Automatic Revision of Metabolic Networks through Logical Analysis of Experimental Data

Oliver Ray[1], Ken Whelan[2], and Ross King[2]

[1] University of Bristol, Bristol, BS8 1UB, UK
oray@cs.bris.ac.uk
[2] University of Aberystwyth, Ceredigion, SY23 3DB, UK
{knw,rdk}@aber.ac.uk

Abstract. This paper presents a nonmonotonic ILP approach for the automatic revision of metabolic networks through the logical analysis of experimental data. The method extends previous work in two respects: by suggesting revisions that involve both the addition and removal of information; and by suggesting revisions that involve a combination of gene function, enzyme inhibition and metabolic reactions. Our proposal is based on a new declarative theory of metabolism expressed in a nonmonotonic logic programming formalism. With respect to this theory, a mixture of abductive and inductive inference is used to compute a set of minimal revisions needed to make a given network consistent with some observed data. In this way, we describe how a reasoning system called XHAIL was able to usefully revise a state-of-the-art metabolic network in order to better account for real-world experimental data acquired by an autonomous laboratory platform known as the Robot Scientist.

1 Introduction

Metabolic networks are formal descriptions of the enzyme-catalysed biochemical transformations that mediate the breakdown and synthesis of molecules within a living cell. Logic programs are useful for representing and reasoning about such networks as they provide an expressive relational language with efficient computational support for deduction, abduction and induction. Moreover, the recent development of nonmonotonic learning systems such as XHAIL (eXtended Hybrid Abductive Inductive Learning) [5] means that the full potential of logic programs with both classical and default negation can be now exploited for the representation and inference of defaults and exceptions under uncertainty.

This paper introduces a logical theory of metabolism which can be used to compute a set of minimal revisions needed to make a given network consistent with observed data. Using this theory, we describe how XHAIL correctly revised a state-of-the-art metabolic model to better account for real-world data acquired by an autonomous laboratory platform known as the Robot Scientist [1]. The method improves upon previous work by suggesting revisions that involve both the addition and removal of information; and by suggesting revisions that involve a combination of gene function, enzyme inhibition and metabolic reactions.

Presented at the 19th International Conference on Inductive Logic Programming (ILP'09).

2 Background

Metabolic networks are collections of interconnected biochemical reactions that mediate the synthesis and breakdown of essential compounds within a cell. These reactions are catalysed by specific enzymes whose amino acid sequences are specified in regions of the host genome called Open Reading Frames (ORFs). The activity of particular pathways within a network are controlled by regulating the expression of those genes on which the associated ORFs are located. One such pathway is exemplified in Figure 1, below, which shows the Aromatic Amino Acid (AAA) biosynthesis pathway of the yeast *S. cerevisiae*.

Nodes in the graph below are metabolites involved in the transformation of the start compound Glycerate-2-phosphate into the amino acids Tyrosine, Phenylalanine, and Tryptophan. Arrows are chemical reactions from substrates to products. 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 ORFs (in green). The single dashed line shows the inhibition of an enzyme (YBR249C) by a metabolite (C00082). The double dashed line represents the cellular membrane, which separates the cell cytosol from the growth medium.

ORFs shown above each other are iso-enzymes catalysing the same reaction, while ORFs next to each other are enzyme-complexes. All reactions take place in the cytosol using nutrients imported from the medium; and they proceed at a standard rate (within 1 day), except for the importation of two italicised compounds (C01179 & C00166), which take longer (between 1 and 2 days).

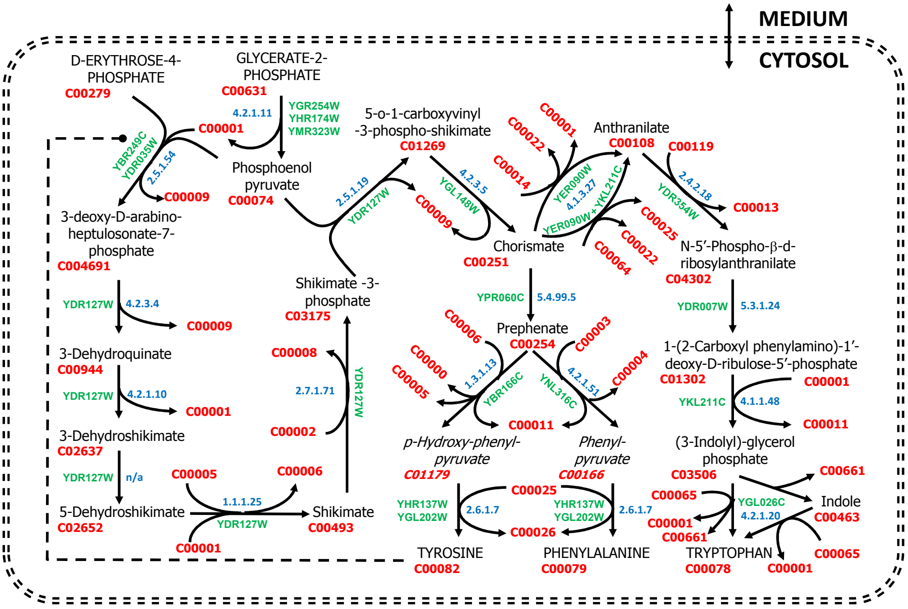

Fig. 1. Aromatic Amino Acid (AAA) biosynthesis pathway of the yeast *S. cerevisiae*

Over time, metabolic models must be revised as discrepancies emerge between predicted and observed results. This is usually done by hand, as for the AAA pathway, which was derived from the KEGG database, but was manually tuned to better explain the results of initial semi-automated growth experiments where different strains of yeast (from which certain ORFs are knocked out) are cultured in various growth media (to which certain nutrients are added) [2].

Now, our goal is to mechanise this revision process by applying XHAIL [5] to data obtained by a fully-autonomous improved Robot Scientist [1].

In brief, XHAIL is a nonmonotonic ILP system that takes a background theory B and a set of examples E, to return a set of hypotheses H that entail E with respect to B. The hypothesis space is controlled by a set M of mode declarations [3] that allow the user to constrain which literals may appear in the heads and bodies of hypothesis clauses. A compression heuristic [3] then selects between competing hypotheses by preferring solutions with fewer literals.

3 Approach

Our metabolic theory has the following basic types: ORFs and metabolites, which are denoted by their KEGG identifiers; enzyme-complexes and reactions, which are given unique integer identifiers; days and experiments, which are also represented by integers; and extra-cellular or intra-cellular compartments, of which we only consider the growth medium and cell cytosol.

The additional nutrients and knockout strains in each growth experiment are represented by ground atoms of the form `additional_nutrient(e,m)` and `knockout(e,o)`, for some particular experiment `e`, ORF `o`, and metabolite `m`. In addition, a minimal set of growth nutrients common to all experiments are represented by ground atoms of the form `start_compound(m)`.

By definition, any metabolite `Met` that is a start compound or additional nutrient is in the compartment `medium` on any `Day` in any experiment `Exp`:

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

Each enzyme-complex is given an integer identifier `c`. Then, for each reaction catalysed by `c`, one fact is added to the model of the form `catalyst(r,c)`, where `r` is the corresponding reaction identifier. Also, for each ORF `o` needed in the complex `c`, one fact is added to the model of the form `component(o,c)`.

Any enzyme inhibition in the AAA pathway, is represented by a ground atom of the form `inhibitor(c,m)`. Metabolites that are essential to cell growth, like the three amino acids, are specified as such by ground atoms of the form `essential_compound(m)`.

Cell development is arrested if an essential metabolite is not in the cytosol but growth is predicted otherwise. An enzyme-complex is deleted if a component ORF 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 biological practice as it ensures all newly introduced reactions are at least 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 ORFs 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).
```

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).

Positive and negative examples, which correspond to results about the growth and non-growth of the yeast in an experiment `e` on a day `d`, are denoted by ground literals of the form `observed_growth(e,d)` or `¬observed_growth(e,d)` where, purely for convenience, we use classical negation.

Previous work [6] describes how proof-of-principle tests on artificial data were used to demonstrate XHAIL's ability to revise a model by adding and removing information. In the next section, we describe the result of using real data from 40 experiments conducted by the Robot Scientist to revise the state-of-the-art AAA model in Figure 1.

4 Results

Upon submitting the observed growth results to XHAIL, they were immediately found to be inconsistent with the AAA model, thereby suggesting that a revision was necessary. After experimenting with the language bias for about one hour, we obtained around half a dozen hypotheses that achieved logical consistency between predicted and observed growth. On closer examination, these hypotheses turned out to be different combinations of four basic conjectures relating to the metabolites Anthranilate and Indole:

a. Anthranilate Import: There was 1 abductive conjecture `inhibited(Exp,25,1)` stating that the import of Anthranilate (which is mediated by a hypothetical enzyme with id 25) is blocked on day 1 in all experiments. This can be understood as meaning that the import of Anthranilate is a slow reaction analogous to the import of C01179 and C00166. Following a more detailed study of the raw growth data, we believe is indeed the case and have updated our model accordingly:

b. Enzyme Complex: There were 3 alternative abductive conjectures of the form `component("YER090W",7)`, `component("YER090W",8)` or `component("YER090W",9)` stating that YER090W is needed as part of a complex in any of 3 reactions, 2.4.2.18, 5.3.1.24 or 4.1.1.48 (which are catalysed by complexes 7,8 and 9, respectively) immediately following the Anthranilate Synthase step (4.1.3.27) in the Tryptophan pathway. This is plausible as reaction 4.1.1.48 is catalysed by YKL211C, which is already known to form a complex with YER090W in reaction 4.1.3.27. But an extensive literature survey, encompassing recent genome-wide protein interaction studies, has so far proved inconclusive.

c. Indole Contamination: In addition, there was 1 inductive conjecture of the form `predicted_growth(Exp,Day):-additional_nutrient(Exp,"C00463")` which can be understood as stating that the use of Indole as an additional nutrient leads to biased growth readings. This is plausible as Indole has a distinctive yellow colour much darker than the other nutrients. We wondered if this property might be confusing the optically measured growth readings. But, a more detailed analysis of the raw growth curves observed on Indole enriched media do not support this conclusion.

d. Indole Contamination: Finally, there was 1 inductive conjecture of the form `additional_nutrient(Exp,"C00078" ):-additional_nutrient(Exp,"C00463")` which can be understood as stating that the Robot Scientist's 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 the two may be hard to separate. In any event, using Mass Spectrometry, we have verified that our Indole was indeed contaminated with Tryptophan; and we will obviously take steps to avoid this source of error in future experiments.

5 Related Work

Our approach builds upon earlier Robot Scientist work [2] which used Progol5 [4] to rediscover ORF-enzyme mappings removed from the AAA pathway . 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 nested list representation of reactions over a 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 model is restricted to learning additional ORF-enzyme mappings in single gene deletion experiments. By contrast, the XHAIL model employs a completely declarative representation, adapted from [7] 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 can simultaneously add or remove reactions, inhibitions and complexes.

6 Conclusions

This paper presented a logical method for the automatic revision of metabolic networks through abductive and inductive analysis of experimental data. First we showed how a nonmonotonic logic programming formalism can be used to declaratively model metabolic reactions with arbitrary reactants catalysed by multiple enzymes subject to inhibitory feedbacks. Then, we described how the XHAIL reasoning system was used to revise a state-of-the-art AAA pathway in the light of real-world data obtained by an autonomous Robot Scientist. The results have been tested biologically and have thereby led to improvements in both or metabolic model and our experimental setup.

References

1. 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.
2. 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.
3. S. Muggleton. Inverse Entailment and Progol. *New Gen. Comp.*, 13:245–286, 1995.
4. S. Muggleton and C. Bryant. Theory Completion Using Inverse Entailment. In *Proc. 10th Int. Conf. on ILP*, volume 1866 of *LNCS*, pages 130–146. Springer, 2000.
5. O. Ray. Nonmonotonic Abductive Inductive Learning. *Journal of App. Logic*, 2009.
6. 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.
7. K. Whelan and R. King. Using a logical model to predict the growth of yeast. *BMC Bioinformatics*, 9(97), 2008.