Representation of probabilistic scientific knowledge

Larisa N. Soldatova*[1], Andrey Rzhetsky[2], Kurt De Grave[3] and Ross D. King[4]

[1] Department of Information Systems and Computing, Brunel University, London, UK.
[2] Department of Medicine & Department of Human Genetics, the University of Chicago, US.
[3] Department of Computer Science, KU Leuven, Belgium.
[4] Manchester Institute of Biotechnology, the University of Manchester, UK.

Email: Larisa N. Soldatova*- larisa.soldatova@brunel.ac.uk; Andrey Rzhetsky - arzhetsk@medicine.bsd.uchicago.edu; Kurt De Grave - Kurt.DeGrave@cs.kuleuven.be; Ross D. King - robotscientist1@gmail.com;

*Corresponding author

Abstract

The theory of probability is widely used in biomedical research for data analysis and modelling. In previous work the probabilities of the research hypotheses have been recorded as experimental metadata. The ontology HELO is designed to support probabilistic reasoning, and provides semantic descriptors for reporting on research that involves operations with probabilities. HELO explicitly links research statements such as hypotheses, models, laws, conclusions, etc. to the associated probabilities of these statements being true. HELO enables the explicit semantic representation and accurate recording of probabilities in hypotheses, as well as the inference methods used to generate and update those hypotheses. We demonstrate the utility of HELO on three worked examples: changes in the probability of the hypothesis that sirtuins regulate human life span; changes in the probability of hypotheses about gene functions in the *S. cerevisiae* aromatic amino acid pathway; and the use of active learning in drug design (quantitative structure activity relation learning), where a strategy for the selection of compounds with the highest probability of improving on the best known compound was used. HELO is open source and available at https://github.com/larisa-soldatova/HELO

KEYWORDS: ontology; knowledge representation; probabilistic reasoning

1 Introduction

"All knowledge resolves itself into probability".

David Hume, in a treatise of Human Nature (1888), 181-182.

Scientific knowledge is inherently uncertain: experimental observations may be corrupted by noise, and no matter how many times a theory has been tested there is still the possibility that new experimental observations will refute it — as famously happened to Newtonian mechanics. Probability theory has from its conception been utilized to represent this uncertainty in scientific knowledge. However the role of probability theory has proved controversial, with for example the great philosopher of science Karl Popper arguing that probabilities cannot be applied to scientific theories on the grounds that an infinite number of theories can explain any scientific data, therefore their *a priori* probabilities are zero. This view is now generally disregarded and a Bayesian approach to the use of probabilities in science is widely accepted. In Bayesian reasoning *a priori* probability estimates for hypotheses are updated through observation of additional evidence [1]. The Bayesian approach is arguably the only rational method for updating beliefs [2, 3].

Despite the undoubted importance of probabilities in science it is unfortunately the case that conventional knowledge representations in bio-medicine are insufficient to support probabilistic reasoning. The best available representation, in our view, is the Evidence Code Ontology (ECO) [4]. ECO enables the recording of evidence that supports scientific statements, e.g. experimental evidence, sequence similarity, curator inference; and also by what method the evidence was obtained, e.g. through computational combinatorial analysis, inference from background knowledge, non-traceable author statement. This information enables researchers to qualitatively evaluate the degree of uncertainty of scientific statements. However, such evaluations are rarely recorded, not checked for consistency with other relevant evaluations, and therefore are difficult to use for probabilistic reasoning. There is a need for a resource that would enable the explicit quantitative recording of probabilities associated with research statements. To address this need we propose the ontology HELO (HypothEsis and Law Ontology) that supports probabilistic reasoning about bio-medical research statements.

2 HELO aims

The HELO ontology was originally designed to support development of Robot Scientists, these are physically implemented laboratory automation systems that exploit techniques from the field of artificial intelligence to execute cycles of scientific experimentation.

A probability that a research statement is true may vary greatly depending on the source of the statement.

While experimental data from good laboratories are likely to be true, even research statements extracted from very high impact journals are not necessarily valid. C.G. Bengley and L.M. Ellis in their recent article in *Nature* report that scientific findings have been confirmed only in 6 out of 53 "landmark" studies in haematology and oncology [5]. This is consistent with results in other areas. For example Prinz *et. al* report that only 25% of published pre-clinical studies could be validated [6]. The authors stressed that validation attempts could fail for various reasons, including technical differences. HELO aims to provide a framework for the recording of probabilities that research statements are true, and for probabilistic reasoning with such statements.

3 The key HELO classes

3.1 Research statements

The HELO representation of research statements is based on the representation of research hypothesis as PREDICATE($entity_i, entity_j$) defined in an ontology LABORS, where *predicate* is a relation and *entity* is a class or instance defined in a domain ontology [7]. HELO enables one to formulate complex research statements, where basic (atomic) statements like PREDICATE($entity_i, entity_j$) are combined by logical operators $\wedge, \vee, \neg, \rightarrow, \leftrightarrow$. Entities that form research statements may be replaced by more generic entities (parent classes) and/or be specialized by their properties: individual gene names could be replaced with classes from Gene Ontology (GO) [8]; specific environmental factors could be replaced with general terms such as increased/decreased temperature, carbon source, addition of drugs, etc.; and measurable phenotypes could be replaced with general terms such as relate to growth, cell shape, etc. The following complex statement about yeast strains could be then represented: *"If all genes with lactase activity are deleted from a yeast strain and if this strain is grown in medium with lactose as the sole carbon source, then the phenotype will be no growth."* This could be expressed in logic using terms defined in various ontologies as:

$$
\begin{aligned}
&(((\forall gene, \forall yeast\_strain, \forall x \mid \\
&\text{HAS-FUNCTION}(gene, lactase\_activity) \wedge \\
&\text{HAS-PART}(yeast\_strain, gene) \wedge \\
&\text{IS-A}(process, deletion) \wedge \\
&\text{HAS-PARTICIPANT}(gene, deletion) \wedge \\
&\text{HAS-OUTPUT}(deletion, yeast\_strain) \wedge \\
&\text{HAS-PART}(growth\_medium, lactose) \wedge
\end{aligned}
$$

HAS-FUNCTION$(lactose, carbon_source) \wedge$

HAS-PART$(growth_medium, x) \wedge$

IS-A$(phenotype, no_growth) \wedge$

$\neg$HAS-FUNCTION$(x, carbon_source))$

In combination with a logical model of metabolism these statements would enable deduction of the fact:

$\rightarrow$ (HAS-QUALITY$(yeast_strain, no_growth)$).

HELO defines a hierarchy of research statements: *research hypothesis*, *hypotheses set* (a collection of hypotheses with a total probability 1, it usually combines research hypotheses, negative hypotheses, and alternative hypotheses, see [7] for more detail), *assumption*, *conclusion*, *scientific law* (models and generic rules, including Bayes rule), *theorem* (including Bayes theorem). Research laws may be represented as production rules ($statement_i \rightarrow statement_j$), where statements correspond to hypotheses, evidence, conclusions. For example,

INTERACT-PHYSICALLY$(gene - product_i, gene - product_j)$

$\rightarrow$ INTERACT-EPISTATICALLY$(gene_i, gene_j)$.

Research laws may be models that are produced for example by the Eureka system that outputs laws of nature [9].

HELO is designed to consistently accommodate scientific hypotheses and laws collected from different sources: interviews with scientists, web pages, research papers, databases, program codes. Any research statement in HELO has an associated probability of being true.

3.2 Probability

Probabilistic reasoning is essential in biomedicine, e.g. the Ontology of Adverse Events (OAE) models *causal adverse event probability* (an information content entity that represents a probability that an adverse event is caused (induced) by a medical intervention) [10], the Mass Spectrometry (MS) structured controlled vocabulary developed by the HUPO Proteomics Standards Initiative models *modification probability* (a priori probability of a modification) [11]. However, the concept of a probability is not modeled consistently

in biomedical ontologies. MS models the class *probability* as a subclass of the class *modification parameters*. The Parasite Experiment Ontology (PEO) defines the concept as a subclass of the class *statistical measure* and *data collection* (a statistical way of expressing knowledge or belief that an event will occur or has occurred) [12]. Computational Neuroscience Ontology (CNO) defines *probability* as the subclass of the class *model parameter* [13]. CNO has the class *mathematical concept* ("a thing that represents the different mathematical concepts used to represent models"), but for some reason probability is not considered as a mathematical concept. The Semanticscience Integrated Ontology (SIO) [14] defines the class *probability* as the subclass of the class *description*. SIO has the class *mathematical entity*, but again probability is not considered as such.

HELO follows the theory of probability [15], [16] and defines the class *probability* as a subclass of the class *mathematical function* to enable mathematical operations with probabilities (a mathematical function expressing knowledge or belief that a research statement is true). This definition covers frequentist probabilities (taken as a limiting frequency of experimental observations), and also Bayesian "subjective" probabilities, or beliefs.

Reliable statistical estimates of the probability of a statement being true are often unavailable. In the subjective Bayesian framework human experts are expected to provide priors that capture scientific background knowledge and intuition. Obtaining such probabilities is notoriously difficult and there is an extensive literature on the subject. Once prior probabilities are given the probabilities of scientific statements may be then iteratively updated (increased or decreased) with new evidence. It is important to record these changes in value and how they were inferred.

HELO enables the recording of how probabilities were obtained. The class *method of probability estimation* has subclasses *Bayesian inference*, *expert estimation*, *statistical calculation*, *deduction*, *abduction*, *induction*, *homological inference*; and linked to the class *procedure* that records a specific algorithm implementation for obtaining a probability of a research statement. The class *probability* has the subclasses *prior probability* and *posterior probability*. A *research statement* is linked to an associated *probability* via the functional relation HAS-PROBABILTY.

HELO imports from SIO the relations REFUTES, SUPPORTS, DISPUTES to link research statements, and the relations HAS-DISPUTING-EVIDENCE, HAS-REFUTING-EVIDENCE, HAS-SUPPORTING-EVIDENCE to link research statements and evidence (see Fig.1).

3.3 An ontology of the theory of probability

HELO defines the key entities of the theory of probability to enable logically consistent recording of operations that involve probabilities. HELO includes such classes as *variable*, *probability distribution function*, *probability mass function*, *mean*, *variance*, and such qualities as *independent*, *random (variable)*, *joint (probability)*. In order to organize these classes into a hierarchical system, HELO imports the following top-level classes: *continuant* (from BFO [17]), *information content entity* (from IAO [18]), *plan specification* (from OBI [19]), *procedure* (from LABORS [20]), *representation* (from LABORS and SIO [14]) (see Fig.2).

Additionally, the class *random event* is defined as an upper class, because the concept of an *event* ontologically differs from the notion of a *process* or any other notion. It may involve a process and participants and it has an associated time point, e.g. the end of the process. The theory of probability deals with *random events* defined on a *sample space* of all possible outcomes of a random event.

HELO is expressed in OWL-DL. It has been checked for logical consistency with the reasoners HermiT 1.3.6 and FaCT^{++}. HELO is open source and available at https://github.com/larisa-soldatova/HELO.

4 Worked examples

4.1 The S. cerevisiae aromatic amino acid pathway

This example demonstrates how a probability that a research hypothesis is true is used for automated experimentation.

King *et al.* (2009) demonstrated the full automation of scientific discovery [20]. The Robot Scientist "Adam" employed abductive inference to formulate a set of 8 hypotheses based on its logical model of the *S. cerevisiae* aromatic amino acid (AAA) pathway concerning which gene had been deleted (see Supplementary Information in [20] for more detail). The *prior probability* of each hypothesis from the set of being correct, (using a uniform distribution) was 1/8. Adam then planned and executed cycles of auxotrophic experiments to test these hypotheses. Each cycle resulted in the rejection of one or more hypotheses, and the probabilities of the remaining hypotheses were increased with each cycle. The experiments were executed until only one hypothesis was left. The *posterior probability* of the remaining hypothesis was 1 and all of the others - 0.

In making its decision about which experiment to execute in each cycle Adam used the probabilities of the hypotheses being true, the cost of the compounds required in the experiments to test those hypotheses, and the predicted information gain in testing the hypotheses. Previously, probabilities of research hypotheses were represented and recorded as associated with the experiment's metadata [21]. HELO enables the direct recording of *prior* and *posterior probabilities* as properties of research hypotheses. This makes the

representation of probabilistic knowledge explicit, and streamlines probabilistic reasoning, decision making, and automated experimentation.

4.2 Sirtuins

We use the example of sirtuins as an example of how to utilise HELO for probabilistic representation of research statements. We are interested in recording and automating the argumentation involved in the sirtuin case, both to direct our own research into aging, but also as an exemplar of biological reasoning. This example is typical in how the probability of scientific statements varies over time with the observation of new experiments. The example also illustrates the use of homologous inference, which is the basis of much biological reasoning, and which is essentially probabilistic.

Sirtuins are highly conserved NAD^+ - dependent deacetylases that are believed to play a role in regulating lifespan in many organisms. The potential role of sirtuins in extending human lifespan has led to extensive research into the human gene SIRT1 and its orthologs. For example in 2001 Tissenbaum & Guarente showed that increased dosage of the SIRT1 homolog extends lifespan in the nematode (*Caenorhabditis elegans*) [22]. Increasing sirtuin level through genetic manipulation has been observed to extend lifespan in *C. elegans*, the yeast (*Saccharomyces cerevisiae*), the fruitfly (*Drosophila melanogaster*), and the mouse (*Mus musculus*) [23, 24]. This research sparked commercial interest and in 2008 Sirtris Pharmaceuticals Inc., working on exploiting sirtuin modulation for the treatment of human disease, was bought by GlaxoSmithKline for approximately USD 720 million.

This excitement about the potential of sirtuins suffered a major setback in 2011 when Burnett *et al.* reported that overexpressing the sirtuin gene in two model organisms, *C. elegans* and *D. melanogaster* did not in fact boost longevity as had been previously reported [25]. The situation changed again in 2012 when Kanfi *et al.* reported that the sirtuin SIRT6 regulates lifespan in male mice, but not in female ones [23]. Therefore the probability of the research hypothesis that sirtuins regulate organism lifespan has increased and decreased over the last decade.

The primary research hypothesis h_1 we are interested in is: "SIRT1 regulates human life span" (SIRT1 is a sirtuin gene in humans). HELO enables the recording of this research statement:

$$\text{IS-A}(\textit{human}, \textit{organism}) \wedge$$
$$\text{HAS-QUALITY}(\textit{organism}, \textit{life-span}) \wedge$$
$$\text{REGULATES}(\textit{SIRT1}, \textit{life-span}).$$

However, it is difficult to directly test this hypothesis, so most evidence relating to it comes from laboratory experiments using model eukaryotes. For example a hypothesis h_2 is about *C. elegans*:

IS-A(*C. elegans*, *organism*)∧
HAS-QUALITY(*organism*, *life-span*)∧
REGULATES(*SIRT2*, *life-span*).

and a hypothesis h_3 is about *S. cerevisiae*:

IS-A(*S. cerevisiae*, *organism*)∧
HAS-QUALITY(*organism*, *life-span*)∧
REGULATES(*SIRT2*, *life-span*).

The evidence about h_2 and h_3 is then related to h_1 by probabilistic reasoning (homological inference):

HAS-PROBABILITY$(h_2, p_2) \rightarrow$ HAS-PROBABILITY(h_1, p_{12})
HAS-PROBABILITY$(h_3, p_3) \rightarrow$ HAS-PROBABILITY(h_1, p_{13}).

SIR2 is the *S. cerevisiae* homolog of SIRT1. The research hypothesis h_3 "SIR2 regulates yeast life span" is very well supported by the scientific literature. The dataset yeast70 is a subset of GeneWays 7.0 database [26]. The GeneWays 7.0 database was produced through automated analysis of 368,331 full-text research articles and 8,039,972 article abstracts from the PubMed database, using the GeneWays system. The database covers a wide spectrum of molecular interactions, such as bind, phosphorylate, glycosylate, and activate (nearly 500 relations in total). The dataset yeast70 has 1,135 sentences containing the keyword "aging" and yeast gene names, 492 of them contain SIR2, the gene SIR1 is not mentioned, and SIR3 is mentioned 42 times. Examining these papers suggests that the probability of h_2 is close to 1.0.

The probability of scientific hypotheses changes with new evidence, and we wish to use HELO to represent this. For example Burnett *et al.* results directly decreased the probability of the hypotheses regarding the function of the SIRT1 homologs in *Caenorhabditis elegans* and *Drosophila melanogaster*, and these indirectly

decreased the probability of h_1. The situation changed again in 2012 with Kanfi *et al.* where the evidence directly increased the probability of the hypotheses regarding the function of the SIRT1 homolog in *Mus musculus*, and this indirectly increased the probability of h_1 (see Fig.1).

Of course the weight of the evidence in these papers on h_1 depends on a host of factors other than simply the model species involved: the amount and variety of evidence, its statistical confidence, the lab where the work was done, the publisher, etc. Taking all these into account an expert estimate of the probability that h_1 is held after the publication of the paper [22] is 0.8 (see Fig.1). It should be noted that the exact probability of h_1, say 0.8 or 0.82, is not that critical. What is important is the "ball-park" figure, and the direction of change with new evidence. Our idea is that addition of more and more evidence and inferences to the argument constrain the probabilities to reasonable numbers. It is our contention that all human scientists make such implicit inferences and much is to be gained by making them explicit. In addition these probability can be used for further automated inference and experimentation.

Experiments with a Sir2 deletant strain run within the Robot Scientists project showed no difference between the wild type, while yeast strains with NAD^+ grew to a significantly higher biomass than the wild type. The experiments demonstrate that Sir2 functions differently from other NAD^+ genes, and this indirectly supports the hypothesis h_1. It is clear that further experimentation is required to accept or reject the hypothesis h_1.

4.3 Active learning for drug discovery

This example demonstrates the recording of probabilities in drug discovery experiments. The goal of these experiments was to find the best compound (with respect to a biomedical assay, e.g. for treating cancer) without having to test all the compounds against the assay. This involved learning quantitative structure activity relationships (QSARs). These are functions that take as input the structure of a compound and output an estimate of how well the compound will perform in a biomedical assay. The investigation was computational and used existing assay results.

The task of finding the best instance (e.g. compounds, parameters) as evaluated on an unknown target function (e.g. high biological activity, minimal costs) using limited resources (e.g. time) is important to many scientific disciplines. In drug discovery it is not sufficient to find just a single best compound or "lead" as several leads improve the chances of finding a compound that passes toxicology tests. The challenge therefore is to identify the k best performing instances (= compounds in this context) using as few experiments as possible. We refer to this task as *active k-optimization*.

We applied machine learning to solve the *active k-optimization* task and to propose the best candidates for screening [27]. We considered several selection strategies for the best instances: Cox and John's lower confidence bound criterion [28] (we refer to it as the optimistic strategy), the most probable improvement (MPI) of the current solution strategy [29], the maximum expected improvement (MEI) strategy, and also the random choice (see [27] for more detail).

These strategies were evaluated on the US National Cancer Institute 60 anticancer drug screen (NCI60) dataset [30]. This repository contains measurements of the inhibitory power of tens of thousands of chemical compounds against 59 different cancer cell lines (one of the originally 60 cell lines was evicted because it was essentially a replicate of another one [31]). NCI reports the negative log-concentration required for 50% cancer cell growth inhibition (pGI_{50}) as well as cytostatic and cytotoxic effect measures, but we only used the pGI_{50}.

The goal is to find compounds in a library that have a high pGI_{50}, and to do so using as few pGI_{50} measurements as possible. The program bootstraps by selecting 10 random compounds and measuring their pGI_{50}. In each subsequent step, a current QSAR model is fitted to all available pGI_{50} values. The model is used to predict the pGI_{50} for all remaining (untested) compounds in the library. The model is a Gaussian process, which outputs a (Normal) distribution for the pGI_{50} value rather than only a point prediction. This enables the implementation of the previously listed strategies. For example, for the MPI strategy, one computes the probability that a compound has a pGI_{50} which is larger than the current k-th best one. The compound with the highest probability is selected for the next measurement of pGI_{50}.

Table 1 illustrates MPI for a particular cell line 786-0, for a specific bootstrap, and for $k = 1$. The first column of the table shows the number of known pGI_{50} values at that time. $P1$ is the probability, given the current evidence, that a particular compound NSC 642567 will have a pGI_{50} better than the best bootstrap compound. The subsequent column shows what is the probability $P2$ that NSC 642567 has a better pGI_{50} than the current best value. The third column shows the highest such probability $P3$ for any of the compounds remaining in the library.

Each computational experiment was repeated 20 times and the results were averaged. Overall, on the NCI60 datasets, the optimistic strategy was most robust. In all situations considered, it performed either best or not significantly worse than the best strategy (see [27] for detail and diagrams). The performance of MPI is competitive for medium experimental budgets, but it may fail to find more than one good compound when constrained to low budgets, and it does not optimally exploit high budgets. MEI is a very good strategy when about 10 compounds are needed. The random selection strategy performs worse than all

other selection methods in all settings. Actively choosing compounds substantially speeds up the finding of the compounds with high pGI_{50}.

HELO enables the recording of these important results in a semantically defined way. The following semantic descriptors are required for the reporting of this study: Gaussian distribution, zero mean, variance, prior belief, posterior probability, random variable, likelihood, estimated probability. HELO contains exact matching terms or equivalent synonyms of the required semantic descriptors, for example HAS-VALUE(mean,0) is equivalent to zero mean.

5 Conclusion

Scientific knowledge is inherently uncertain. There is therefore a need for a representation that focuses on the probabilistic features of research statements, and supports probabilistic reasoning. In order to address this need we proposed the ontology HELO that supports probabilistic reasoning over uncertain scientific statements. HELO defines a hierarchy of typical research statements and links them to their associated probabilities, and methods of obtaining those probabilities. We demonstrated HELO on the representation of scientific belief that sirtuins regulate organism life span, and regarding deleted genes in the *S. cerevisiae* aromatic amino acid pathway. In both cases the probability of research statements changed with new evidence, and it is clearly important to employ the most updated probability estimate for making decisions about research involving these genes. The active learning for drug discovery study is based on operations with probabilities. The probabilities of having a high pGI_{50} were iteratively computed for all compounds in the library, and the best compounds were chosen for further study. HELO enables accurate recording of supporting and refuting evidence of research statements, and how they participate in the process of updating probability values.

HELO is specifically designed to support the cycles of automatic scientific discovery that incorporate text mining, machine learning, robotic automation, and knowledge representation, and may be of use for other areas of research that involve probabilistic reasoning.

Author's contributions

Larisa N. Soldatova originated the idea of ontological representation of the key entities of the theory of probability and worked on the ontology HELO. Andrey Rzhetsky originated the idea of annotating research

statements extracted from natural language text with semantic descriptors indicating the level of truthfulness of those statements, and also the recording of competing and contradictory statements. Kurt De Grave applied HELO for the reporting on the active learning for high throughput screening study. Ross D. King originated the idea of using probabilities of hypotheses for the choice of experiments in automated experimentation. He also contributed to the development of HELO and other worked examples.

Acknowledgements

This work was partially supported by grant BB/F008228/1 from the UK Biotechnology & Biological Sciences Research Council, from the European Commission under the FP7 Collaborative Programme, UNICELLSYS, KU Leuven GOA/08/008 and ERC Starting Grant 240186.

References

1. Barber D: *Bayesian Reasoning and Machine Learning*. Cambridge University Press 2012.
2. Gillies D: *Philosophical Theories of Probability*. Routledge 2000.
3. Howson C, Urbach P: *Scientific reasoning: The Bayesian approach*. Chicago: Open Court 1993.
4. **Evidence Code Ontology** [www.obofoundry.org/cgi-bin/detail.cgi?id=evidence_code].
5. Bengley CG, Ellis LM: **Raise standards for preclinical cancer research**. *Nature* 2012, **483**:531–533.
6. Prinz F, Gilliand DG: **Drug development and clinical trials - the path to an approved cancer drug**. *Natue Rev.Clin. Oncol.* 2012, **22**.
7. Soldatova L, Rzhetsky A: **Representation research hypotheses**. *J. of Biomedical Semantic* 2011, **2**(2):S9.
8. **Gene Ontology** [www.geneontology.org].
9. Schmidt M, Lipson H: **Distilling Free-Form Natural Laws from Experimental Data**. *Science* 2009, :81–85.
10. **Ontology of Adverse Events** [www.oae-ontology.org].
11. **Mass Spectrometry** [bioportal.bioontology.org/ontologies/1105].
12. **Parasite Experiment Ontology** [bioportal.bioontology.org/ontologies/1335].
13. **Computational Neuroscience Ontology** [bioportal.bioontology.org/ontologies/3003].
14. **Semanticscience Integrated Ontology** [code.google.com/p/semanticscience/wiki/SIO].
15. Ross SM: *Introduction to probability Models*. Academic Press 2007.
16. Lee PM: *Bayesian Statistics*. New York: Hodder Arnold 2004.
17. **Basic Formal Ontology** [www.ifomis.org/bfo].
18. **Information Artifact Ontology** [code.google.com/p/information-artifact-ontology].
19. **Ontology for Biomedical Investigations** [obi-ontology.org].
20. King R, Whelan K, Jones F, et al: **Functional genomic hypothesis generation and experimentation by a robot scientist**. *Nature* 2004, **427**:247–252.
21. King R, Rowland J, Oliver SG, Young M, Aubrey W, Byrne E, Liakata M, Markham M, Pir P, Soldatova LN, Sparkes A, Whelan K, Clare A: **The Automation of Science**. *Science* 2009, **324**(5923):85–89.
22. Tissenbaum HA, Guarente L: **Increased dosage of a sir-2 gene extends lifespan in Caenorhabditis elegans**. *Nature* 2001, **410**:227–230.
23. Kanfi Y, Naiman S, Amir G, et al: **The sirtuin SIRT6 regulates lifespan in male mice**. *Nature* 2008, **483**(7388):218–21.
24. Michan S, Sinclair S: **Sirtuins in mammals: insights into their biological function**. *Biochem.* 2007, **404**:1–13.
25. Burnett C, Valentini S, Cabreiro F, et al: **Absence of effects of Sir2 overexpression on lifespan in C. elegans and Drosophila**. *Nature* 2011, **477**:482–485.
26. Iossifov I, Rodriguez-Esteban R, Mayzus I, Millen K, Rzhetsky A: **Looking at cerebellar malformations through text-mined interactomes of mice and humans**. *PLoS Comput Biol* 2009, **5**(e1000559).
27. De Grave K, Ramon J, De Raedt L: **Active Learning for High Throughput Screening**. In *Proceedings of the Eleventh International Conference on Discovery Science, Volume 5255 of* Lecture Notes in Computer Science, Springer 2008:185–196.
28. Cox D, John S: **SDO: a statistical method for global optimization.** In *Multidisciplinary Design Optimization*. Edited by Hampton VA, Philadelphia, PA: SIAM 1997:315–329.
29. Kushner H: **A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise**. *Journal of Basic Engineering* 1964, :97–106.
30. Shoemaker R: **The NCI60 human tumor cell line anticanser drug screen**. *Nat. Rev. Canser* 2006, **6**:813–823.
31. Nishizuka S, et al.: **Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays**. *PNAS* 2003, **100**(24):14229–14234.

Figures

Figure 1 - An example of the HELO representation of a research statement.

The figure shows the representation of the values of the prior and posterior probabilities of the research statement about sirtuins, and also the supporting and refuting evidence.

Figure 2 - An overview of the ontology HELO.

The figure shows the top-level classes of HELO and some of their extentions.

Tables

Table 1 - The probabilities that the selected compounds have high GI_{50}