Towards a Logical Analysis of Biochemical Reactions

Patrick Doherty and **Steve Kertes** and **Martin Magnusson** and **Andrzej Szalas**[1]

Abstract.

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

1 A Logical Model for the Analysis of Biochemical Reactions

The analysis of biochemical pathways has been considered in numerous papers (see, e.g., [1, 2]). In this paper, a bipartite graph representation of chemical reactions will be used (see, e.g., [2]).

It is assumed that any reaction is specified by:

$$n : \mathrm{c_1} + \ldots + \mathrm{c_k} \xrightarrow{\alpha(\mathrm{n})} \mathrm{c'_1} + \ldots + \mathrm{c'_l}, \tag{1}$$

where n is a label (name) of the reaction, $c_1, \ldots, c_k$ are reactants (inputs for n), $c'_1, \ldots, c'_l$ are products of n and $\alpha(n)$ is a formula that specifies additional conditions necessary for the reaction, such as temperature, pressure, presence of catalyzers, etc.

In this paper, the classical first-order logic is used for specifying reactions. Reaction nodes will be represented explicitly, while information about available compounds will be given via a suitable relation. Consequently, it is assumed that the following symbols are available:

1. constants and variables:
   - constants naming reactions (n, n'), compounds ($c, c', h2o, co2$ etc.), and reaction nodes (r, r')
   - variables representing reactions (N, N'), compounds (C, C'), and reaction nodes (R, R')

   for constants and variables we also use indices, when necessary
2. relation symbols reflecting static information:
   - $in(C, N)$ meaning that compound C is needed for reaction N
   - $out(N, C)$ meaning that compound C is a product of reaction N
3. relation symbols reflecting dynamic information:
   - $prec(R, R')$ meaning that reaction node R precedes reaction node R'
   - $chain(R, R')$ meaning that there is a chain of reactions $R, R_1, R_2, \ldots, R_k, R'$ such that $prec(R, R_1), prec(R_1, R_2), \ldots, prec(R_{k-1}, R_k), prec(R_k, R')$
   - $react(N, R)$ meaning that reaction N actually happened in reaction node R
   - $av(C, R)$ meaning that compound C is available for reaction represented by reaction node R.

[1] Department of Computer and Information Science, SE-581 83 Linköping, Sweden, email: {patdo,g-steke,marma,andsz}@ida.liu.se

Let e, t be any expressions and s any subexpression of e. By $e(s/t)$ we shall mean the expression obtained from e by substituting each occurrence of s by t.

It is assumed that any formula is implicitly universally quantified over all its variables that are not bound by a quantifier.

Any reaction of the form (1) is translated into the formula $react(n, R)$, where the following background theory is assumed:

- static information about reaction n:

  $$in(c_1, n) \wedge \ldots \wedge in(c_k, n) \wedge$$
  $$out(n, c'_1) \wedge \ldots \wedge out(n, c'_l)$$

- linking nodes in graphs with reactions

$$\begin{aligned} & react(n, R) \rightarrow \\ & \quad \alpha(n/R) \wedge \\ & \quad av(c_1, R) \wedge \ldots \wedge av(c_k, R) \wedge \\ & \quad \forall R'.\, \{prec(R, R') \rightarrow av(c'_1, R')\} \wedge \\ & \quad \ldots \wedge \\ & \quad \forall R'.\, \{prec(R, R') \rightarrow av(c'_l, R')\}. \end{aligned} \tag{2}$$

2 Hypotheses Generation Using Strongest Necessary and Weakest Sufficient Conditions

Snc's and wsc's (see e.g. [5] for a definition) provide a powerful means of generating hypotheses using abduction. Suppose one is given a (incomplete) specification of a set of interacting reactions of the form shown in equation (1). 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 $\mathrm{WSC}(\alpha; T; P)$ does just the right thing since we know that $T \wedge \mathrm{WSC}(\alpha; T; P) \models \alpha$ and it is the weakest such formula by definition.

$\mathrm{WSC}(\alpha; T; P)$ actually represents alternative hypotheses for explaining α. If it is put in disjunctive normal form, each of the disjuncts makes $\mathrm{WSC}(\alpha; T; P)$ true and represents a weakest hypothesis. To reason about what each candidate hypothesis might imply in terms of completing the reaction representation, one can add both the background theory T and the candidate hypothesis α' to a logic database and query the database as desired. For this purpose

the approximate databases described in [3] are a suitable alternative that enables reasoning about incomplete and approximate knowledge through the use of rough set techniques.

Observe that:

- wsc corresponds to a weakest abduction expressed in a given target language. For example, consider

  $\textsc{Wsc}(av(fin, so3); Th; L)$,

  where Th is the theory expressing properties of given reactions, as constructed in Section 1. Then
  - if the target language L consists of av only, the resulting wsc expresses what compound's availability makes the required output of reaction node fin feasible
  - if the target language consists of $react$, then the resulting wsc expresses reactions necessary to make the required output of fin feasible
- snc allows one to infer facts from negative information. In fact, snc expresses what would be possible under a given set of hypotheses. For example, if a certain side product has not been observed, a reaction can be excluded from the set of hypotheses.

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 [5], and obtaining logically equivalent first-order formulas by applying the $\textsc{Dls}^*$ algorithm described, e.g., in [4].

3 A Metabolic Pathway Example

Consider a fragment of the aromatic amino acid pathway of yeast, shown in Figure 1 (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, i.e., that av holds. For example, $av(ydr127w, r)$ is necessary, when the label of r is "YDR127W".

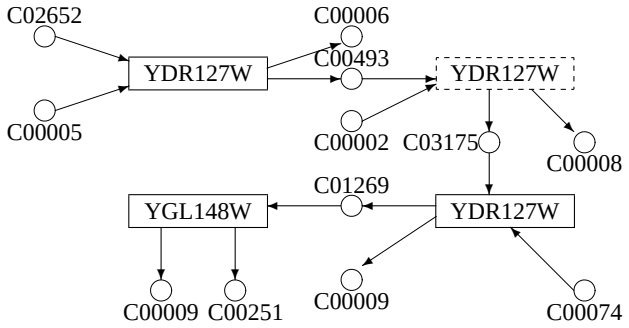

Figure 1. A fragment of the aromatic amino acid pathway of yeast.

Figure 1 depicts the following reactions:

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

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 as defined in Section 1. 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:

$$\begin{aligned} &react(n_1, R) \rightarrow \\ &\quad av(ydr127w, R) \wedge \\ &\quad av(c00493, R) \wedge av(c00002, R) \wedge \\ &\quad \forall R'.\,[prec(R, R') \rightarrow av(c03175, R')] \wedge \\ &\quad \forall R'.\,[prec(R, R') \rightarrow av(c00008, R')]. \end{aligned}$$

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_3. Let the considered knowledge base be denoted by $\textsc{Kdb}$.

We can now consider, e.g., $\textsc{Wsc}(\alpha; \textsc{Kdb}; av)$, where

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

providing one with the weakest requirement expressed in terms of av only, making α true, provided that the background theory given by $\textsc{Kdb}$ holds.

In our case, the generated hypotheses will contain the disjunct $av_+(c00002, r_2)$, reflecting, among others, sufficient conditions for $prec$. The $\textsc{Snc}(\alpha; \textsc{Kdb}; \{out\})$ will contain the disjunct

$$out_+(N, c03175) \wedge out_+(N, c00008),$$

reflecting, among others, necessary conditions for $prec$. 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] Y. Deville, D. Gilbert, J. van Helden, and S. Wodak, 'An overview of data models for the analysis of biochemical pathways', *Briefings in Bioinformatics*, **4**(3), 246–259, (2003).

[3] P. Doherty, J. Kachniarz, and A. Szałas, 'Using contextually closed queries for local closed-world reasoning in rough knowledge databases', in *Rough-Neuro Computing: Techniques for Computing with Words*, eds., S.K. Pal, L. Polkowski, and A. Skowron, Cognitive Technologies, pp. 219–250. Springer–Verlag, (2003).

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

[5] 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).