

Hazard Estimation and Method Comparison with OWL-Encoded Toxicity Decision Trees

Leonid L. Chepelev¹, Dana Klassen¹, and Michel Dumontier^{1,2,3},

¹ Department of Biology, ² Institute of Biochemistry, and ³ School of Computer Science, Carleton University, 1125 Colonel By Drive, K1S 5B6, Ottawa, Canada
{leonid.chepelev, dana.klassen, michel.dumontier}@gmail.com

Abstract. Industrial and regulatory evaluation of chemical toxicity is often done via statistical analysis of chemical features focusing on chemical structure and function. One popular method to characterize chemical toxicity involves the development of decision trees based on large sets of empirical toxicological data where chemicals are assigned toxicity or activity classes. In this paper, we describe the representation of decision trees as OWL ontologies that can be used to carry out initial evaluation of toxicity and activity of prospective chemical products. We further discuss how trees derived from different datasets can be semantically compared by examining the logical equivalence of the toxicity and bioactivity classes in different trees. Taken together, this initial work forms the basis for continued investigation into OWL-driven semantic framework for toxicity evaluation.

Keywords: Chemical Hazard Estimation, Computational Toxicology, Decision Trees

1 Introduction

Our industrialized society relies on millions of diverse chemical entities in applications as broad as energy production, combating disease, and manufacturing. As novel chemicals are developed and as industrial processes evolve, we become heavily exposed to an increasingly diverse pool of environmental pollutants and their poorly characterized by-products. The resource commitment necessary to fully characterize the toxicity of even a single chemical entity experimentally is very substantial. Since the pool of chemicals in need of routine toxicity screening by organizations such as environmental protection agencies and pharmaceutical companies is practically infinite and the resources for this task are often scarce, alternative means of toxicity screening are often applied to prioritize compound screening or alert chemical researchers to the potential adverse effects of their molecule of interest, especially in the early stages of compound development.

Such predictive *in silico* approaches may be broadly characterized into two major categories: data-driven systems and expert systems [1]. Data-driven systems involve the generation of mathematical models (regression, neural network, or any other method) to correlate computed or observed physicochemical molecular properties to their experimentally obtained functional characteristics, such as toxicity, binding

affinity for a given enzyme, or biological activity of a given type. The result of data-driven systems are quantitative structure-activity relationships (QSAR) that are specific to the class of compounds represented in the training set and are often difficult to logically interpret or integrate even for a human operator.

Expert-based systems, on the other hand, strive to capture the knowledge of human toxicology experts into machine-readable models with the aim of automating chemical classification and chemical information analysis. Expert-based systems can take a number of forms, among which rule-based and decision tree-based systems are quite prominent. Rule-based systems rely on the formulation of a number of independent rules that can be integrated to construct a logical conclusion about the toxicity or activity of a given compound. Decision tree-based systems involve the sequential execution of a series of logical tests, with each branch point of the tree containing a logical test, and each leading either to a final classification, or a deferral to further tests (Fig. 1).

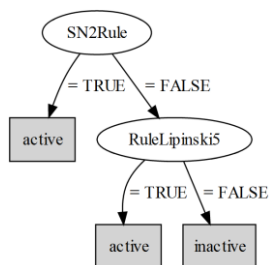


Fig. 1. A simple toxicity decision tree: at each branching point, a rule is evaluated, and based on the outcome of this rule, either a final activity decision is made, or judgment is deferred to another node.

Since their introduction four decades ago [2], decision tree-based toxicity and activity prediction systems have gained acceptance by academics and industrial researchers alike, finding applications in predicting molecular properties such as mutagenicity, toxicity, and skin sensitization among others [3]. Furthermore, automated objective methods have appeared to emulate the work of human experts by creating decision trees in which rules and tree structures are drawn based on the analysis of empirical toxicity data [4]. Aside from simplifying and automating classification efforts, and unlike data-driven toxicology prediction systems, decision tree-encoded expert knowledge is understandable to humans and machines alike.

Unfortunately, the potential of OWL ontologies to formally capture and enact such expert-based decisions in chemical toxicology and many other fields has not yet been fully realized. Consequently, the decision frameworks and the supporting databases for making such decisions are still largely fragmented along discipline, software, and institutional divides. Since biological and chemical information is increasingly standardized and integrated into the Semantic Web through initiatives like Bio2RDF [5], we find ourselves at the point where OWL-based formalization of expert rule bases and decision trees, combined with ready access to vast amounts of linked data can yield unprecedented, tangible benefits in integrated bioactivity and toxicity prediction and predictive method comparison and integration.

In this work, we demonstrate the automated generation of biologically relevant decision trees and their subsequent representation as OWL ontologies. We show how the OWL ontologies can be used for classification over RDF-based linked data and discuss the potential for the application of OWL-based decision trees on large RDF chemical knowledgebases. Finally, we demonstrate the automated logical comparison and integration of bioactivity/toxicity classes on the example of automatically derived decision trees for drug-likeness and toxicity prediction. We believe that this work is an important initial development in the formalization, standardization, and integration of computational toxicology resources and predictive classification methods.

2 Methods

In order to explore the practical utility of decision trees for predictive chemical toxicology, we first built decision trees using a popular toolkit with experimental and molecular features from a chemical carcinogenicity dataset. These trees were converted to OWL ontologies, which were used in classification of RDF-based data using automated reasoning. Finally, we demonstrate the possibility of inferring toxicity/bioactivity class logical equivalence for different OWL-based decision trees.

2.1 Data Sources and Data Preparation

Our analysis made use of empirically and theoretically derived datasets. A carcinogenic toxicity dataset, from which 1400 chemical entities were selected, was obtained from the ToxCast database [6]. These compounds were either active or inactive with respect to single cell mutagenicity. Then, 318 non-redundant features for each molecule were computed using the ToxTree API [7] to determine a Boolean value for each feature: true for feature presence and false for absence. These features corresponded to rules at decision tree branch points: true if satisfied, and false if not.

Features for the Rule of Five training set, consisting of 7000 compounds selected from HMDB [10], were computed using the Chemistry Development Kit [8], and the drug-likeness attribute was derived using the logical tests outlined by Lipinski [9]. Software and data are available upon request.

2.2 Decision Tree Construction and Validation

Weka [11] was used to construct and validate binary decision trees using the experimental and computed feature information. Decision trees were constructed using the J48 algorithm [4]. We applied ten-fold cross-validation to derive a set of statistical measures of tree predictive ability. Though these statistical measures are not directly relevant for this work, they have been included as annotations on resultant OWL-encoded decision trees for completeness. For the purposes of discussion in this work, we generated five decision trees: Lipinski Rule of Five, modified Lipinski Rule of Five, as well as trees resulting from different partitions of the ToxCast datasets.

2.3 Representation of Decision Trees as OWL Ontologies

OWL ontologies were constructed using the OWL API [12] from the decision tree graphs represented with the DOT graph description language. Each decision node is represented as being equivalent to a class expression involving the parent decision node intersected with a restriction on the attribute value (true/false) that the parent node represents (e.g. contains an alcohol moiety). For example, given three substances (A, B and C), where A is the parent substance and B and C are defined with respect to the exact value of the parent feature X, and given Substance classes, 'has attribute' object property, and 'has value' functional datatype property, the equivalent class expressions corresponding to *Substance B* and *Substance C* are:

Substance B EquivalentClass

Substance A and 'has attribute' some (Attribute X and 'has value' true)

Substance C EquivalentClass

Substance A and 'has attribute' some (Attribute X and 'has value' false)

EquivalentClass axioms were added to terminal nodes corresponding to the final classification, e.g. *toxic* or *non-toxic*. This enabled us to reflect both the structure of the decision tree and the formal axioms leading to the classification of a given chemical entity into a given biological functional class. We did not include covering axioms (e.g. A can have the disjoint subclasses B or C) because we would like to avoid inconsistencies in some manually created trees where multiple classification outcomes may be possible and the most hazardous classification outcome is selected.

2.4 Ontology Integration and Comparison

For direct comparison of simple ontologies to logically identify predicted toxicity and bioactivity class equivalence, we used the Pellet reasoner through the OWL API in Java. We fused ontologies through a direct import and carried out ontology classification using Pellet [13]. In cases where an equivalence or subclass relationship between the final bioactivity or toxicity classes was identified, we noted this relationship directly.

2.5 Chemical Classification

Molecular entities were instantiated using conventions set out by the Chemical Entity Semantic Specification (CHESS) [14] and the Chemical Information Ontology (CHEMINF) [15]. These entities annotated with chemical feature data were classified using Pellet through the OWL API into the predicted toxicity classes using our automatically generated OWL-based decision trees.

3 Results and Discussion

3.1 OWL-Based Decision Trees: Rule of Five

The first task that we addressed with our automated OWL ontology decision tree generator was the construction of simple ontologies where the classification rules involved the evaluation of numerical values associated with various molecular descriptors. This is a fairly common mode of preliminary screening of large compound datasets in initial stages of cheminformatics analysis. The decision tree generated by Weka using computed data reproduced the Rule of Five criteria (Fig. 2).

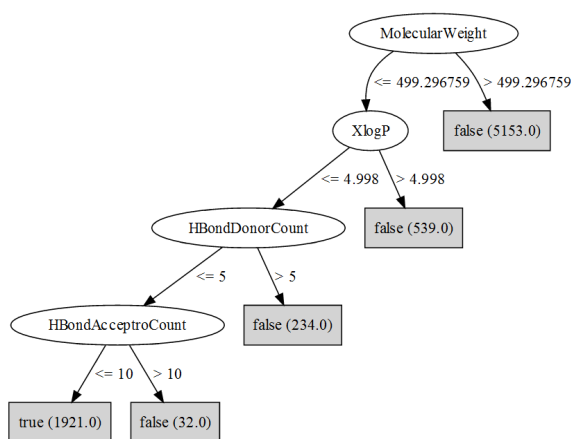


Fig. 2. A decision tree generated from a computationally derived dataset of drug-like compounds. Drug-like compound classification is indicated as *true*. Correctly classified molecule counts are given in brackets. No classification was incorrect.

There was little surprise that the Rule of Five criteria (used as an example, not a practical application) which we imposed in the computationally derived dataset were perfectly returned to us after data-based decision tree construction in Weka. However, this had demonstrated to us that, given a sufficient amount of data with low levels of noise, one could successfully derive meaningful and useful numerical cutoff-based decision trees which could subsequently be converted to predictive ontologies.

In order to carry out the conversion, we have followed the scheme indicated in Section 2.3 to obtain a set of substance classes that followed numerical cutoff rules, such as the following.

Substance_N1:

Substance_N0 and has_attribute some (MolecularWeight and has_value some double[<= "500"^^double])

As a result of applying our generator, we have obtained an ontology that perfectly captured the Rule of Five decision tree (Fig. 3).

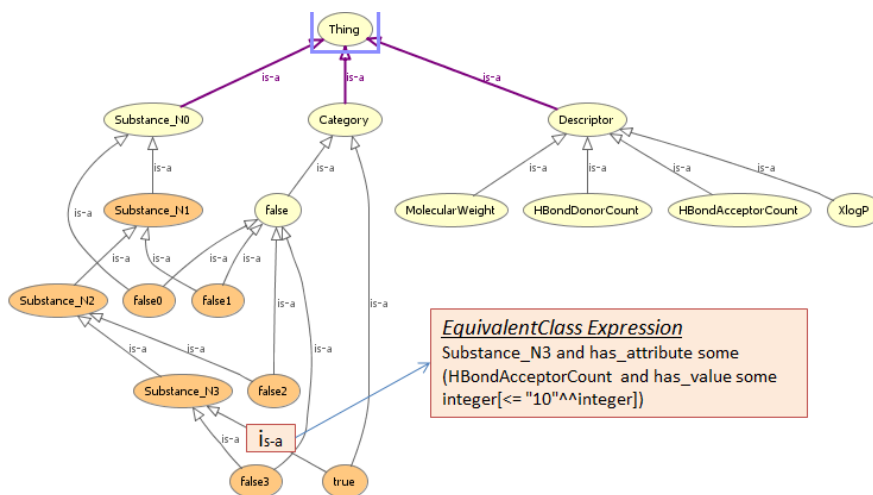


Fig. 3. The structure of an automatically generated OWL representation of a Rule of Five tree (Fig. 2).

3.2 OWL-Based Decision Trees: Large-Scale Boolean Feature-Based Trees

Unfortunately, biological information is often a subject to extensive variation, whether due to noise in experimental conditions or the abundance of the variable parameters that may differ even within a single laboratory and experiment. Compounding this is the limited experimental data availability to characterize most forms of biological activity, especially for experiments that are not high-throughput at inception. As a result, the real-world data is rarely as neatly classifiable as in the decision tree above. However, our primary concern in this work has been the proof of principle for the utility of OWL-based decision trees. To this end we have been able to generate a number of useable trees with the full 318-feature set (not shown due to complexity), as well as the more presentation-friendly limited feature sets (Fig. 4).

Upon closer examination of such increasingly complex decision trees, we have identified several unanticipated classification challenges. The greatest surprise has come from the identification of the logical equivalence of several branches within some of the generated trees. While that was considered completely plausible at the level of the individual nodes, the subsequent identification of the logical equivalence of the final toxicity and bioactivity classifications upon the application of reasoners to our generated ontologies has led to some concerns over the validity and applicability of our approach. Clearly, the equivalence of the class of toxic compounds to the non-toxic compounds is not an anticipated or desirable effect for an ontology used to replace the existing classification systems. Further, in order to make the decision tree more transparent, we needed a way to trace the logical path taken to activity classification leaves, while still preserving broad activity classification capacity.

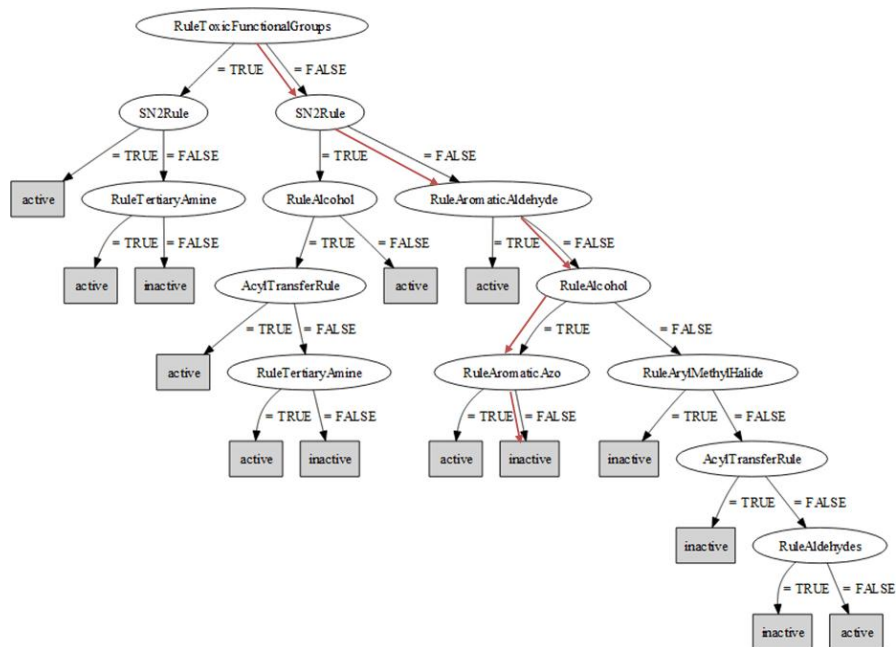


Fig. 4. A simplified carcinogenic toxicity decision tree generated from a ToxCast dataset, using a restricted set of chemical features for ease of presentation. Note the repetition of some rules at multiple decision tree nodes. The path taken to classify acetaminophen, as detected with the explanation functionality of Protégé, has been highlighted with red arrows.

After careful consideration of the logical explanation of the equivalence of these practically distinct classes, we identified the cause of the problem to lie in the repetition of rules within a single decision tree and the lack of the distinction between the nodes that executed rules in a particular order. As such, it was quite possible to arrive at a situation where, having ignored the context of the rest of the tree, the classifier technically correctly assigned class equivalence between the toxic and non-toxic compounds simply because parts of the paths taken to these classifications were similar, while the other parts were not mutually exclusive.

To rectify this problem, we have recognized that node-specific classification rule tracking had to be implemented. Thus, we amended our generator to include a local set of node-specific classification features within a given ontology. This translated into alterations to substance classifications, as follows.

Substance_N6:

Substance_N0 and has_attribute some (RuleToxicFunctionalGroups_N0 and has_value value false)

Note that what used to be the *RuleToxicFunctionalGroups* descriptor became the *RuleToxicFunctionalGroups_N0* descriptor. This amendment was effective in solving our misclassification problem. However, the introduction of ontology-specific descriptors would negate our ability to integrate and compare the different ontologies, as well as to draw on existing repositories of chemical entities annotated with the general standard descriptors and features. To rectify ontology comparison deficiency,

we have created versions of our decision tree ontologies where node-specific rules were explicitly defined as subclasses of their generic counterparts. Similarly, node-specific activity leaves were introduced to enable tracing classification paths. Thus, although we had to artificially distinguish activity categories and rules, we were still able to query for the compounds falling into the general activity classes, as well as to trace classification paths, important in e.g. automated toxicity tree comparison.

3.3 Chemical Entity Classification

While the above amendment permitted comparison between multiple ontologies and still avoided erroneous class equivalence conclusions, it did not address drawing on existing data repositories, as there is no direct inference that if a general rule bears a particular value, there exists an instance of its subclass that bears the same value. The first intuitive suggestion to clear this task is to modify our generator to also create ontologies where the general rules were specified as subclasses of the node-specific rules. This has allowed us to automatically make the necessary inference to import data from existing chemical knowledge repositories in RDF.

However, upon carrying out the classification within such ontologies, we have been unpleasantly surprised to find out that due to the introduced equivalences at the data level, some of our instances were capable of adopting both, active and inactive classifications. In order to rectify this problem, we defined node-specific final classifications (e.g. *active_N3*) which were declared to be subclasses of the general final classes (e.g. *active* and *inactive*) (Fig. 5).

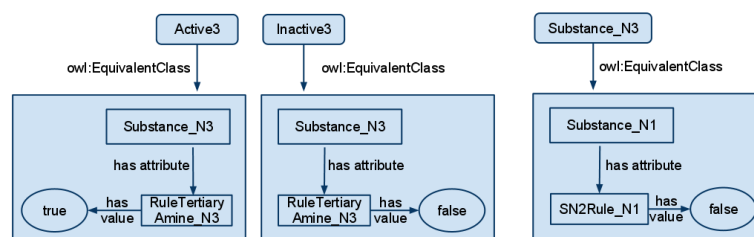


Fig. 5. A fragment of the final, classification-friendly decision tree.

Classification was successfully carried out by querying whether a given instance belonged to one of these general classes. Using thus constructed decision tree-based ontologies, we encountered no problems classifying numerous RDF-encoded molecules bearing the requisite information. A sample OWL model is available [18].

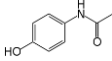
Structure	Acyl Transfer Rule	Rule Alcohol	Rule Aldehydes	Rule Aromatic Aldehyde	Rule Aromatic Azo	Rule Aryl Methyl Halide	Rule Tertiary Amine	Rule Toxic Functional Groups	SN2 Rule
	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

Fig. 6. Relevant features of acetaminophen used in classification.

As an example, consider the case of acetaminophen, a known non-carcinogen. Its attributes (Fig. 6) were imported from its CHESS [15] representation and correctly classified as *inactive* according to the decision tree presented earlier (Fig. 4). This classification was also reproduced using numerical trees (omitted for brevity).

Further, unlike the traditional classification systems, which are essentially black boxes, our approach has allowed us *to automatically trace the exact route taken to classifying acetaminophen as a non-carcinogen*, using the explanation feature of Protégé [16]. The fact that we created artificially distinct activity classes in our trees did not prevent us from querying for chemical activity in terms of general categories.

3.4 Ontology Integration and Concept Comparison

Thanks to the automatically generated ontology structure (Section 3.2), it was possible to integrate and compare multiple predictive toxicology ontologies in order to identify equivalence or subclass relationships between their toxicity and bioactivity classifications. Perhaps the easiest to demonstrate is the integration of two Rule of Five-based ontologies. In one set, one of the requirements for a compound to be *drug-like* was a molecular weight less than 500 Da (Fig. 2), while in another, *small drug-like* compounds were introduced, with a molecular weight under 250 Da. Simple import of one ontology into the other and classification with Pellet resulted in *small drug-like* compounds inferred to be a subclass of *drug-like* compounds.

4 Conclusions

4.1 Significance

In this work, we have demonstrated for the first time the automated construction and practical application of OWL-encoded decision trees in chemical toxicology. The OWL ontologies that we generate can capture numerical cutoff-based rules, as well as Boolean-based rules, and can be used to represent both, automatically and expert-generated decision trees. Using our approach, decision trees that form the basis for predictive chemical toxicology classification and are either manually (expert-based systems) or algorithmically (data-based systems) generated can be routinely converted to OWL ontologies. Due to the explicit and formal specification of concepts within these ontologies, toxicity and bioactivity classes can be exposed for comparison and logical integration. In addition to this, these ontologies can also be easily applied to classify chemical entities in the rapidly growing knowledgebases of RDF-encoded chemical information. In replacing framework-, software-, and domain-specific classification engines with standard OWL ontologies, we allow for the chemical toxicology efforts to break free of their respective boundaries and support their current shift towards the Semantic Web technologies. As this shift occurs, we are confident that the work we present here will play an important role in informing future efforts in integrating and analyzing the future Chemical Semantic Web to support open, transparent, and reproducible chemical toxicology research.

4.2 Future Applications and Developments

This work marks a first step towards an OWL-based predictive toxicology framework that is currently under development. In this framework, ontologies capture the decision tree-based toxicology and bioactivity mathematical models are generated on the fly from linked open data. These ontology-specified models will subsequently be accessible for further automated classification of large collections of semantically represented chemical entities. Preliminary results point to the possibility of logically comparing formalized decision trees of multiple types so as to provide explanations for [16] and to identify points of equivalence of toxicity and bioactivity classes. Finally, the capture of classification statistics presents an interesting avenue to explore probabilistic reasoning [17] using description logics which would be well suited for toxicity prediction within a set of confidence intervals.

Acknowledgments. The authors are financially supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada, a Health Canada grant and CANARIE.

References

1. Helma, C.: In Silico predictive toxicology: The state-of-the-art and strategies to predict human health effects. *Curr. Opin. Drug Discov. Devel.* 8, 27-31 (2005)
2. Cramer, G.M., Ford, R.A., Hall, R.L.: Estimation of Toxic Hazard - A Decision Tree Approach. *J. Cosmet. Toxicol.* 16, 255-276 (1978)
3. Kroes, R., et al.: Structure based thresholds of toxicological concern. *Food Chem. Toxicol.* 42, 65-83 (2004)
4. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1999)
5. Belleau, F. et al.: Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J. Biomed. Inform.* 41, 706-716 (2008)
6. Carcinogenic Potency Database. http://www.epa.gov/ncct/dsstox/sdf_cpdbas.html
7. Patlewicz, G. et al.: An evaluation of the implementation of the Cramer classification scheme in the Toxtree software. *SAR QSAR Environ. Res.* 19, 495-524 (2008)
8. Steinbeck, C. et al.: The Chemistry Development Kit (CDK) *J. Chem. Inf. Comput. Sci.* 43, 493-500 (2003)
9. Lipinski, C.A. et al.: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development. *Adv. Drug. Del. Rev.* 46, 3-26 (2001)
10. Wishart, D.S., et al.: HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.* 37, D603-D610 (2009)
11. Hall, M. et al.: The WEKA Data Mining Software: An Update. *SIGKDD Explorations.* 11, 10-18 (2009)
12. Horridge, M., Bechhofer, S.: The OWL API. OWLED 2009, 6th OWL Experiences and Directions Workshop. Chantilly, Virginia, USA. (2009)
13. Sirin, E., Parsia, B., et al.: Pellet: A practical OWL-DL reasoner. *Software Engineering and the Semantic Web.* 5, 51-53 (2007)
14. Chemical Entity Semantic Specification. <http://semanticscience.org/projects/chess/>
15. CHEMINF. <http://semanticchemistry.googlecode.com/svn/trunk/ontology/cheminf.owl>
16. Horridge, M., Parsia, B., Sattler, U.: Laconic and precise justifications in OWL. In: Proc. of ISWC-08, LNCS. 5318, 323-338 (2008)
17. Klinov, P.: Pronto: A Non-monotonic Probabilistic Description Logic Reasoner. In: The Semantic Web: Research and Applications, LNCS. 5021, 826-830 (2008)
18. Semtox Project Page. <http://semanticscience.org/projects/semtox/>