source (e.g. motility assay evidence) or HELO [17] which is more suitable for probabilistic knowledge representation. Furthermore, even textual information can be translated into

probabilistic graphical models and using approximations for the most probable explanations creating link between the available data and probabilistic techniques.

The QSF system offers the following solutions for the specific challenges.

1) *Integration* For a general biological model the integrative functions are necessary, therefore the essential disease, genetic, protein, pathway and substance information must be included. Based on linked open data Chem2Bio2RDF [7], the QSF framework is able to integrate and perform inference using the relevant information sources. Furthermore, it is possible to expand the framework with further sources.

2) *Accessibility and scaling* To cope with computational complexity, the relevant data sources are stored and processed in one server. Due to RDF compression techniques and proper selection of the relevant properties, entities and links the overall size of a minimal, but representative data covering the a given problem in life sciences remains in the few GB range.

3) *Filtering* SPARQL cannot handle well negation and complex filtering, but in the RDF-based QSF framework complex logic can be applied to focus and control the inference process.

4) *Transparency* Large-scale data visualization techniques like Cytoscape [18] allow users to create and explore graphs, which are natural, transparent, intuitive way for representing biological data. The QSF allows the export of the results of evidence propagation as paths in knowledge graphs, which provides easy access and interpretation for biologists.

5) *Bayesian inference* Using probabilistic methods, quantitative evidences can be constructed for the input queries based on the sources. For example, for a drug candidate the acceptance rate based on the trial number is a possible approximation for a drug-disease association. The QSF allows the approximation of a Bayesian inference technique. Many bioinformatics tools are able to incorporate and provide p-values, but QSF provides a full-fledged approximation for a Bayesian inference.

6) *Bayesian fusion* Integrating multiple sources or using data from similar, repeated experiments have significant advantages, due to the cancellation effect of method or setup specific biases. Calculating pseudo-Bayesian posteriors for such data can result a significant improvement in case of such an uncertain space as the life sciences [19].

V. APPLICATION OF QUANTITATIVE SEMANTIC FUSION

Translation of questions with complex biomedical background to a formal query is often difficult and require special technical knowledge. To support this process, the graphical user interface of the QSF endowed with the following functionalities:

1) *Selecting targets*: Having computationally approachable targets (i.e. answers) is essential for constructing a query,

which in itself can be seen as a modeling activity. In QSF, evidences are entered, then propagated throughout the network, thus selection of targets does not influence the results, but essentially influence interpretability.

2) *Context-sensitive inference*: Planning which sources and paths should be taken into account in the inference can be controlled by the query and the results can be affected by them.

3) *Collecting inputs*: The proper input is key for inference and it is recommended to collect as much data as possible from different sources to overcome the biases of the data by canceling out the random effects.

4) *Converting evidences*: Providing probabilities for the inputs is challenging in many cases, and recommendations or analytic solutions are needed for this step.

5) *Applying conditions*: Filtering is essential within the inference, e.g. via negation even complex questions can be asked.

6) *Interpreting the results*: It is essential to check the paths of the evidences to find possible anomalies and to get a better understanding of the dominating effects in behind the results.

VI. CONCLUSION

Automated reasoning using life science data is challenging due to the fragmentation of the data located in separate databases, but fortunately semantic web technologies and the Linked Open Data approach provide sufficient background to access these data. However, current approaches are very limited in reasoning efficiently based on this knowledge: (1) the native usage of semantic data as inference in graph database lacks the ability of to incorporate uncertainty, (2) network approaches using diffusion-based inference methods lack semantic control within the inference, (3) kernel fusion based prioritization methods cannot directly manage relational data. Additionally, biases of life science data complicate the problem further, because the sources are uncertain, incomplete and unevenly distributed. Our group developed a system, approximating a full-fledged probabilistic inference, which is grounded in an underlying semantic graph database. Furthermore, we developed a graph-based query language to support the translation of complex biomedical queries, allowing semantic influence over the inference process. I developed multiple models and manually evaluated various inference schemes in this system to refine existing applicability and explore new directions for its development [20]. Tools integrating different information sources, handling high-dimensional weak evidences, providing semantic control for the inference and supporting the interpretation of the results, such as the QSF framework, could provide new possibilities for cooperative reasoning of experts and machines.

ACKNOWLEDGMENT

The author would like to thank the advices and guidance of Peter Antal, and the continuous help of Andras Gezsi and Gabor Guta. The research has been supported by the European Union, co-financed by the European Social Fund (EFOP-3.6.2-16-2017-00013).

REFERENCES

- [1] S. Decker, S. Melnik, F. Van Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, and I. Horrocks, "The semantic web: The roles of xml and rdf," *IEEE Internet computing*, vol. 4, no. 5, pp. 63–73, 2000.
- [2] S. Bechhofer, "Owl: Web ontology language," in *Encyclopedia of Database Systems*. Springer, 2009, pp. 2008–2009.
- [3] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker *et al.*, "Pubchem substance and compound databases," *Nucleic acids research*, vol. 44, no. D1, pp. D1202–D1213, 2015.
- [4] S. Jupp, J. Malone, J. Bolleman, M. Brandizi, M. Davies, L. Garcia, A. Gaulton, S. Gehant, C. Laibe, N. Redaschi *et al.*, "The ebi rdf platform: linked open data for the life sciences," *Bioinformatics*, vol. 30, no. 9, pp. 1338–1339, 2014.
- [5] E. Ong, Z. Xiang, B. Zhao, Y. Liu, Y. Lin, J. Zheng, C. Mungall, M. Courtot, A. Ruttenberg, and Y. He, "Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration," *Nucleic acids research*, vol. 45, no. D1, pp. D347–D352, 2016.
- [6] L. Yu, "Linked open data," in *A Developers Guide to the Semantic Web*. Springer, 2011, pp. 409–466.
- [7] B. Chen, X. Dong, D. Jiao, H. Wang, Q. Zhu, Y. Ding, and D. J. Wild, "Chem2bio2rdf: a semantic framework for linking and data mining chemogenomic and systems chemical biology data," *BMC bioinformatics*, vol. 11, no. 1, p. 255, 2010.
- [8] A. Abele, J. McCrae, P. Buitelaar, A. Jentzsch, and R. Cyganiak, "Linking open data cloud diagram (2017)," 2017.
- [9] R. Hoffmann and A. Valencia, "Life cycles of successful genes," *TRENDS in Genetics*, vol. 19, no. 2, pp. 79–81, 2003.
- [10] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte *et al.*, "The chEMBL database in 2017," *Nucleic acids research*, vol. 45, no. D1, pp. D945–D954, 2016.
- [11] L. Osherovich, "Hedging against academic risk," *SciBX: Science-Business eXchange*, vol. 4, no. 15, 2011.
- [12] F. Prinz, T. Schlange, and K. Asadullah, "Believe it or not: how much can we rely on published data on potential drug targets?" *Nature reviews Drug discovery*, vol. 10, no. 9, pp. 712–712, 2011.
- [13] L.-C. Tranchevent, A. Ardeshirdavani, S. ElShal, D. Alcaide, J. Aerts, D. Auboeuf, and Y. Moreau, "Candidate gene prioritization with endeavour," *Nucleic acids research*, vol. 44, no. W1, pp. W117–W121, 2016.
- [14] G. Valentini, A. Paccanaro, H. Caniza, A. E. Romero, and M. Re, "An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods," *Artificial Intelligence in Medicine*, vol. 61, no. 2, pp. 63–78, 2014.
- [15] B. Bolgár and P. Antal, "Vb-mk-lmf: fusion of drugs, targets and interactions using variational bayesian multiple kernel logistic matrix factorization," *BMC bioinformatics*, vol. 18, no. 1, p. 440, 2017.
- [16] M. C. Chibucos, C. J. Mungall, R. Balakrishnan, K. R. Christie, R. P. Huntley, O. White, J. A. Blake, S. E. Lewis, and M. Giglio, "Standardized description of scientific evidence using the evidence ontology (eco)," *Database*, vol. 2014, 2014.
- [17] L. N. Soldatova, A. Rzhetsky, K. De Grave, and R. D. King, "Representation of probabilistic scientific knowledge," *Journal of biomedical semantics*, vol. 4, no. 1, p. S7, 2013.
- [18] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome research*, vol. 13, no. 11, pp. 2498–2504, 2003.
- [19] M. A. Province and I. B. Borecki, "Gathering the gold dust: methods for assessing the aggregate impact of small effect genes in genomic scans." in *Pacific Symposium on Biocomputing*, vol. 13, 2008, pp. 190–200.
- [20] B. Bence, "Large-scale data and knowledge fusion in aging research," Scientific Student Conference, 2017, 2017.