Towards a Logical Analysis of Biochemical Reactions

Patrick Doherty and Steve Kertes and Martin Magnusson and Andrzej Szalas

Linköpings universitet, Artificial Intelligence and Integrated Computer Systems (AIICS)
Department of Computer and Information Science, SE-587 23 Linköping, Sweden

1 Introduction

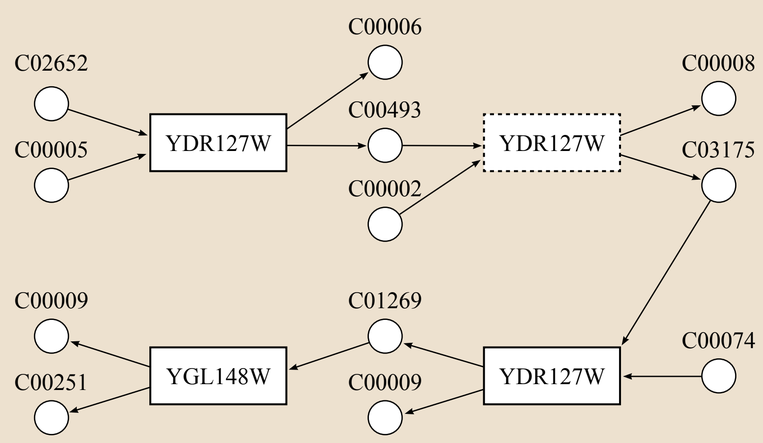
We provide a logical model of biochemical reactions and show how hypothesis generation using weakest sufficient (wsc) and strongest necessary conditions (snc) may be used to provide additional information in the context of an incomplete model of metabolic pathways.

2 Hypotheses Generation

Suppose one is given a (incomplete) specification of a set of interacting reactions. We would use this set of formulas as the background theory T. Suppose additionally, that a number of observations are made referring to reactions known to have occurred, or compounds known to be available for participation in a reaction, etc. Let α denote the formula representing these observations. Generally, it will not be the case that $T \models \alpha$ because T only provides an incomplete specification of the reactions.

We would like to generate a formula (candidate hypotheses) ϕ in a restricted subset of the language of reactions P such that ϕ together with the background theory T does entail the observations α. It is important that we do not over commit otherwise we could just as easily choose α itself as the hypothesis which wouldn't do much good. In fact, the $\text{WSC}(\alpha; T; P)$ does just the right thing since we know that $T \wedge \text{WSC}(\alpha; T; P) \models \alpha$ and it is the weakest such formula by definition.

The wsc's and snc's can be calculated efficiently for a large class of formulas by expressing them as second-order formulas, using an equivalence proved in [3], and obtaining logically equivalent first-order formulas by applying the DLS* algorithm described, e.g., in [2].



3 A Metabolic Pathway Example

Consider a fragment of the aromatic amino acid pathway of yeast shown in the figure (this is a fragment of a larger structure used in [1]).

In the graph there are two types of nodes: *compound nodes* (depicted by circles) and *reaction nodes* (depicted by rectangles). An edge from a compound node to a reaction node denotes a substrate. An edge from a reaction node to a compound node denotes a product of the reaction. We additionally allow conditions placed in the boxes and in this case rectangles are labelled with enzyme names, meaning that a respective enzyme is to be available for reaction.

The figure depicts the following reactions:

$$\mathrm{n}_1 : \mathrm{C02652} + \mathrm{C00005} \overset{\mathrm{YDR127W}}{\longrightarrow} \mathrm{C00006} + \mathrm{C00493}$$
$$\mathrm{n}_3 : \mathrm{C03175} + \mathrm{C00074} \overset{\mathrm{YDR127W}}{\longrightarrow} \mathrm{C01269} + \mathrm{C00009}$$
$$\mathrm{n}_4 : \mathrm{C01269} \overset{\mathrm{YGL148W}}{\longrightarrow} \mathrm{C00009} + \mathrm{C00251}.$$

4 Example Continued

It is assumed that reaction

$$\mathrm{n}_2 : \mathrm{C00493} + \mathrm{C00002} \overset{\mathrm{YDR127W}}{\longrightarrow} \mathrm{C03175} + \mathrm{C00008}$$

depicted by the dashed box is, in fact, missing.

The above set of reactions is expressed by formulas using the relation symbol $prec(R, R')$, meaning that reaction node R directly precedes reaction node R', and $av(C, R)$, meaning that compound C is available for reaction represented by reaction node R.

For example, the first reaction is expressed by:

$$\begin{aligned} &react(n_1, R) \rightarrow \\ &\quad av(ydr127w, R) \wedge \\ &\quad av(c02652, R) \wedge av(c00005, R) \wedge \\ &\quad \forall R'.\,[prec(R, R') \rightarrow av(c00006, R')] \wedge \\ &\quad \forall R'.\,[prec(R, R') \rightarrow av(c00493, R')]. \end{aligned}$$

The missing reaction is also present, among many other reactions, in the database, and is expressed by:

$$\begin{aligned} &react(n_2, R) \rightarrow \\ &\quad av(ydr127w, R) \wedge \\ &\quad av(c00493, R) \wedge av(c00002, R) \wedge \\ &\quad \forall R'.\,[prec(R, R') \rightarrow av(c03175, R')] \wedge \\ &\quad \forall R'.\,[prec(R, R') \rightarrow av(c00008, R')]. \end{aligned}$$

We assume that the underlying database contains partial information about the observed chain of reactions:

$$react(n_1, r_1) \wedge react(n_3, r_3) \wedge react(n_4, r_4)$$

together with a description of reactions n_1, n_3, n_4 and many other reactions, including n_2. Let the considered knowledge base be denoted by KDB.

We can now consider, e.g., $\text{WSC}(\alpha; \text{KDB}; av)$, where

$$\alpha \overset{\text{def}}{\equiv} \exists N.[react(N, r_2) \wedge prec(r_1, r_2) \wedge prec(r_2, r_3)],$$

providing one with the weakest requirement expressed in terms of av only, making α true, provided that the background theory given by KDB holds.

In our case, the generated hypotheses will contain the disjunct $av(c00002, r_2)$, reflecting sufficient conditions for reaction r_2 occuring under the $prec$ constraints. The $\text{SNC}(\alpha; \text{KDB}; \{out\})$ will contain the disjunct

$$out(N, c03175) \wedge out(N, c00008),$$

reflecting necessary conditions for the reaction and $prec$ constraints. If one of the compounds $c03175$, $c00008$ has not been observed during the reaction chain, one can reject the hypothesis that reaction N in node r_2 was n_2.

REFERENCES

[1] C.H. Bryant, S.H. Muggleton, S.G. Oliver, D.B. Kell, P. Reiser, and R.D. King, 'Combining inductive logic programming, active learning and robotics to discover the function of genes', *Linkoping Electronic Articles in Computer and Information Science*, **6**(12), (2001).

[2] P. Doherty, W. Łukaszewicz, and A. Szałas, 'Computing circumscription revisited', *Journal of Automated Reasoning*, **18**(3), 297–336, (1997).

[3] P. Doherty, W. Łukaszewicz, and A. Szałas, 'Computing strongest necessary and weakest sufficient conditions of first-order formulas', *International Joint Conference on AI (IJCAI'2001)*, 145 – 151, (2000).