Towards the Automation of Scientific Method

Oliver Ray
Department of Computer Science
University of Bristol
Bristol, BS8 1UB, UK
oray@cs.bris.ac.uk

Amanda Clare and **Maria Liakata** and **Larisa Soldatova** and **Ken Whelan** and **Ross King**
Department of Computer Science
University of Aberystwyth
Ceredigion, SY23 3DB, UK
{afc,mal,lss,knw,rdk}@aber.ac.uk

Abstract

The automation of scientific method is a subject of increasing intellectual and practical interest, with potentially great benefits to science and society. This paper discusses four key challenges in this task and explains how they have been addressed within a functional genomics project known as the Robot Scientist. In so doing, it describes how abduction and induction have enabled the automatic revision of metabolic models through a synthesis of cutting edge artificial intelligence and laboratory robotics. Our aim is to summarise the progress which has already been made and to set out an agenda for further technological and social changes that are needed to turn the automation of science into a truly useful reality.

1 Introduction

Scientific method is the continual cycle of experiment and analysis used in all branches of science to develop rational models of empirical data. Analyses lead to models that might account for observed data. Experiments lead to data that could test inferred models. Each builds on the other.

Since the first national academies of science were founded in the latter half of the 17th century, the scientific method has transformed our understanding of ourselves and of the world around us. In so doing, it has led to unprecedented cultural developments that continue right up to the present day.

But we believe a new scientific revolution is in the making with the potential to further accelerate progress in all areas of science. It being driven by the recent explosion of robotics, computing, and the internet, which are now starting to have a major impact on the practice of science.

Many branches of science are now heavily reliant on industrial-strength experimental apparatus and high-throughput computational techniques: as can be clearly seen in a spectrum of disciplines ranging from particle physics through the biological sciences all the way up to astronomy.

In addition, the world-wide-web is changing the way that scientists communicate and collaborate. The rapid growth of on-line publishing and indexing sites has made scientific papers more accessible; and a corresponding surge in public data repositories has done the same for results.

The enthusiastic adoption of web-based tools like blogs, wikis, RSS feeds, social networks and grid computing has led to a trend dubbed *Science 2.0* [Waldrop, 2008] of greater cooperation among scientists: with some even promoting a fully public *Open Notebook* approach [Bradley, 2007].

Yet this vision of science in the 21st century is beginning to attract some intense controversy as the focus shifts from its obvious technological benefits to some of its profound social implications that fundamentally challenge the way scientists have worked for more than three hundred years.

The strong pressure on human scientists and institutions to enhance their reputations and increase their wealth through exclusive publications and patents promotes some amount of secrecy which many are duly worried may be compromised by the new agenda of transparency.

By contrast, we are concerned with just one issue: how can technology help us to do better science? We are interested in human affairs only to the extent that they help or hinder the scientific process. Our goal is to better understand and support the roles of both man and machine.

While this may well raise some sensitive issues and hard decisions, we believe the potential benefits are simply too great to ignore. Instead, we take inspiration from Alhazen who, nearly a millennium ago, wrote in defence of what we now call scientific method:

> "Truth is sought for its own sake. And those who are engaged upon the quest for anything for its own sake are not interested in other things. Finding the truth is difficult, and the road to it is rough."

In fact, we do appreciate that pragmatic considerations must play a key role in science, just as they do in any other field of endeavour. Our point is that such factors must be clearly understood and fully integrated into the scientific life-cycle in a principled and properly supported way.

From the outset, it is clear that a major obstacle to the more effective use of technology in science is due to the way that scientific results are reported and recorded; which for over three centuries has been exclusively done through published journal articles and unpublished paper notebooks.

Just representing this information in a more formal way would greatly facilitate the use of computational tools and help reduce the ambiguities of natural language. In turn, this would make it easier to store, access, inspect, query, validate and reuse analytical models and experimental data.

Moreover, the use of formal knowledge representations would allow for semantic markup and ontologies that would enable the exploitation of more sophisticated reasoning services based on emerging Semantic Web technologies [Berners-Lee *et al.*, 2001; Hendler, 2003].

This insight is leading to a view of *Semantic Science* [Poole *et al.*, 2008] which advocates the publication of theories and data in machine-readable form. The idea is that theories will be used to automatically generate predictions from data, while data will be used to automatically evaluate theories.

While the realisation of these aims would already provide significant advantages over the traditional practice of science, we believe it is only the first step in what will prove to be a far more profound integration of machine-based support into all aspects of the scientific lifecycle.

To fully exploit such technologies, the experimental and analytical phases of the scientific process must be formalised along with the resulting data and theories. This will bring rigour to the specification of scientific protocols and will enable the automation of their design and execution.

Our vision is to develop an autonomous cycle of scientific activity that combines theoretical analyses and experimental interventions in way that can be understood and guided by humans; and we argue much this is possible using methods from artificial intelligence and laboratory robotics.

This paper discusses four challenges in the automation of science and explains how they have been addressed in a project called the Robot Scientist [King *et al.*, 2004; 2009]. In so doing, it describes how abduction and induction were used to revise a state-of-the-art scientific model.

The paper is structured as follows. Section 2 introduces the Robot Scientist application domain. Section 3 outlines the techniques we have developed for formalising scientific data and models. Sections 4 and 5 describe the hard- and soft-ware we have developed for performing scientific experiments and analyses. Section 6 summarises what has been achieved and what is still to do in terms of closing the loop.

2 Selecting a Domain

To completely mechanise just one iteration of scientific method, in any domain, raises so many technical challenges that, at least for the time being, the automation of science will most likely have to be studied in the context of just a few well-defined target applications.

The Robot Scientist project focuses on a certain type of microbial growth experiment involving yeast, *S. cerevisiae*. Each such experiment measures the growth of a yeast strain (from which some genes have been removed) in a growth medium (to which some nutrients have been added).

These experiments are a classic method for understanding the function of deleted genes and the role that they play in the living cell. Furthermore, several additional factors mean this specific domain is an ideal target for automation from the perspective of both experiment and analysis:

First is the availability of yeast strains and growth nutrients needed as inputs in the experiments. Second is the availability of laboratory hardware needed to carry out the main steps of the experiments. Third is the availability of a public databases with information needed to interpret the experiments.

Even though the experimental and analytic aspects of this domain are relatively well understood, the work we describe below will show the considerable difficulties and significant benefits arising from their automation. Fortunately, much of this work will be relevant in other domains too.

3 Formalising the Domain

Having chosen an appropriate domain, any relevant data, models, experiments and analyses must be represented in enough detail for machines to perform the required tasks. As a matter of fact, this level of formality is desirable to ensure objectivity whether automation is intended or not.

While some aspects of science are highly mathematical, very often the description of experiments are treated much too informally. Moreover, even when scientific data and models are formally specified, they often leave out a lot of important assumptions and meta-data.

In the Robot Scientist project we utilise formal ontologies and logical representations to ensure the accurate encoding of all necessary information. More precisely, we have developed two separate ontologies to characterise various high-level and low-level elements of an experimental investigation.

First there is a scientific EXPeriments Ontology (EXPO) [Soldatova and King, 2006] for experimental design and methodology. Second there is an ontology of EXperimental ACTions (EXACT) [Soldatova *et al.*, 2008] for experimental protocols and physical manipulations.

Two example fragments of these ontologies are shown for a typical experiment in Figures 1 and 2, respectively. Currently, EXPO meta-data is automatically stored for every experiment conducted by the Robot Scientist.

Other interesting applications of these ontologies can be found in [Soldatova and King, 2006] and [Soldatova *et al.*, 2008] where they were used to critique three arbitrary studies on the phylogeny of solenodons, the mass of the top quark, and the preparation of yeast gene deletion cassettes.

⟨scientific experiment⟩ :	
⟨administration info⟩ :	
⟨title⟩ :	Robot scientist
⟨ID⟩ :	exp200401113-0001
⟨classification by domain⟩ :	
⟨DDC(Dewey)⟩ :	576 Microbiology
⟨research hypothesis⟩ :	
⟨natural language⟩ :	Knocked out gene named "yer152c" (= met8) has the function named "2.6.1.39" (=2-aminoadipate:2-oxoglutarate aminotransferase)
⟨artificial language⟩ :	encodes(yer152c, 2.6.1.39)
⟨null hypothesis⟩ :	
⟨artificial language⟩ :	¬encodes(yer152c, 2.6.1.39)
⟨alternative hypothesis⟩ :	
⟨natural language⟩ :	⟨time effect⟩ : maturation effect (incubator too cold) ⟨object effect⟩ : no entry of metabolite into the cells ⟨object effect⟩ : cross contamination
⟨domain model⟩ :	
⟨language⟩ :	Prolog.
⟨reference⟩ :	Whelan, K.E. & King, R.D. (2005). Using a logical model to predict the growth behaviour of yeast cell cultures. Tech. Report, UWA-DCS-05-045.
⟨experimental design⟩ :	
⟨subject⟩ :	The Robot Scientist
⟨object⟩ :	S. cerevisiae
⟨experimental model⟩ :	
⟨factor⟩ :	Strains: wild-type [Mat A, by4741]; knockout [yer152c]
⟨factor⟩ :	Metabolites: minimal media; additional compound [xxx]

Figure 1: EXPO annotation of a micro-biological experiment (a fragment, from [Soldatova and King, 2006])

⟨operating procedure⟩ :	grow yeast culture
⟨pre-condition⟩ :	sealed yeast colonies plate located_in cold room
⟨pre-condition⟩ :	YPD media bottle located_in cold room
⟨experiment action⟩ :	move 1
...	
⟨experiment action⟩ :	add 17
⟨component 1⟩ :	single yeast colony
⟨start volume⟩ :	small volume
⟨start container⟩ :	sealed yeast single colonies plate
⟨end container⟩ :	YPD conical flask
⟨equipment⟩ :	inoculating loop
⟨experiment action⟩ :	rename 18
⟨old name⟩ :	YPD conical flask
⟨new name⟩ :	yeast culture flask
⟨end location⟩ :	incubator
⟨experiment action⟩ :	move 19
⟨object⟩ :	yeast culture flask
⟨start location⟩ :	laminar flow hood
⟨end location⟩ :	incubator
⟨experiment action⟩ :	incubate 20
⟨object⟩ :	yeast culture flask
⟨start equipment⟩ :	shaking incubator
⟨rpm⟩ :	200
⟨tempo⟩ :	30'C
⟨time interval⟩ :	12-24h
⟨goal⟩ :	grow yeast until medium becomes cloudy
⟨post-condition⟩ :	yeast culture located_in incubator

Figure 2: EXACT annotation of a micro-biological experiment (a fragment, from [Soldatova *et al.*, 2008])

To automate the analytical phase of the scientific method, it is also necessary to obtain a formal description of the target domain. For the Robot Scientist project we have developed a detailed model of the *S. cerevisiae* metabolism [Whelan and King, 2008], part of which is shown in Figure 3.

Nodes in this figure represent metabolites involved in the transformation of the compound Glycerate-2-phosphate into the amino acids Tyrosine, Phenylalanine, and Tryptophan. Arrows denote chemical reactions converting their substrates (the tails) into products (the heads).

Each node is labelled with a KEGG identifier (in red); and each arrow is annotated with a 4-part EC number (in blue) and a set of genes (in green) called an enzyme-complex. The singly dashed line denotes the inhibition of an enzyme by a metabolite. The doubly dashed line is the cell membrane.

All reactions are assumed to take place in the cell cytosol using nutrients imported from the growth medium; and they are all assumed to proceed at a standard rate (within 1 day), except for the import of two italicised compounds, which take longer (between 1 and 2 days).

The additional nutrients and knockout genes used in each experiment are denoted by atoms of the form `additional_nutrient(e,m)` and `knockout(e,g)`, for an experiment `e`, gene `g`, and metabolite `m`. Nutrients common to all experiments are denoted `start_compound(m)`.

By definition, all start compounds and additional nutrients are in the growth medium on any day in any experiment. This is represented by two logical rules which state that a certain metabolite is in a specific compartment on a given day in a particular experiment:

```
in_compartment(Exp,Met,medium,Day) :-
    start_compound(Met).

in_compartment(Exp,Met,medium,Day) :-
    additional_nutrient(Exp,Met).
```

Reactions, genes and complexes are all assigned unique identifiers so that atoms of the form `catalyst(r,c)` and `component(g,c)` can be added to the model to in order to denote the denote the fact that complex `c` catalyses reaction `r` and the fact that gene `g` participates in complex `c`.

The inhibition of a complex `c` by a metabolite `m` is denoted `inhibitor(c,m)`. Any metabolites that are essential to the cell growth, such as the three amino acids, are specified as such by adding ground atoms of the form `essential_compound(m)`.

Cell development is arrested if an essential metabolite is not in the cytosol. But, if development is not arrested, then growth is predicted. A complex is deleted if a component gene is knocked out; and it is inhibited if some inhibitor is present (in high concentration) as an additional nutrient:

```
arrested(Exp,Day) :-
    essential_compound(Met),
    not in_compartment(Exp,Met,cytosol,Day).

predicted_growth(Exp,Day) :-
    not arrested(Exp,Day).

deleted(Exp,Cid) :-
    component(Orf,Cid),
    knockout(Exp,Orf).

inhibited(Exp,Cid) :-
    inhibitor(Cid,Met),
    additional_nutrient(Exp,Met).
```

To complete our background theory, it remains to give a logical encoding of the metabolic reactions. To facilitate the addition and removal of reactions, they are each given one of three degrees of belief: $certain$ (i.e., definitely in the model), $retractable$ (i.e., initially in the model, but can later be excluded), or $assertable$ (i.e., initially out of the model, but can later be included). Note that this allows us to consider reactions from related pathways or organisms for inclusion in a revised network; which is common practice as it ensures all newly introduced reactions are biologically feasible.

For every reaction, one rule is added to the theory for each product. Each rule states that the product will be in its compartment if (i) all substrates are in their respective compartments, (ii) there is an enzyme-complex catalysing the reaction whose activity is not inhibited and whose genes are not deleted, (iii) sufficient time has passed for the reaction to complete, and (iv) the reaction has not been excluded (if it is retractable) or it has been included (if it is assertable). As an example, the following is one of two rules produced for reaction 2.5.1.19 with id 31, assuming it is retractable:

```
in_compartment(Exp,"C01269",cytosol,Day) :-
    in_compartment(Exp,"C00074",cytosol,Day),
    in_compartment(Exp,"C03175",cytosol,Day),
    catalyst(31,Cid),
    not inhibited(Exp,Cid),
    not deleted(Exp,Cid),
    Day >= 1,
    not exclude(31).
```

As explained in [Ray *et al.*, 2009], for every start compound and additional nutrient, `m`, we assume there is an import reaction which takes `m` from the `medium` into the `cytosol`; and to each reaction with no known catalysts, we attribute an `unknown` catalyst (so all reactions are assumed to proceed in the absence of evidence to the contrary). As a result, our model of the AAA pathway actually contains 22 metabolic reactions and 23 import reactions.

4 Performing Physical Experiments

The physical mechanisation of scientific experiments poses significant technical challenges. The Robot Scientist is an autonomous laboratory platform that is able to conduct growth experiments with no human assistance. As shown by the schematic in Figure 4, the hardware includes freezers, incubators, liquid handlers, plate readers, centrifuges, washers, robot arms and environment sensors, all contained in an air-filtered enclosure. The system is able to extract particular strains of yeast from the freezer, culture them in rich growth media, harvest the cells, inoculate them into defined media, and monitor the resulting growth over a period of several days by taking regular optical density readings. In addition, it annotates all experiments with appropriate meta-data and runs them many times to ensure statistically significant results. Moreover, it schedules all these activities in a way that makes efficient use of available resources. Further details can be found in [King *et al.*, 2009].

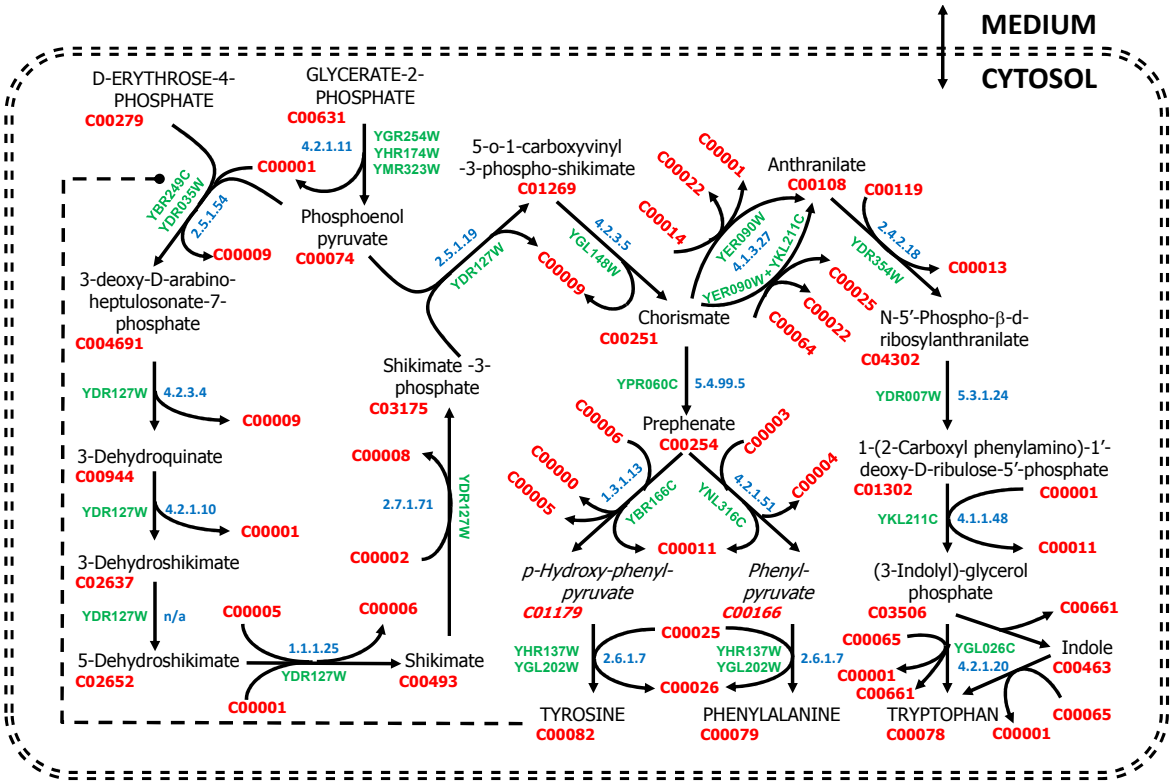

Figure 3: Aromatic Amino Acid (AAA) biosynthesis pathway of *S. cerevisiae*

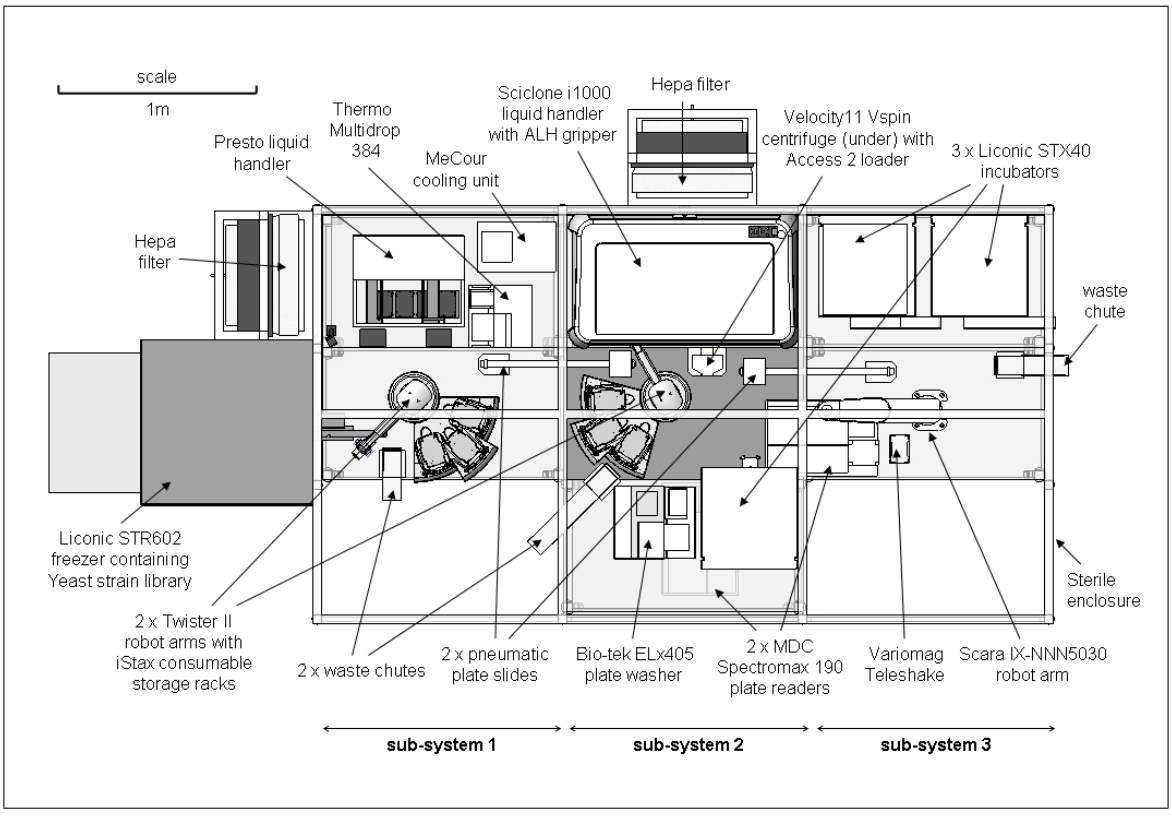

Figure 4: Schematic view of Robot Scientist hardware

5 Performing Theoretical Analyses

Over the course of time, scientific models are often revised as discrepancies begin to emerge between observed and predicted results. This is usually accomplished by hand, as in the case of the AAA pathway of Figure 3, which was mainly derived from the KEGG database, but was manually crafted in places. Our goal is to automate this revision process using computational tools.

In recent work, we have used a non-monotonic reasoning system called eXtended Hybrid Abductive Inductive Learning (XHAIL) [Ray, 2009] to do just this. The main advantage of XHAIL over other systems is that it provides well-defined semantics and proof procedure for abduction and induction over extended logic programs (which include operators for classical and default negation).

Non-monotonicity is useful as it allows to reason with defaults and exceptions and to hypothesise revisions involving the removal of information from an incorrect model as well as the addition of information into an incomplete model. XHAIL achieves this by generalising traditional techniques of language and search bias to the non-monotonic case.

First, the hypothesis space is controlled by a set of mode declarations [Muggleton, 1995] that allow the user to constrain which literals may appear in the heads and bodies of hypothesis clauses. Second, a compression heuristic [Muggleton, 1995] is used to select between competing hypotheses by preferring solutions with the fewest number of literals.

After utilising some artificial data in [Ray *et al.*, 2009] to validate the capability of XHAIL for adding and removing metabolic reactions, enzyme inhibitions or gene functions, we applied the method to try and revise our state-of-the-art AAA model in the light of real growth data collected by the Robot Scientist.

Upon submitting the observed growth results to XHAIL, we discovered they were inconsistent with the AAA model, thereby suggesting that a revision was necessary. By systematically varying the language bias, we very quickly obtained several hypotheses that achieved logical consistency between predicted and observed growth.

XHAIL produced two hypotheses of interest. The first one, was interpreted as stating that the import of anthranilate is a slow reaction (just like the import of phenyl-pyruvate and P-hydroxy-phenyl-pyruvate). The second one, was interpreted as stating that our source of Indole (`"C00463"`) is contaminated with Tryptophan (`"C00078"`). This is plausible as Indole can be synthesised from Tryptophan (by essentially reversing reaction 4.2.1.20) but any unconverted Tryptophan may be hard to separate off. Using mass spectrometry we have verified that was indeed the case.

It turns out these two hypotheses are sufficient to restore logical consistency with all of the growth measurements made by the Robot Scientist. As a result of this work, the first hypothesis has enabled us to improve our state-of-the-art AAA model; and the second hypothesis has alerted us to a source of systematic experimental error that we shall take care to avoid in future work.

6 Closing the Loop

The final step involves the integration of our hardware and software in a completely automated cycle of scientific discovery that can run indefinitely. Even though we have automated one iteration of the scientific cycle, it is not yet possible to loop arbitrarily without human assistance. In order to progress, we urgently need to develop methods for designing experiments to test competing hypotheses.

7 Related Work

Our approach builds upon earlier Robot Scientist work, which used a reasoning system called Progol5 [Muggleton and Bryant, 2000] to rediscover gene-enzyme mappings removed from the AAA pathway [King *et al.*, 2004]. However, XHAIL overcomes several key limitations of Progol5, including its inability to reason hypothetically through negation and its inability to infer more than one clause in response to any given example. For these reasons the logical model used by Progol5 employs a complex list-based representation of reactions over a logic program in which all negations are restricted to built-in predicates and where most of the code is devoted to procedural issues such as pruning the search tree, avoidance of cyclic computations, and efficient sorting of data structures. As a result, the earlier Progol5 model is restricted to the learning of new gene-enzyme mappings in single gene deletion experiments.

By contrast, the XHAIL model employs a completely declarative representation, adapted from [Whelan and King, 2008] but extended with support for enzyme-complexes, which imposes no a-priori constraints on the learning task and can be applied to multiple gene deletion experiments and simultaneously add or remove reactions, inhibitions and complexes. We note that this flexibility depends upon the ability of XHAIL to reason nonmonotinically and to infer multiple clauses in response to a single example. Moreover, XHAIL's compression heuristic is also appropriate because it ensures the revisions are in some sense minimal.

Related work in [Juvan *et al.*, 2005; Dworschak *et al.*, 2008; Papatheodorou *et al.*, 2005] and [Baral *et al.*, 2004] apply logic-based approaches to the identification of genetic regulatory networks and signalling pathways. But they do not incorporate enzyme and reaction data. Numerical techniques based on ordinary differential equations for analysing metabolic flux [Varma and Palsson, 1994] allow quantitative simulation of metabolic networks and can be used for parameter estimation. But they cannot be used to suggest structural refinements to metabolic networks.

8 Conclusions

This paper described recent progress in automating scientific method within a functional genomics domain. In particular, it described how a state-of-the-art metabolic network model was successfully revised through the abductive and inductive analysis of experimental data acquired by a Robot Scientist. This demonstrates the significant benefits to be gained by integrating abduction and induction in a biological context and it highlights the utility of non-monotonic logical reasoning to enable both the addition and removal of information.

Acknowledgments

Part of this work was funded by the BBSRC, EPSRC and RCUK. We acknowledge Wayne Aubrey for his help with Figure 2. We thank Kate Martin for performing the mass spectrometry experiments. We are grateful to Andrew Sparkes for preparing Figure 4.

References

[Baral *et al.*, 2004] C. Baral, K. Chancellor, N. Tran, N.L. Tran, A. Joy, and M. Berens. A knowledge based approach for representing and reasoning about signaling networks. In *Proc. 12th Int. Conf. on Intelligent Systems for Molecular Biology*, pages 15–22, 2004.

[Berners-Lee *et al.*, 2001] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001.

[Bradley, 2007] J. Bradley. Open notebook science using blogs and wikis. *Nature Prepress*, 2007.

[Dworschak *et al.*, 2008] S. Dworschak, S. Grell, V. Nikiforova, T. Schaub, and J. Selbig. Modeling Biological Networks by Action Languages via Answer Set Programming. *Constraints*, 13(1/2):21–65, 2008.

[Hendler, 2003] J. Hendler. Science and the semantic web. *Science*, 299(5606):520–521, 2003.

[Juvan *et al.*, 2005] P. Juvan, J. Demsar, G. Shaulsky, and B. Zupan. GenePath: from mutations to genetic networks and back. *Nucleic Acids Res.*, 33, 2005.

[King *et al.*, 2004] R. King, K. Whelan, F. Jones, P. Reiser, C. Bryant, S. Muggleton, D. Kell, and S. Oliver. Functional Genomic Hypothesis Generation and Experimentation by a Robot Scientist. *Nature*, 427:247–252, 2004.

[King *et al.*, 2009] R. King, J. Rowland, S. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. Soldatova, A. Sparkes, K. Whelan, and A. Clare. The automation of science. *Science*, 324(5923):85–89, 2009.

[Muggleton and Bryant, 2000] S. Muggleton and C. Bryant. Theory Completion Using Inverse Entailment. In *Proc. of the 10th Int. Conf. on Inductive Logic Programming*, volume 1866 of *LNCS*, pages 130–146. Springer, 2000.

[Muggleton, 1995] S. Muggleton. Inverse Entailment and Progol. *New Generation Comp.*, 13(3-4):245–286, 1995.

[Papatheodorou *et al.*, 2005] I. Papatheodorou, A. Kakas, and M. Sergot. Inference of gene relations from microarray data by abduction. In *Proc. 8th Int. Conf. on Logic Programming and Nonmonotonic Reasoning*, volume 3662 of *LNCS*, pages 389–393. Springer, 2005.

[Poole *et al.*, 2008] D. Poole, C. Smyth, and R. Sharma. Semantic science: Ontologies, data and probabilistic theories. In *Uncertainty Reasoning for the Semantic Web I*, volume 5327 of *LNCS*, pages 26–40. Springer, 2008.

[Ray *et al.*, 2009] O. Ray, K. Whelan, and R. King. A nonmonotonic logical approach for modelling and revising metabolic networks. In *Proc. 3rd Int. Conf. on Complex, Intelligent and Software Intensive Systems*, pages 825–829. IEEE, 2009.

[Ray, 2009] O. Ray. Nonmonotonic Abductive Inductive Learning. *Journal of Applied Logic*, 3(7), 2009.

[Soldatova and King, 2006] L. Soldatova and R. King. An ontology of scientific experiments. *Journal of the Royal Society, Interface*, 3(11):795–803, 2006.

[Soldatova *et al.*, 2008] L. Soldatova, W. Aubrey, R. King, and A Clare. The EXACT description of biomedical protocols. *Bioinformatics*, 24(13):295–303, 2008.

[Varma and Palsson, 1994] A. Varma and B. Palsson. Metabolic flux balancing: Basic concepts, scientific and practical use. *Nature Biotechnology*, 12:994–998, 1994.

[Waldrop, 2008] M. Waldrop. Science 2.0: Great new tool, or great risk? *Scientific American*, 2008.

[Whelan and King, 2008] K. Whelan and R. King. Using a logical model to predict the growth of yeast. *BMC Bioinformatics*, 9(97), 2008.