A new database for drug-discovery address key-issues in mining of knowledge

Ole Kristian Ekseth and Svein-Olav Hvasshovd

Department of Computer Science (IDI)
NTNU, Trondheim, Norway

Abstract. The life of individuals are strongly influenced by their health. An example concerns salinity resistant plants, an invention which may alleviate issues of climate change and rising sea levels. A different issue conserns drug discovery for humans, such as accurate and inexpensive cures available for the poor, personalized drugs, etc. In drug discovery the applied strategy is to combine domain experts with data made accessible through off-the-shelf software, and from the latter expect to identify new drugs. While computational drug-discovery is known to be working when number of candidate-factors are sufficiently small, the established methods and software are unfeasible for mining in big-data knowledge bases.

In this paper we address the above issue. We present an holistic approach for searches in big-data with complex relations. We demonstrate how our novel strategies for integration of large heterogeneous datasets results in knowledge discovery. In our work we address issues of: *semantics*, *entity similarity*, *clustering*, *data-engine*, *hypothesis testing*, and *user-interfaces*. To verify our approach we implement data from 37 external data-resources, resulting in a database with more than 30 million bio-medical relationships. When we compare our findings with existing literature we observe how our holistic approach for big-data mining discover 1000+ novel candidates for drug interaction. To address key-issues in knowledge discovery we have constructed 10+ new software-approaches for data-mining, tools which enable the development of a new method for mining of big-data. To enable reuse of our approaches, they are available from: `http://www.knittingTools.org/`, `http://www.knittingTools.org/gui_lib_mine.cgi`, `https://bitbucket.org/oekseth/mine-data-analysis/downloads/`, and `https://bitbucket.org/oekseth/hplysis-cluster-analysis-software`.

1 Introduction

In life-science a recurring task is to understand how and why entities relate: to construct a hypothesis which translates discrete observations into a conceptual figure capturing core-traits of an evaluated subject, as exemplified in Fig. 3 for the research of [1]. An example concerns the effects of *Cytoplasmic Phospholipase A2 (cPLA$_2$)* enzyme which is associated to a number of diseases, such as

Alzheimer [2] and *Rhematism* [3]. In knowledge-discovery researchers use manual approaches to identify candidate interactions, as exemplified in [4] where the authors use literature to manually construct a "heterogeneous network with 351 node" [4].

In contrast to established approaches for data-mining, an understanding of drug-interactions require the analysis of possible interactions, as exemplified in Fig. 1. While the "PubMed" database [5] contains "more than 27 million citations for biomedical literature" [5], the "Unified Protein Resources (UniProt)" describes more than 47 million protein sequences [6]. "The Economist" asserts that 50 per-cent of published research-literature are erroneous [7], hence the established use of manually selected research findings to identify new drug candidates is challenging.

The high cost of drug development discourage the development of drugs for the poor [8]. The cost of developing a single drug vary from $802 million to $2.2 billion [9]. The drug-company of "AstraZeneca" spend on average $11+ billion ([10,11]) on each accepted drug. The main-cost of drug development is the number of failed drugs [12], *e.g.*, as observed by [13]: "only, one in 5,000 medicines makes it to the marked" [14]. Of importance is to address the above issues in drug discovery, *i.e.*, as "today's pharmaceutical industry cannot sustain sufficient innovation" [15] with today's cost of drug-development. Hence, the importance of accurate tools for knowledge discovery.

In this paper we relate the above perspectives, demonstrating how a new holistic approach for mining of big data enable user-interactive drug discovery. In the method and associated software we unify the approaches of *user-centric* and *software-centric* approaches for data-mining, as depicted in Fig. 5. What we assert is that an holistic approach which increase accuracy and performance of *data from disparate sources*, *software for mining*, and *tacit understanding*, is sufficient to address major issues in drug discovery, a view supported by [15]. "R&D efficiency represents the ability of an R&D system to translate inputs (for example, ideas, investments, effort) into defined outputs (for example, internal milestones that represent resolved uncertainty for a given project or product launches), generally over a defined period of time" [15]. The ensemble of methods and software, summarized in Fig. 5, address challenges which have prevented established semantic data-bases from knowledge discovery, *e.g.*, as observed with respect to the issues encountered by [17,**?**,19].

In the work we have identified and addressed the issues of:

1. Disparate data: automatic approaches to unify distinctively different data, where results are exemplified in Fig. 1.
2. Execution-time: high-performance software for accurate and large-scale data-mining, as exemplified in Fig. 4;
3. User searches: interactive real-time data-mining which stimulate use of tacit knowledge, as exemplified in Fig. 3 and Fig. 6.

The remainder of the paper is organised as follows. In section 2 we briefly survey related approaches, before we in section 3 describe the approach. In the result-section 4 we identify evaluate/discuss how the holistic approach address

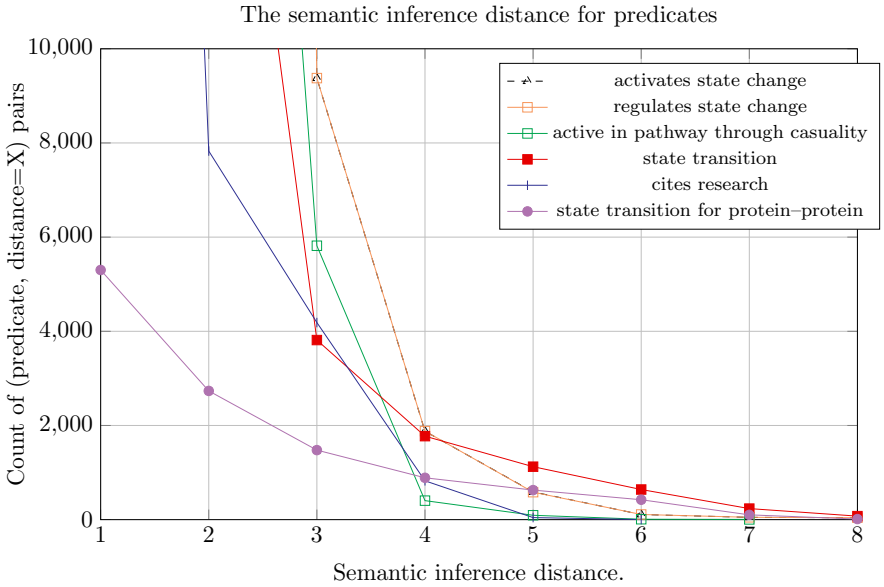

Fig. 1: Semantic inference and knowledge discovery. The above figure count each inferred relationship for a subset of the predicates in our database In the figure predicates at a *semantic inference distance* ≥ 2 capture inferences not known in public drug-discovery databases, hence there are 1000+ identified candidates for knowledge discovery. To increase accuracy of predictions the predicates, depicted as legends, are constructed from a unification of data from "UniProt" [6] and "BioPax" [16].

current issues in big-data mining. This paper ends with a brief summary of observations in section 6.

2 Related Work

A challenge in data-mining conserns the slow performance of software, as observed for [20] in Fig. 4. A possible explanation of the latter is an unawareness of high-performance software implementation strategies [21]. To exemplify, the major efforts in "systems biology is on developing fundamental computational and informatics tools" [22], an assertion motivated by how "a concerted effort to bring all the useful tools for pathway analysis in a common platform is still missing" [23]. When combining the observations of ([22,**?**]) and Fig. 4 we realize how poor-performing software represents a hurdle in knowledge discovery. To summarize, we observe that established approaches for data-mining suffers from:

1. Disparate data: insufficient data-coverage and prediction, *e.g.*, in [23,24,25,26,17];
2. Execution-time: high query response-time, *e.g.*, in [20,**?**];

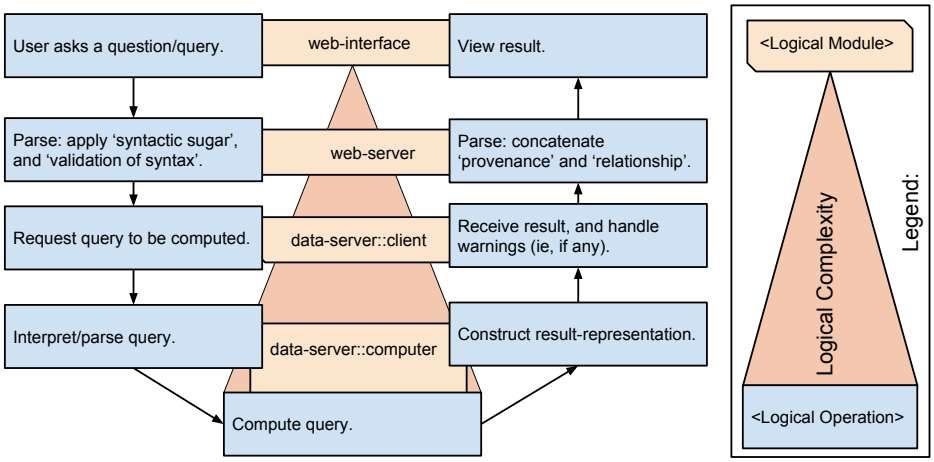

Fig. 2: Computational complexity versus user-interaction in knowledge searches. The above figure relates user-searches to the computational inference process. This reflects the flow of information, and complexity of software, for user-queries supported at `www.knittingTools.org`. While major research efforts are invested in improving performance of software categorized in the upper part of the figure, their return on investment is limited, *i.e.*, as captured from the figure. In contrast the majority of our efforts are invested on improving the software modules depicted in the bottommost part in the above figure, as exemplified in Fig. 5.

3. User searches: user-interfaces which limits domain-experts from accurate data-searches and result-interpretation, *e.g.*, in [18,28,29,30].

In below we briefly examine the above issues, focusing on issues concerned with *disparate data* and *execution-time*.

2.1 Execution-Time: Tools for similarity, feature-selection and clustering

There are more than 10^6 research-works concerned with data-mining[1]. An example concerns the k-means cluster-algorithm, where new permutations are published every year, *e.g.*, with respect to [31]. The work of [32] observes how existing software for data-analysis under-utilizes computer-hardware. A popular software-tool for cluster-analysis is the "cluster C" software [20]. From Fig. 4 we observe how the approach manages to outperform the software of [20] by a factor of 100,000x+. While [33] provide a GPU-optimized implementation of "DB-SCAN" [34], the GPU implementations limited support for user-defined parameters result in inaccurate cluster-predictions for numerous data-sets [35], *e.g.*,

[1] Observation from searching on "Google Scholar" for terms such as *PCA*, *k-means*, *Sum of Squared Error (SSE)*, *spearman*, *Euclid*, *correlation*, *similarity*, etc.

with respect to issues in *missing data* and similarity-metrics. In order to evaluate accuracy of cluster-algorithms, application of feature-selection, and many-dimensional hypothesis-testing, metrics for cluster-consistency are used [36]. Examples of cluster-consistency metrics are "Silhouette" [37], "Sum of Squared Error (SSE)" [38], "Rand's Index" [39] and "Rands Adjusted Index" [40], etc. Therefore, accurate software for data mining need to be optimized both with respect to number and execution-time of integrated metrics.

2.2 Disparate data and Execution-Time: Engines for data-access

A major challenge in big-data analytics concerns the slow performance of database-engines [41,23,22,42]. To exemplify, the authors of [23] asserts that there is no sound computational framework for database-management. The work of [43] observes how "big data analytics requires technologies to efficiently process large quantities of data" [43]. To address the performance lag in database-engines current approaches seeks to pre-compute queries [17,44,45], reduce RDF-dataset through slicing [46], etc. However, a prevailing issue concerns the high time-cost of queries: the search-engine of [47] use more than 13 minutes to evaluate a simple query. What may be argued is that the choice of accurate data-engines may address the performance issue. There is a large number of different data-engines for high-performance querying of semantic data [48,49,50,**?**,**?**]. One of these is the "Sesame" data-engine [53], a data-engine which is unable to provide real-time query-answer-time to simple queries [54]. Our earlier work [55] identifies how the established B-tree ([56,41]) data-structure results in a 10,000x+ performance-delay when compared to accessing data stored in an *in-memory 2d-sparse data-structure*, as discussed in [55]. In our [55] we demonstrates how a *2d sparse data-structure* may be used as an alternative to established data-engines, a work which observe how application of a *2d sparse data-structure* outperforms MySQL by 10,000,000x+ for important bio-medical queries.

3 Method: A holistic method for knowledge discovery

In the integration of real-time user access to 30 million bio-medical relationships we have faced the challenges described in research, as exemplified in section 2. From the works of others we realize that it is not feasible to follow the established strategies. To exemplify, major efforts by [17,**?**,19] are placed on translating data-formats into RDF. However, their approaches have not resulted in knowledge discovery. In the unification of data-resources we have addressed issues in assimilation of the graph-structured "BioPax" [16] formats and evidence-annotations in "OBO" [60], *i.e.*, where latter by definition is not supported by the "SPAR-QL" query-language. An example of erroneous name-mappings is seen for an entity with name "HDR" asserted by [61,**?**] to be an exact synonym of the "gata3" gene. In contrast the established view is that "HDR" describes a mechanism in cells [63]. The latter example is one of many fallacies observed in

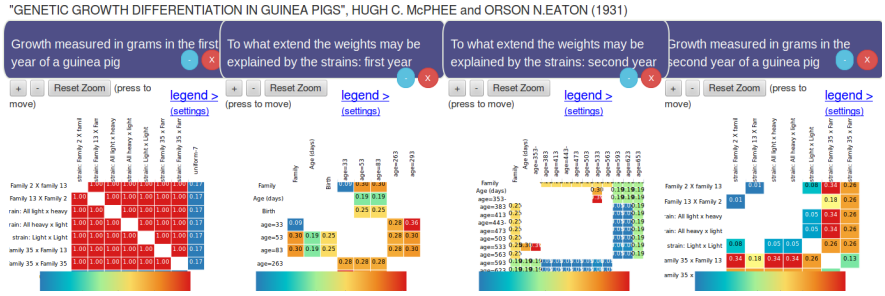

Fig. 3: Our support for knowledge-inferences in filtered data. The above figure exemplify how our approach enable knowledge-discovery, as described in [57]. Each of the sub-figures represent distinct data-sets capturing different hypothesis in [1]. The differences and similarities between the sub-figures provide clues of how guinea pigs develop. Importantly, the above separation between entities reflect the findings in [1], hence our interactive data-mining approach provide support for accurate and fast data-filtering of user-defined data-sets.

multi-origin databases, hence integration of data need to take care when using assertions from unreliable sources.

A different aspect concerns the execution-time of user-queries, exemplified in Fig. 2. To address the high time-cost of translating external data-bases into RDF, and searching RDF data-stores, we have designed a data-engine which accepts semantic relationships. When measuring the response-time of queries we observe how our new data-engine address issues in execution-time, as described in our [55]. The/Our semantic data-engine address issues such as:

1. Disparate data: integration of evidence annotations, hence less relationships to investigate during evidence-centered user-queries;
2. Execution-time: memory-cache aware data-searches, effective use of SSE [64], memory-tiling [65], etc;
3. User searches: pre-computation of statistics, and ranks, for database-vertices enable accurate suggestion when users type name of entities, as exemplified in Fig. 6.

The above described strategy exemplify approaches which reduce search-time without introducing erroneous heuristics, *e.g.*, in contrast to [18]. Fig. 5 presents a summary of the approaches undertaken to optimize the performance aspects of bio-medical knowledge discovery, hence a holistic method for data-mining. To exemplify, we from Fig. 5 observe how the holistic approach address issues in *disparate data* through a combination of manually curated rules (to address quality issues in data-resources, *e.g.*, the "HDR" use-case), application of clustering to unify entities both with respect to their database-resource (*e.g.*, "uniprot"), etc.

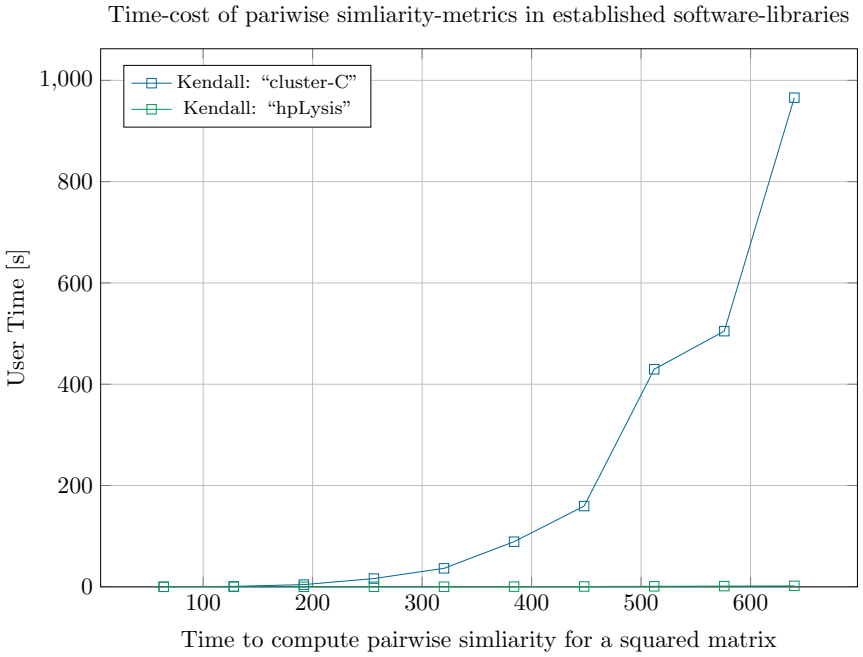

Fig. 4: Time-difference of our hpLysis software versus established approaches. The above figure capture the performance-difference of different strategies to compute the pairwise similarity metric of"Kendall's Tau" [58,59]. While the bottommost legend represents the time-cost of our hpLysis software, the topmost legend capture the time-cost of the popular "Cluster C" library [20]. The figure demonstrates how the hpLysis approach out-performs [20] by a factor of 100,000x+.

4 Result

The holistic approach for drug-discovery, introduced in this paper, is constructed to relate domain-experts to accurate interactions minted from big data. From below sub-sections we assert that the approach manages to correctly address the issues described in section 2.

4.1 Disparate data: Data-access in the bio-medical domain

There are more than 220 different knowledge sources in the bio-medical domain [66]. Correctness and usability are characteristics which describe the most popular tools for knowledge integration. An example tool is *cPath*, written by [66], which is built around a MySQL database [67]. From the performance measurements of data-structures in sub-section 4.4 we observe how semantic searches through MySQL results in a 10,000,000x+ performance-delay. The best tools provide access to data which has been manually curated by field experts, such

as the Reactome tool [26] or the BioGrid tool [25]. The back-bone of the tools is often the Gene Ontology (GO) [68], which is used to define the lexicographic order of the genes and proteins. Translating compartmentalized knowledge into an ontology for reasoning, such as the RDF format, is seen in [24,45]. The problem with both approaches is the performance and quality issues in knowledge discovery.

4.2 Execution-Time: Relationship between implementation, execution-time, and their influence

In data-mining the execution-time of software may render high-quality analytical approaches useless, *e.g.*, as inferred for large data-sets in Fig. 4. The application of established implementation-strategies results in under-performing code due to the challenges of compilers to identify strategies for performance tuning (as it otherwise would not have been a time-difference between different software implementations). In below we list a subset of observations from the holistic optimization of approaches for data-mining:

1. Search-time: the test-cases listed in sub-section 4.4 relate the time-cost of semantic searches in our novel database-engine to the established use of B-trees ([56,41]), observing a time-difference of 10,000x+;
2. Data-mining: Fig. 4 compare the time-cost of strategies for computing "Kendall's Tau". When our approach is compared to the popular "cluster-C" software [20], a time-difference of 100,000x+ is observed;
3. Software complexity: Fig. 5 examplify how accurate and fast user-searches involve steps in data-curation, analysis of semantic similarity, and construction of user-interfaces.

The above observations indicates that the application of low-level optimization strategies is a central part in efforts for mining of big-data, reflecting observations in section 2, hence the importance of an holistic optimization strategy.

4.3 User searches: application centered perspective

A sound interaction between software-tools and human domain-experts is seen as an essential part of knowledge discovery. In below we exemplify a subset of the strategies we have applied in the holistic approach (Fig. 5):

1. Semantic user-interactions: Fig. 6 exemplify how users are provided with support for both semantic queries (topmost sub-figures), interactive exploration (sub-figures in the middle) and signature queries (bottom-right sub-figure);
2. Testing hypothesis: in Fig. 3 we observe how a combination of the "MINE" metric [69] and our web-based framework for data-mining facilitate knowledge discovery on filtered query-subsets.

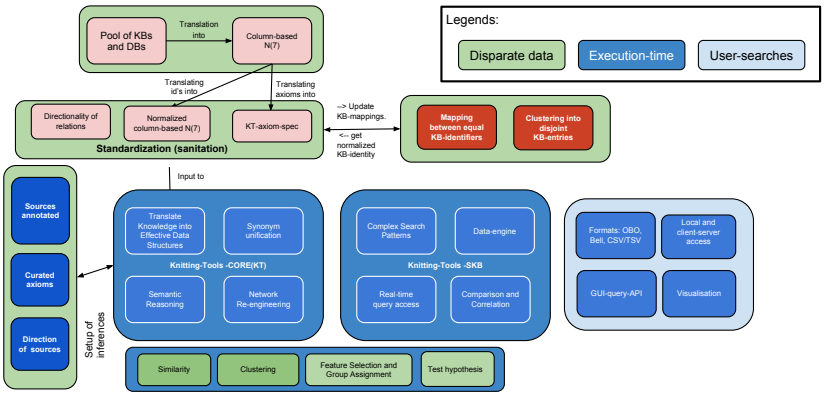

Fig. 5: An holistic approach for data-mining. The above figure depicts a collection of labelled boxes, such as *Pools of KBs and DBs* and *Visualization*. The legend-text (top-right) describe the classification of the different background rectangles, as discussed in section 3. An example of a *classification* concerns the process of handling *disparate data*, where tasks for data-parsing, entity optimization, and format-unification, are combined into an automated approach. The size of the background-boxes reflect their computational complexity, as discussed in Fig. 2. The uniqueness of our approach concerns how we relate the existing strategies into one unified model, thereby avoiding overheads associated with generalised approaches (such as RDF centered integration-strategies). To exemplify, when approaches such as [19] apply *Standardization* they use standardized rules for all of the integrated data-resources, hence entities from different data-resources are syntactically correct while semantically inaccurate.

4.4 Reproducibility: interfaces to validate and elaborate our approaches for data-mining

The results, summarized in this paper, may be re-produced through application of our software, as listed in below:

1. Semantics and data: `http://www.knittingTools.org/`;
2. MINE data-mining: `http://www.knittingTools.org/gui_lib_mine.cgi`;
3. MINE high-performance software: `https://bitbucket.org/oekseth/mine-data-analysis/downloads/`;
4. Software for data-analysis: `https://bitbucket.org/oekseth/hplysis-cluster-analysis-software`.

5 Use-cases: how the holistic approach improves drug discovery

The holistic approach represented in this paper manages to address issues in big-data analysis. The term *big data* depends on the complexity of data, algorithms, and use-cases to be evaluated, *e.g.*, where [70] asserts that a study of 857 proteins implies a *large-scale analysis*. This section therefore seeks to address strategies for:

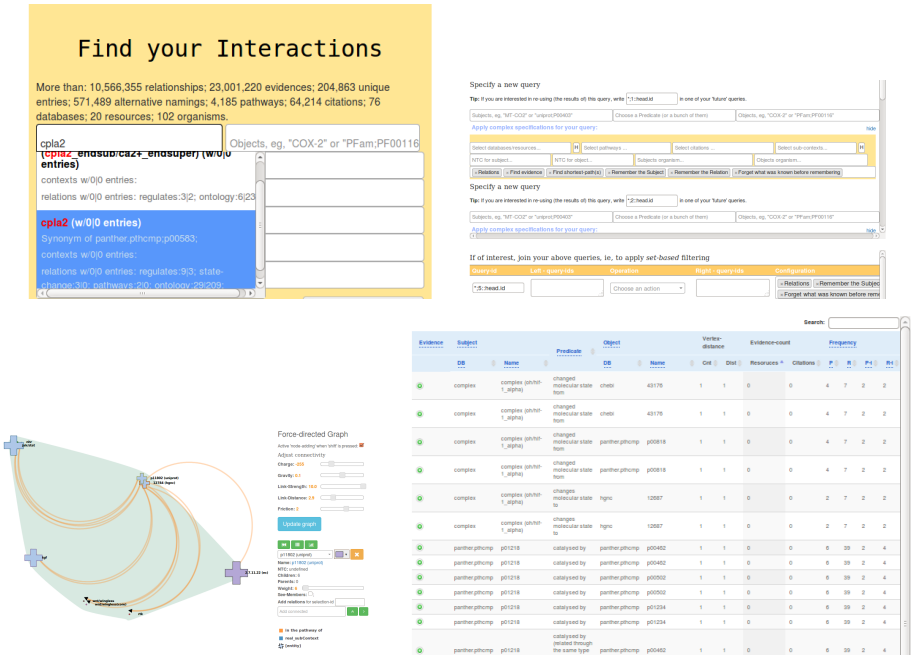

Fig. 6: A user-interface for semantic queries. The above sub-figures depict the user-interface for semantic searches, a user-interface designed for domain-experts without interest in programming. The two uppermost sub-figures depicts the query-form for submitting conditional questions. The drop-down menu observed in the top-left figure exemplifies the support for auto-complete. In the middle-right sub-figure the query-results is seen. To handle the frequent issue of 1000+ identified relationships, the table include filter-options. When a user filters a subset of the query-result, it is visualized in the middle-left sub-figure.

1. Disparate data: algorithms and data which may be accurately queried;
2. Execution-time: why the enabled performance-increase is important in drug discovery;
3. User searches: how domain-experts may identify accurate prediction from complex data.

5.1 Disparate data and user-searches

The "Knitting-Tools" web-server includes a number of pre-computed use-cases (`http://knittingtools.org/examples.cgi`). The results have been manually investigated and verified. Accuracy of predictions depends on data (Fig. 1) and accessibility (Fig. 6). Below use-cases provides a brief introduction to how complex searches may be applied on big-data.

Use-case(1): *What is known for "notch2"?* (`http://knittingtools.org/query.cgi?queryID=get_allRelations_forA_vertex`). The use-case illustrates

an exploratory search to identify all relationships, synonyms and provenance (*e.g.*, the set of databases) describing a vertex of interest, *e.g.*, the "notch2". Amplifies the use of *basic search functionality* to fetch relationships in KnittingTools, both for visual evaluation (Fig. 6), and as a prior data-gathering step before application of software for pattern identification (sub-section 5.2 and sub-section 5.3).

Use-case(2): *What is known for "cdk4", and how was this known?* (`http://knittingtools.org/query.cgi?queryID=get_allRelations_forA_vertex_evidence_and_synonyms`). Extends *use-case(1)* with logic to fetch the synonymous vertices (for each vertex in the set of identified relations) and the provenance for each relationships. Provides insight into why the identified relationships were predicted, *i.e.*, their provenance.

question(3): *Identify the regulations associated to the important event of apoptosis (i.e., 'controlled cell death').* (`http://knittingtools.org/query.cgi?queryID=intro_basic_bio_2`). The query identifies relationships associated to *pathways* and *regulations* for chemical entities, proteins, genes, and pathways. In the result provenance is associated to each relationship, a provenance which becomes visible when *clicking* the green-plus button in the result-table (Fig. 6).

5.2 Pattern identification and usability

The MINE software combine an highly accurate algorithm for pattern identification [69] with a web-interface for interactive data-exploration (Fig. 3). To evaluate the applicability of the MINE web-based software (`http://www.knittingTools.org/gui_lib_mine.cgi`) the data-sets of [71], [72], and [1] are evaluated. While the data-set by [1,71] provide explanation factors for growth of guinea pigs, [72] analysis the variation in the guinea pigs goat-spots. The conclusions presented by the authors are supported through application of the MINE web-interface.

5.3 Execution-Time and knowledge discovery

For large data-sets the prediction accuracy relates directly to the execution-time of data-mining software, *i.e.*, a users are otherwise needed to explore smaller data-samples and test fewer number of hypothesis. The below paragraphs exemplifies how our proposed approach increase accuracy of large-scale data-analysis.

Application(1). *Large-scale ontology engineering* (`https://bitbucket.org/oekseth/hplysis-ontology/`). We have developed a new *hpLysis-onto* software for high-performance engineering of bio-medical ontology. Ontologies are used in a large number of application, *e.g.*, to identify similarities of gene products from experimental outcomes [73] The *hpLysis-onto* software address performance issues in computation of transitive closures and transitive reductions, an issue hampering analysis of large and complex data-sets. For the task to compute transitive closures for all vertices, the software of [74] consumes more than one day on the 24 MB "Gene Ontology" [68]. In contrast, the *hpLysis-onto* software

manages to answer the latter query in less than one second, hence a significant improvement in performance.

Application(2). *Large scale Semantic similarity* (`https://bitbucket.org/oekseth/hplysis-cluster-analysis-software`). The *hpLysis* software is updated with a new high-performance library for computation of 20+ semantic similarity metrics. "Generally speaking, semantic similarity measures involve the GO tree topology, information content of GO terms, or a combination of both" [75]. The software proposed by [75] takes several hours to complete. The *hpLysis-semantic* software improves the performance of established software approaches by 1000x+, *i.e.*, without reducing the prediction accuracy. The latter is enabled through increased utilization of computer memory hardware. Semantic similarity-metrics are used to identify important traits in data-sets [76], *e.g.*, to (1) relate hypothetical assumptions to gene-expression-levels [3] and (2) with respect to "Word Sense Disambiguation" (WSD) for automated analysis of text-corpuses [77].

Application(3). *Many-dimensional data-analysis* ([36]). The *hpLysis* software provides an API for high-performance computation of 20+ cluster-algorithms, 320+ pairwise similarity-metrics, 10+ metrics for string-similarity, and 20+ metrics for cluster-validity. The work enables a performance-improvement of 600x+ for pairwise similarity-metrics such as "Canberra" and "Cosine", while 100,000x+ performance-improvement when compared to "Kendall's Tau" (Fig. 4). In large-scale data-analysis the execution-time severely hamper the types of relationship which may be explored, *e.g.*, when analysing gene-expressions data-sets for possible interactions, when using pathway-relationships (sub-section 5.1), application of ontology annotations for similarity-assessment, mining of bibliometric data-bases (*e.g.*, in [78]), etc. Therefore, *hpLysis* improves both accuracy of data-generation and analysis of user-defined data (Fig. 6).

6 Conlusion and Future Work

We have presented both a method, a database, and 10+ software, for data-mining. This paper argue that the holistic approach, which captures an ensemble of approaches and high performance software, manages to overcome the current hurdles in big-data drug-discovery. Fig. 6 exemplify how domain-experts may interact with our real-time support for querying 30+ million bio-medical relationships. In order to evaluate the quality of the approach we investigate the number of accurate and unique relationships identified in our approach, exemplified in Fig. 1. The 1000+ novel candidate interactions which are identified highlight the ability of our approach to automatically identify relationship which are not known in literature. Through an optimized data-engine the relationships are accessible for users in real-time, exemplified in Fig. 3.

In this paper we have described an approach to unify our semantic interface (`www.knittingTools.org`) with our high-performance software application (*e.g.*, `https://bitbucket.org/oekseth/hplysis-cluster-analysis-software`). Through concrete use-cases we have exemplified how the approach address is-

sues in *disparate data*, *execution-time*, and *user searches*, ie, parameters which are critical in discovery of knowledge. From the examples we observe how our method and 10+ novel software approaches address issues in big-data drug-discovery. Therefore, we assert that our novel holistic approach may influence strategies for mining of big-data.

6.1 Future Work

We plan to address the weakness of the user-interfaces and the unknown quality of our knowledge inferences. In order to improve our user-interfaces we are now initiating efforts in usability testing for different target groups. Similarily, we have initiated efforts to evaluate the drug-impact of our putative knowledge discoveries. Both of the issues require year-long lab-experiments, hence the importance of quality and performance enabled through our novel method and software.

Acknowledgements

The authors would like to thank MD K.I. Ekseth at UIO, Dr. O.V. Solberg at SINTEF, Dr. S.A. Aase at GE Healthcare, MD B.H. Helleberg at NTNU–medical, Dr. Y. Dahl, Dr. T. Aalberg, Dr. J.C. Meyer, and K.T. Dragland at NTNU, and the High Performance Computing Group at NTNU for their support.

References

1. McPhee, H.C., et al.: Genetic growth differentiation in guinea pigs. Technical report, United States Department of Agriculture, Economic Research Service (1931)
2. Stephenson, D.T., Lemere, C.A., Selkoe, D.J., Clemens, J.A.: Cytosolic phospholipase a 2 (cpla 2) immunoreactivity is elevated in alzheimer's disease brain. Neurobiology of disease **3**(1), 51–63 (1996)
3. Feuerherm, A.J., Johansen, B.: Rheumatoid arthritis treatment. US Patent App. 13/783,088 (2013)
4. Zhao, M., Yang, C.C.: Mining online heterogeneous healthcare networks for drug repositioning. In: Healthcare Informatics (ICHI), 2016 IEEE International Conference On, pp. 106–112 (2016). IEEE
5. for Biotechnology Information, N.C.: PubMed data-base for biomedical litterature. `https://www.ncbi.nlm.nih.gov/pubmed/` (2017)
6. Consortium, U., *et al.*: Uniprot: the universal protein knowledgebase. Nucleic acids research **45**(D1), 158–169 (2017)
7. Economist, T.: How science goes wrong: Trouble at the lab. The Economist **409**(8858), 21–24 (2013)
8. Cuatrecasas, P.: Drug discovery in jeopardy. Journal of Clinical Investigation **116**(11), 2837 (2006)
9. DiMasi, J.A., Grabowski, H.G., Hansen, R.W.: Innovation in the pharmaceutical industry: new estimates of r&d costs. Journal of health economics **47**, 20–33 (2016)
10. Al-Huniti, N.: Quantitative Decision-Makingin Drug Development. `http://www.phuse.eu/download.aspx?type=cms&docID=5334` (2013)

11. Herper, M.: The Truly Staggering Cost Of Inventing New Drugs. https://www.forbes.com/sites/matthewherper/2012/02/10/the-truly-staggering-cost-of-inventing-new-drugs/#1a0129244a94 (2012)
12. DiMasi, J.A., Grabowski, H.G., Hansen, R.W.: The cost of drug development. New England Journal of Medicine **372**(20), 1972–1972 (2015)
13. Association, C.B.R., et al.: Fact sheet: New drug development process. FDA Special Consumer Report
14. Thomas, K.: The price of health: the cost of developing new medicines. https://www.theguardian.com/healthcare-network/2016/mar/30/new-drugs-development-costs-pharma (2016)
15. Paul, S.M., Mytelka, D.S., Dunwiddie, C.T., Persinger, C.C., Munos, B.H., Lindborg, S.R., Schacht, A.L.: How to improve r&d productivity: the pharmaceutical industry's grand challenge. Nature reviews. Drug discovery **9**(3), 203 (2010)
16. Demir, E., Cary, M.P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'Eustachio, P., Schaefer, C., Luciano, J., *et al.*: The biopax community standard for pathway data sharing. Nature biotechnology **28**(9), 935–942 (2010)
17. Blonde, W.: Metarel, an ontology facilitating advanced querying of biomedical knowledge. PhD thesis, Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Ghent, Belgium (2012)
18. Antezana, E., Blond, W., Egaña, M., Rutherford, A., Stevens, R., De Baets, B., Mironov, V., Kuiper, M.: Biogateway: a semantic systems biology tool for the life sciences. BMC bioinformatics **10**(10), 11 (2009)
19. Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., Morissette, J.: Bio2rdf: towards a mashup to build bioinformatics knowledge systems. Journal of biomedical informatics **41**(5), 706–716 (2008)
20. de Hoon, M.J., Imoto, S., Nolan, J., Miyano, S.: Open source clustering software. Bioinformatics **20**(9), 1453–1454 (2004)
21. Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H.: The erato systems biology workbench: enabling interaction and exchange between software tools for computational biology (2002)
22. Butcher, E.C., Berg, E.L., Kunkel, E.J.: Systems biology in drug discovery. Nature biotechnology **22**(10), 1253 (2004)
23. Chowdhury, S., Sarkar, R.R.: Comparison of human cell signaling pathway databasesevolution, drawbacks and challenges. Database **2015** (2015)
24. Beisswanger, E., Lee, V., Kim, J.-J., Rebholz-Schuhmann, D., Splendiani, A., Dameron, O., Schulz, S., Hahn, U., *et al.*: Gene regulation ontology (gro): design principles and use cases. Studies in health technology and informatics **136**, 9 (2008)
25. Stark, C., Breitkreutz, B.-J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X., *et al.*: The biogrid interaction database: 2011 update. Nucleic acids research **39**(suppl 1), 698–704 (2011)
26. Croft, D., OKelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., *et al.*: Reactome: a database of reactions, pathways and biological processes. Nucleic acids research **39**(suppl 1), 691–697 (2011)
27. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.*: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**(Oct), 2825–2830 (2011)

28. Clustergrammer, J.: Clustergrammer heatmap visualization. http://amp.pharm.mssm.edu/clustergrammer/
29. Metsalu, T., Vilo, J.: Clustvis: a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. Nucleic Acids Research **43**(W1), 566–570 (2015). doi:10.1093/nar/gkv468
30. Tan, C.M., Chen, E.Y., Dannenfelser, R., Clark, N.R., Maayan, A.: Network2canvas: network visualization on a canvas with enrichment analysis. Bioinformatics **29**(15), 1872–1878 (2013). doi:10.1093/bioinformatics/btt319
31. Chen, Y., Zeng, Y., Luo, F., Yuan, Z.: A new algorithm to optimize maximal information coefficient. PloS one **11**(6), 0157567 (2016)
32. Mekkat, V., Natarajan, R., Hsu, W.-C., Zhai, A.: Performance characterization of data mining benchmarks. In: Proceedings of the 2010 Workshop on Interaction Between Compilers and Computer Architecture, p. 11 (2010). ACM
33. Andrade, G., Ramos, G., Madeira, D., Sachetto, R., Ferreira, R., Rocha, L.: G-dbscan: A gpu accelerated algorithm for density-based clustering. Procedia Computer Science **18**, 369–378 (2013)
34. Ester, M., Kriegel, H.-P., Sander, J., Xu, X., *et al.*: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd, vol. 96, pp. 226–231 (1996)
35. Ekseth, O.K., Hvasshovd, S.-O.: How an optimized DB-SCAN implementation reduce execution-time and memory-requirements for large data-sets. Accepted for publication (2017)
36. Ole Kristian Ekseth: hpLysis: a high-performance software-library for big-data machine-learning. https://bitbucket.org/oekseth/hplysis-cluster-analysis-software/. Online; accessed 06. June 2017
37. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
38. Lloyd, S.: Least squares quantization in pcm. IEEE transactions on information theory **28**(2), 129–137 (1982)
39. Rand, W.M.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association **66**(336), 846–850 (1971)
40. Yeung, K.Y., Ruzzo, W.L.: Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data. Bioinformatics **17**(9), 763–774 (2001)
41. Jagadish, H., Olken, F.: Database management for life sciences research. ACM SIGMOD Record **33**(2), 15–20 (2004)
42. Eltabakh, M.Y., Ouzzani, M., Aref, W.G., Elmagarmid, A.K., Laura-Silva, Y., Arshad, M.U., Salt, D., Baxter, I.: Managing biological data using bdbms. In: Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference On, pp. 1600–1603 (2008). IEEE
43. Lau, L., Yang-Turner, F., Karacapilidis, N.: Requirements for big data analytics supporting decision making: A sensemaking perspective. In: Mastering Data-Intensive Collaboration and Decision Making, pp. 49–70. Springer, ??? (2014)
44. Blond, W., Antezana, E., Mironov, V., Schulz, S., Kuiper, M., Baets, B.D.: Using the relation ontology Metarel for modelling Linked Data as multi-digraphs (2012)
45. Blond, W., Mironov, V., Antezana, E., Venkatesan, A., Baets, B.D., Kuiper, M.: Reasoning with bio-ontologies: using relational closure rules to enable practical querying. Oxford Bioinformatics **27**, 1562–1568 (2011). doi:10.1093/bioinformatics/btr164

46. Marx, E., Shekarpour, S., Soru, T., Braşoveanu, A.M., Saleem, M., Baron, C., Weichselbraun, A., Lehmann, J., Ngomo, A.-C.N., Auer, S.: Torpedo: Improving the state-of-the-art rdf dataset slicing. In: Semantic Computing (ICSC), 2017 IEEE 11th International Conference On, pp. 149–156 (2017). IEEE
47. Papanikolaou, N., Pavlopoulos, G.A., Pafilis, E., Theodosiou, T., Schneider, R., Satagopam, V.P., Ouzounis, C.A., Eliopoulos, A.G., Promponas, V.J., Iliopoulos, I.: Biotextquest+: a knowledge integration platform for literature mining and concept discovery. Bioinformatics **30**(22), 3249–3256 (2014)
48. Kolpakov, F., Poroikov, V., Sharipov, R., Kondrakhin, Y., Zakharov, A., Lagunin, A., Milanesi, L., Kel, A.: Cyclonetan integrated database on cell cycle regulation and carcinogenesis. Nucleic acids research **35**(suppl_1), 550–556 (2007)
49. Demir, E., Babur, Ö., Rodchenkov, I., Aksoy, B.A., Fukuda, K.I., Gross, B., Sümer, O.S., Bader, G.D., Sander, C.: Using biological pathway data with paxtools. PLoS computational biology **9**(9), 1003194 (2013)
50. Masseroli, M., Pinoli, P., Venco, F., Kaitoua, A., Jalili, V., Palluzzi, F., Muller, H., Ceri, S.: Genometric query language: a novel approach to large-scale genomic data management. Bioinformatics **31**(12), 1881–1888 (2015)
51. Mironov, V., Seethappan, N., Blond, W., Antezana, E., Splendiani, A., Kuiper, M.: Gauging triple stores with actual biological data. BMC bioinformatics **13**(1), 3 (2012)
52. Wylot, M., Cudré-Mauroux, P.: Diplocloud: Efficient and scalable management of rdf data in the cloud. IEEE Transactions on Knowledge and Data Engineering **28**(3), 659–674 (2016)
53. Huysmans, M., Richelle, J., Wodak, S.J.: Sesam: a relational database for structure and sequence of macromolecules. Proteins: Structure, Function, and Bioinformatics **11**(1), 59–76 (1991)
54. Guo, Y., Pan, Z., Heflin, J.: Lubm: A benchmark for owl knowledge base systems. Web Semantics: Science, Services and Agents on the World Wide Web **3**(2), 158–182 (2005)
55. Ekseth, O.K., Hvasshovd, S.-O.: hpLysis database-engine: A new data-scheme for fast semantic queries in biomedical databases. Under review: Provides details of the in-memory data-engine: contact oekseth@gmail.com for the paper. (2017)
56. Bayer, R.: Symmetric binary b-trees: Data structure and maintenance algorithms. Acta Informatica **1**, 290–306 (1972). 10.1007/BF00289509
57. Ekseth, K., Hvasshovd, S.: hpLysis MINE: A high-performance approach for computation of the accurate MINE simliarty-metric. `http://www.knittingtools.org/gui_lib_mine.cgi`. Online; accessed 06. June 2017
58. Knight, W.R.: A computer method for calculating kendall's tau with ungrouped data. Journal of the American Statistical Association **61**(314), 436–439 (1966)
59. Kendall, M.G.: Rank correlation methods. (1948)
60. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., *et al.*: The obo foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology **25**(11), 1251–1255 (2007)
61. Hoffmann, R.: A wiki for the life sciences where authorship matters. Nature genetics **40**(9), 1047–1051 (2008)
62. Chawla, K., Tripathi, S., Thommesen, L., Lægreid, A., Kuiper, M.: Tfcheckpoint: a curated compendium of specific dna-binding rna polymerase ii transcription factors. Bioinformatics **29**(19), 2519–2520 (2013)

63. Davis, L., Maizels, N.: Homology-directed repair of dna nicks via pathways distinct from canonical double-strand break repair. Proceedings of the National Academy of Sciences **111**(10), 924–932 (2014)
64. Intel: SSE computer-hardware-low-level parallelism. `https://software.intel.com/sites/landingpage/IntrinsicsGuide/`. Online; accessed 06. June 2017
65. Drepper, U.: What every programmer should know about memory. Red Hat, Inc **11**, 2007 (2007)
66. Cerami, E., Bader, G., Gross, B., Sander, C.: cpath: open source software for collecting, storing, and querying biological pathways. BMC bioinformatics **7**(1), 497 (2006)
67. MySQL: MySQL database engine. `https://www.mysql.com/` (2017)
68. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.*: Gene ontology: tool for the unification of biology. Nature genetics **25**(1), 25–29 (2000)
69. Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., Sabeti, P.C.: Detecting novel associations in large data sets. science **334**(6062), 1518–1524 (2011)
70. Butland, G., Peregrín-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., *et al.*: Interaction network containing conserved and essential protein complexes in escherichia coli. Nature **433**(7025), 531–537 (2005)
71. Peaker, M., Taylor, E.: Sex ratio and litter size in the guinea-pig. Journal of reproduction and fertility **108**(1), 63–67 (1996)
72. Wright, S., Chase, H.B.: On the genetics of the spotted pattern of the guinea pig. Genetics **21**(6), 758 (1936)
73. Pesquita, C., Faria, D., Falcao, A.O., Lord, P., Couto, F.M.: Semantic similarity in biomedical ontologies. PLoS computational biology **5**(7), 1000443 (2009)
74. Antezana, E., Egana, M., Baets, B., Kuiper, M., Mironov, V.: Onto-perl: An api for supporting the development and analysis of bio-ontologies. Bioinformatics (2008)
75. Ehsani, R., Drabløs, F.: Topoicsim: a new semantic similarity measure based on gene ontology. BMC bioinformatics **17**(1), 296 (2016)
76. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE transactions on systems, man, and cybernetics **19**(1), 17–30 (1989)
77. McInnes, B.T., Pedersen, T.: Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. Journal of biomedical informatics **46**(6), 1116–1124 (2013)
78. Aalberg, T., Žumer, M.: The value of marc data, or, challenges of frbrisation. Journal of Documentation **69**(6), 851–872 (2013)