KINETIC MODELS AND QUALITATIVE ABSTRACTION FOR RELATIONAL LEARNING IN SYSTEMS BIOLOGY

Gabriel Synnaeve
E-Motion Team at INRIA, Grenoble, France
gabriel.synnaeve@gmail.com

Katsumi Inoue
National Institute of Informatics, Tokyo, Japan
ki@nii.ac.jp

Andrei Doncescu
LAAS-CNRS 31007, Toulouse, France
andrei.doncescu@laas.fr

Hidetomo Nabeshima
University of Yamanashi, Japan
nabesima@yamanashi.ac.jp

Yoshitaka Kameya, Masakazu Ishihata, Taisuke Sato
Tokyo Institute of Technology, Tokyo, Japan
{*kameya,ishihata,sato*}*@mi.cs.titech.ac.jp*

Keywords: systems biology, discretization, metabolic pathways, inductive logic programming, abduction

Abstract: This paper presents a method for enabling the relational learning or inductive logic programming (ILP) framework to deal with quantitative information from experimental data in systems biology. The study of systems biology through ILP aims at improving the understanding of the physiological state of the cell and the interpretation of the interactions between metabolites and signaling networks. A logical model of the glycolysis and pentose phosphate pathways of *E. Coli* is proposed to support our method description. We explain our original approach to building a symbolic model applied to kinetics based on Michaelis-Menten equation, starting with the discretization of the changes in concentration of some of the metabolites over time into relevant levels. We can then use them in our ILP-based model. Logical formulae on concentrations of some metabolites, which could not be measured during the dynamic state, are produced through logical abduction. Finally, as this results in a large number of hypotheses, they are ranked with an expectation maximization algorithm working on binary decision diagrams.

INTRODUCTION

Nowadays, systems biology represents the key field to explain the functionality of life science. To analyze a biological system it is necessary to find out new mathematical models allowing to explain the evolution of the system in a dynamic context or to deal in a simple manner with the complex situations where the human experience overtakes mathematical reasoning (Kitano, 2002). Many physical and biological phenomena may be represented on an analytical form using dynamical system. Our case study is based on wet biology experiment consisting in applying a pulse of glucose in a small bio-reactor containing *E.Coli* that led to building an ordinary differential equations (ODEs) based simulator. We used high performance liquid chromatography to measure some metabolites concentrations and some others had to be estimated, using a simulated annealing algorithm, since no experimental results were available. So, knowing completely the evolutions of metabolites concentrations of this system, we applied our approach to show its correctness. For that, we took only steady-state values of metabolites concentrations and ran our model.

Several attemps have been done for logic-based approaches to analyze biochemical pathways in Systems Biology. They use action languages (Baral et al., 2004), abduction (Juvan et al., 2005; King et al., 2004; King et al., 2005; Tamaddoni-Nezhad et al., 2006), SAT (Tiwari et al., 2007), inductive logic programming (Doncescu et al., 2007) or answer set programming (Dworschak et al., 2008). All these previous approaches are based on qualitative modeling, and none of them can handle continuous domains appropriately. Temporal logic combined with the representation of kinetic models in stochastic logic programming (SLP) (Fages et al., 2008) have a similar goal using different means: the authors modeled the kinetics of biochemical systems by continuous time Markov chains as input to SLP where we took an approach to discretize (through continuous HMM) concentrations of metabolites first and then use them combined with a logical translation of ODEs-based kinetics as input to ILP. The goal of this research is to incorporate continuous values and kinetics within the logic-based approach to metabolic pathways. In

particular, we enhance an abductive framework proposed in (Inoue et al., 2009), which consists of abductive hypothesis generation and statistical hypothesis evaluation, by enabling us to handle real-valued data obtained from measurement in observations.

For that, we now propose a *loop for learning about a metabolic pathway from experiments* in which we have to (each step corresponds to a section, as in Fig. 1):

1. clusterize continuous concentrations of metabolites over time into discrete levels and discrete timesteps.
2. use them in an ILP-based model of the pathway, in conjunction with a set of knowledge-generating rules, here in the example describing Michaelis-Menten kinetics.
3. sort the resulting abduced facts or inducted rules with our defined metrics.
4. use this ranking for enhancing our knowledge base and *goto* the beginning of this process.

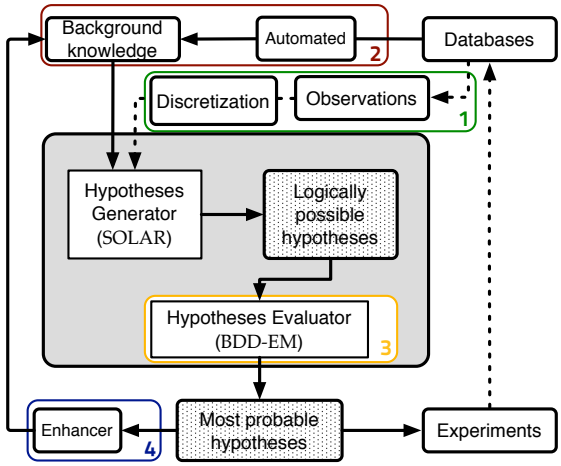

Figure 1: Overview of the complete process

In this paper, we show how this "closed loop" architecture can be applied to an inverse problem: given the measured concentrations of some metabolites in a steady state, we compute the concentrations of metabolites before the dynamic transition to this steady state based on the kinetic modeling. We worked with the beginning of an automated framework (see Fig. 2 for a practical data-centric circuit) to deal with different real world pathways and experiments. It is mainly composed of four tools:

- The combination of an implementation of continuous HMMs (Gauvain and Lee, 1994; Ji et al., 2006) with PY-TSDISC to discretize experimental values.
- KEGG2SYMB, using the KEGG API, that transform pathways from KEGG (Kanehisa and Goto, 2000; Kanehisa et al., 2008) into symbolic models.
- SOLAR, a consequence finding system working on Skipping Ordered Linear tableaux (Nabeshima et al., 2003), which is *complete* for finding minimal explanations, to conduct abduction or induction.
- BDD-EM, an implementation of the expectation-maximization algorithm on binary decision diagrams (Ishihata et al., 2008; Inoue et al., 2009) to rank hypotheses.

We chose to illustrate this method on the conjunction of glycolysis and pentose phosphate pathways for *E.Coli*, simplified the model by keeping 16 relevant reactions and discretized experimental values (16 values) as in section 1. We added the three Michaelis-Menten based rules and the three constraints of unicity for the levels as in section 2. We had 15 unknown levels of concentrations of metabolites before the transition to the steady state (yielding 15×3 levels $= 45$ abducibles). SOLAR, used for abduction, outputs 98 hypotheses that cover all these metabolites. With such a number, picking the right hypotheses should be done in an automated way as we did in section 3.

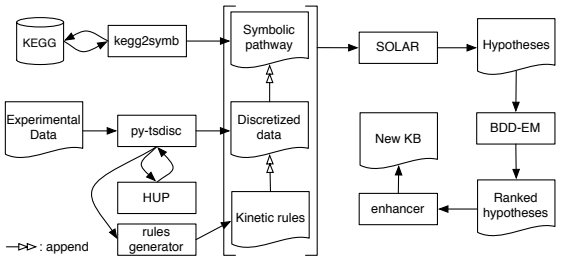

Figure 2: Data-centric schema of the process

1 DISCRETIZATION OF TIME SERIES FROM EXPERIMENTS

In our modeling, we first introduce discrete concentration levels to filter what are the relevant changes of concentration of the metabolites, in regard to hypotheses generation from ILP. We need to be able to infer hypotheses that have a certain level of generality and, for that, we should use intervals instead of single real values. This could have been done with an interval constraints approach (Benhamou, 1994), but we currently choose a discretization approach. Although this gives us less freedom in the logic part as levels

are fixed (as if we have fixed intervals), levels can be handled just as symbols in a logical model of pathways.

Discretizing time series is a research field in which many works (Geurts, 2001; Keogh et al., 2005) have been conducted recently. Our practical problem is that we want to have a statistically relevant (unsupervised) discretization for N metabolites concentrations over time. We also discretize the values of K_m (Michaelis-Menten constants, see (1)), for each reaction, with the same levels. For that purpose, we use a probabilistic model, used in speech recognition and time series analysis: continuous hidden Markov model (HMMs) (Rabiner, 1989). We can therefore compute an appropriate number of levels (that was three for *E.Coli*) in regard to a Bayesian score such as Bayesian Information Criterion (BIC) (Schwarz, 1978) or as the Cheeseman-Stutz score (Cheeseman and Stutz, 1995) or as the variational free energy. This process can be achieved through the following methods all described in (Beal, 2003), respectively: maximum likelihood estimation or maximum a posteriori estimation or through a variational Bayesian method.

We use continuous (Gaussian) HMMs with parameter tying[1]. This is a solution to the problem of sharing the same symbolic levels in all the logic models in order to be able to assign the level of a compound to another and be dealing with the same real values behind the scene. We first prepare N continuous HMMs (one for each metabolite), where each state variable takes a concentration level, and each output variable takes a measurement of concentration and follows a univariate Gaussian distribution. All the HMMs share a state space as well as the parameters in the output variables (i.e. means and variances), so that they produce discrete levels that are corresponding. These relevant discretized levels of concentration are computed through the expectation-maximisation (EM) algorithm with maximum a posteriori (MAP) estimation (Gauvain and Lee, 1994) or through the variational Bayes EM (VB-EM) (Beal, 2003; Ji et al., 2006). We prefer this last method as it is shown (Beal, 2003) that variational free energy provides a more accurate approximation of the marginal log-likelihood than BIC or the Cheeseman-Stutz score.

Then, we use a simple round-mean aggregation of them for time-sampling. We set a maximal number of time steps and look for the better fitting width and alignment for equal-width time intervals. We are currently developping a different process in the direction of discretization of our time series from molecular biology experiments that will discretize time and levels simultaneously but current results are already useable (see Table 1 and Fig. 5) and that is what we based the work presented here on.

[1]Parameter tying is a notion often used in HMMs for speech recognition (Rabiner, 1989) and recently in statistical relational learning (De Raedt, 2008). In our case, the mean and the variance for $X_t^{(n)}$, the output variable at time t in the HMM for the n-th metabolite ($n = 1,\ldots,N$), are tied with the mean and the variance for $X_{t'}^{(n')}$, respectively ($n \neq n'$ and $t \neq t'$).

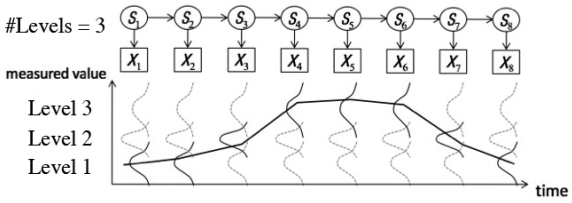

Figure 3: 3-state continuous HMM discretizing one experimental time series, where X_t is the measurement of concentration at time t and S_t is the hidden state that indicates the corresponding discretized level.

2 MODELING OF THE PATHWAYS OF *E.COLI*

To obtain an understanding of the central metabolism, a logical model has been developed according to a kinetic model including the glycolysis and the pentose phosphate pathway for *Escherichia coli* (Chassagnole et al., 2006). The Fig.4 shows the simplified pathway that we modelized logically with relations `reaction(Substrate, Enzyme, Product, Km)`.

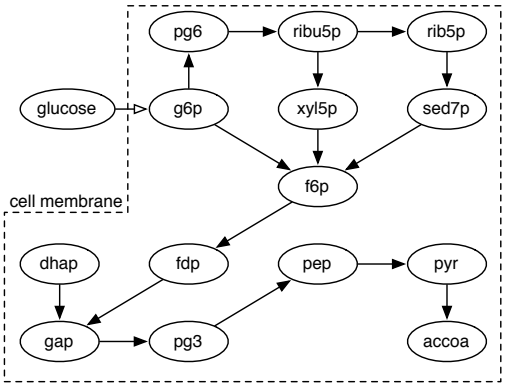

Figure 4: Simplified glycolysis and pentose phosphate pathways for *E.Coli*

The metabolic networks dynamics are in their enzymatic part ruled by the combination of classical kinetics: essentially Michaelis-Menten, Hill and allosteric ones. If we limit our modeling to these kinetics, we can highly simplify their mathematical han-

dling, and that is what we did. We chose to use only Michaelis-Menten kinetics, because we had a pathway simple enough and that it is the more general representation for a non-linear allosteric regulation system. It assumes that the two enzyme binding equilibria are fast when compared to the interconversion of enzyme + substrate (ES) and enzyme + product (EP) compounds. That assumption appears reasonable considering that the dynamics of the experiment were happening in less that a minute: this implies that the effects of genetic regulation of the enzymes included are negligible and so the maximum reaction rates represent the amount and catalytic activity of enzymes.

$$E + S \rightleftharpoons_{k_{-1}}^{k_1} ES \rightarrow^{k_2} E + P$$

$$Michaelis-Menten\ eq.: \frac{d[P]}{dt} = V_m \frac{[S]}{[S] + K_m} \quad (1)$$

If both the substrate (S) and the product (P) are present, neither can saturate the enzyme. For any given concentration of S the fraction of S bound to the enzyme is reduced by increasing the concentration of P and *vice versa*. For any concentration of P, the fraction of P bound to the enzyme is reduced by increasing concentration of S. When we have $S \leftrightarrows P$, we just have to consider reactions for both directions. We consider a time discretization of the chemical rate equation for a reation between a substrate and a product with respective stoechiometric coefficient s and p:

$$s.S \rightarrow p.P : rate = \frac{1}{p} \times \frac{d[P]}{dt} \longrightarrow_{disc.time} \frac{1}{p} \times \frac{\Delta[P]}{\Delta T} \quad (2)$$

$$(1)\ and\ (2) \implies p \times rate = V_m \frac{[S]_T}{[S]_T + K_m}$$

$$\approx \frac{[P]_{T+timestep} - [P]_T}{(T + timestep) - T}$$

We chose to work with a constant timestep :

$$\implies [P]_{T+1} = V_m \frac{[S]_T}{[S]_T + K_m} + [P]_T \quad (3)$$

We can note that the Michaelis-Menten constants (`Km`) are homogenous to a concentration. We can then state `conc(Km, Level, Time)` in our modeling to set them, where `conc` stands for concentration. The experimental response observations of intracellular metabolites to a pulse of glucose were measured in continuous culture employing automatic stopped flow and manual fast sampling techniques in the time-span of seconds and milliseconds after the stimulus with glucose. The extracellular glucose, the intracellular metabolites: glucose-6-phosphate (g6p), fructose-6-phosphate (f6p), fructose1-6bisphosphate (fdp), glyceraldehyde3phosphate (gap), phospho-enolpyruvate (pep), pyruvate (pyr), 6phosphate-gluconate (6pg), glucose-1-phosphate (g1p) as well as the cometabolites: atp, adp, amp, nad, nadh, nadp, nadph were measured using enzymatic methods or high performance liquid chromatography. All the steady-state concentrations *measurements* of the *E.Coli* experiment and their corresponding discrete levels are summarized in Table 1.

#	Metab.	Conc.	Lvl	#	Metab.	Conc.	Lvl
1	glucose	0.055	0	2	g6p	3.480	2
3	f6p	0.600	0	4	fdp	0.272	0
5	gap	0.218	0	6	pep	2.670	2
7	pyr	2.670	2	8	6pg	0.808	1
9	g1p	0.653	0	10	amp	0.955	1
11	adp	0.595	0	12	atp	4.270	2
13	nadp	0.195	0	14	nadph	0.062	0
15	nad	1.470	1	16	nadh	0.100	0

Table 1. Concentrations (mM/L) of the Metabolites and their discretized levels for steady states

Inductive Logic Programming, used for induction or abduction (Mooney, 1997), allows to deal with discrete levels (symbols) and qualitative rules (Doncescu et al., 2007). Given the background knowledge B and an observation E (example), the task of ILP is to find an hypothesis H such that:

- $B \wedge H \models E$ and
- $B \wedge H$ is consistent

Inverse entailment (Inoue, 1992; Muggleton, 1995; Inoue, 2004) enables us to compute H through deduction by using:

- $B \wedge \neg E \models \neg H$ and
- $B \not\models \neg H$

We are here interested in abducing what happens during the dynamical transition based on observations from Table 1. Inverse entailment for abduction is studied in (Inoue, 1992) in which abductive computation can be realized by the consequence finding procedure SOL. In this case, both E and H are sets of literals, so both $\neg E$ and $\neg H$ are clauses. This approach can be further extended for inducing general hypotheses in (Inoue, 2004), which is generalized from (Muggleton, 1995), to allow B, E and H for full clausal theories.

SOLAR can be used as an abductive procedure to infer a hypothesis H in the form of a set of literals. Our logical model is based on the simplified Michaelis-Menten equation **(3)** which has here been represented by three background clauses using the `conc(Compound, Level, Time)` predicate. If we make the approximations for extreme values in:

$$[P]_{T+1} = V_m \frac{[S]_T}{[S]_T + K_m} + [P]_T \quad (3)$$

With only 3 levels, as we have in our discretization of *E.Coli* experiments, we will get the following simple rules:

- $\underline{[S] \ll K_m} \Rightarrow \frac{\Delta[P]}{\Delta T} = \frac{V_m}{K_M} \Rightarrow [P]_{T+1} = [P]_T$ `reaction(S, P, Km)` $\wedge$ `conc(S, 0, T)` $\wedge$ `conc(Km, 2, T)` $\wedge$ `conc(P, L, T)` $\rightarrow$ `conc(P, L, T+1)`

  The concentration of the product will not change between T and T+1 if the reaction is *very* slow.

- $\underline{[S] \simeq K_m} \Rightarrow \frac{\Delta[P]}{\Delta T} = \frac{V_m}{2} \Rightarrow [P]_{T+1} = V_m/2 + [P]_T$ `reaction(S, P, Km)` $\wedge$ `conc(S, L, T)` $\wedge$ `conc(Km, L, T)` $\wedge$ `conc(P, L2, T)` $\rightarrow$ `conc(P, L2, T+1)`

  The concentration change of the product between T and T+1 is not big enough to switch from one level to another. This is an approximation and a handy consequence of our discretization (using a log-scale on real values).

- $\underline{[S] \gg K_m} \Rightarrow \frac{\Delta[P]}{\Delta T} = V_m \Rightarrow [P]_{T+1} = V_m + [P]_T$ `reaction(S, P, Km)` $\wedge$ `conc(S, 2, T)` $\wedge$ `conc(Km, 0, T)` $\wedge$ `conc(P, L, T)` $\rightarrow$ `conc(P, 2, T+1)`

  If the reaction is very quick, it will result in transforming all the substrate into product in one time step.

If we had more than three levels, we would either need more rules (they can be automatically generated) or a general procedure for handling our kinetic model. This last one is a current implementation issue related to SOLAR. Another way to deal with more levels being currently explored consist in the automated generation of kinetics rules w.r.t. the discretization. Furthermore, we made some simplifications in the pathways to be able to use only Michaelis-Menten kinetics, another research topic is to extend our modeling to reactions ruled by other types of kinetics.

We also added constraints about the unicity of levels at a given time to reduce the number of hypotheses while keeping consistency:

- $\neg$`conc(S, 0, T)` $\vee$ $\neg$`conc(S, 1, T)`
- $\neg$`conc(S, 0, T)` $\vee$ $\neg$`conc(S, 2, T)`
- $\neg$`conc(S, 1, T)` $\vee$ $\neg$`conc(S, 2, T)`

Now we set the observations for the 6 metabolites (#2 - #7) from Table 1, which have been possibly affected by the stimulus with glucose, and the abducibles as those literals of the form `conc(_,_,0)`. Using SOLAR, we get 98 hypotheses as: `H76 = conc(g6p,2,0)` $\wedge$ `conc(adp,2,0)` $\wedge$ `conc(fdp,0,0)` $\wedge$ `conc(dhap,0,0)` $\wedge$ `conc(gap,0,0)` $\wedge$ `conc(glucose,2,0)` $\wedge$ `conc(pg3,2,0)` $\wedge$ `conc(pep,2,0)` $\wedge$ `conc(atp,0,0)` $\wedge$ `conc(pyr,2,0)`

3 RANKING HYPOTHESES

(Ishihata et al., 2008) (Ishihata et al., 2008) proposed the BDD-EM algorithm that is an implementation of the expectation maximization algorithm working on binary decision diagram, allowing it to deal with boolean functions. (Inoue et al., 2009) (Inoue et al., 2009) have applied the BDD-EM algorithm to rank hypotheses obtained through abduction. To rank our $H_1, \ldots, H_n$ hypotheses by probability, we consider the finite set of ground atoms $\mathcal{A}$ that contains all the values that can take our `conc(Compound, Level, Time)` and `reaction(Substrate, Product, Km)`. Each of the elements of $\mathcal{A}$ is a boolean variable. One of its subsets is the subset of abducibles Γ composed of all the possible values of `conc(Compounds, Level, 0)`. With $\theta_i = P(A_i)$ *for* $A_i \in \mathcal{A}$, we have to maximize the probability of the disjunction of hypotheses helped with the background knowledge B: $F = (H_1 \vee \cdots \vee H_n) \wedge ground(B)$ to set the good θ parameters (by the BDD-EM algorithm). F can still be too big to be retained as a BDD, so an optimisation F' of its size is obtained through the use of the minimal proofs for B and each H_i. Then, the BDD-EM algorithm computes the probabilities of ground atoms in $\mathcal{A}$ that maximizes the probability of F'. Finally, the probabilities of each hypotheses used for the ranking are computed as the products of the probabilities of literals appearing in each H_i.

To sort our 98 abduced hypotheses, we ran the EM algorithm on the BDDs corresponding to our hypotheses 10,000 times with random initializations. Note that if the comparison of these probabilities with each other is relevant, they should not be taken as absolute probabilities. The 10 most probable abduced hypotheses are the following:

Hyp. #	Probability	Abduced conc. levels at T=0
H76	≈ 1.000	g6p: 2, adp: 2, f6p: 0, fdp: 0, dhap: 0, gap: 0, glucose: 2, pg3: 2, pep: 2, atp: 0, pyr: 2
H41	0.822	the same as H76 except pg3: 0
H56	0.625	the same as H76 except g6p: 0
H70	0.553	the same as H76 except atp: 2
H13	0.515	the same as H56 except adp: 0
H90	0.455	the same as H70 except pg3: 0
H82	0.442	the same as H76 except dhap: 2
H43	0.369	the same as H76 except pyr: 1
H9	0.364	the same as H41 except dhap: 2
H68	0.346	g6p: 0, adp: 0, f6p: 0, fdp: 0, dhap: 0, gap: 0, glucose: 2, pg3: 2, pep: 2, atp: 2

Table 2. 10 most probable hypotheses

These hypotheses are corresponding to our biological knowledge that pyruvate is a bottleneck (Peters-Wendisch et al., 2001) and that the glucose that is to-

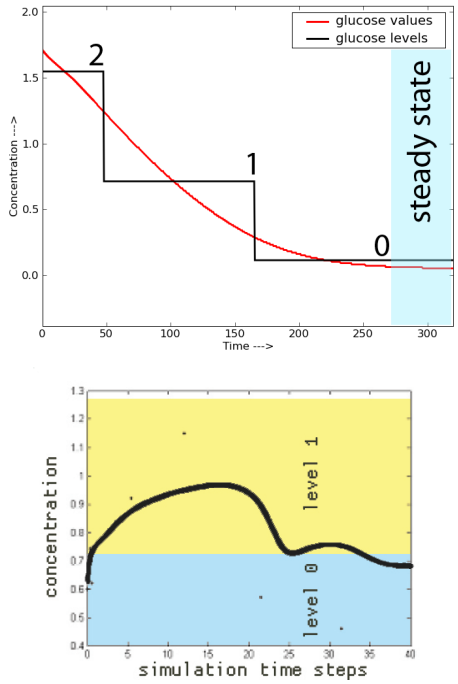

Figure 5: Top: Discretization of the concentration of glucose in the Glycolysis Pathway of *E.Coli* after an initial pulse. Bottom: Simulated evolution of the concentration of fructose-6-phosphate during the whole experiment.

tally consumed (e.g. top plot of Fig. 5 from simulation) was in high concentration at the beginning of the experiment (pulse). It goes along with the very general reaction of glycolysis: $glucose + 2\text{ADP} + 2\text{P} + 2\text{NAD}^+ \rightarrow 2\ pyruvate + 2\text{ATP} + 2(\text{NADH},\text{H}^+) + 2\text{H}_2\text{O}$. Also, for some metabolites, such as fructose-6-phosphate, the levels found through abduction are corresponding to the output of the simulation (e.g bottom plot of Fig.5) with the same low level (0) before and after the dynamic transition.

4 ENHANCING THE KNOWLEDGE BASE

Increasing our knowledge about a system is considered as an iterative process: at first, we consider the background knowledge combined with the observations as our knowledge base. Then we produce hypotheses and we need to use an algorithm to enhance (update) our knowledge base with some of the discovered hypotheses, here: abducibles. Ideally, we would re-run the hypothesis finding process until we cannot find anything new. This is particularly important when working with complex chained reactions and multiple time steps as it can enable deeper learning. This idea of revising the knowledge base is already found in (Ray et al., 2009) with a nonmonotonic approach, but their revision method stays in a qualitative modeling and do not take quantitative aspects into account.

Here, it is needed to pick hypotheses that are consistent with the background knowledge and with each others. For example, if we apply a greedy algorithm (as **Algorithm 1**) that picks hypothesis in decreasing probability order such that the hypothesis add some knowledge and that our enhanced knowledge stays consistent, it prevents from abducing other discoverables than the ones contained in H76. For instance we cannot find concentrations at T=0 for ribu5p, rib5p, sed7p, xyl5p, because if they were abduced, the resulting hypotheses would become inconsistent with H76. Note also that the abducibles added into the knowledge base may reduce the computational cost of later iterations of abduction/induction, but it is comparable to discard some branches of exploration.

Algorithm 1 An algorithm to enhance the knowledge base: most probables firsts

knowledge ← knowledge_base
sorted_hypotheses ← *sort*(hypotheses)
while *length*(discoverable) > 0 &&
length(sorted_hypotheses) > 0 **do**
  tmp ← sorted_hypotheses.*pop*()
  if *contains*(tmp, discoverable) && *consistent*(tmp, knowledge) **then**
    knowledge.*enhance*(tmp)
    discoverable.*remove*(tmp)
  end if
end while

With the explicit functions *length*, *pop* (destructive), and:

- *sort* sorts the hypotheses by decreasing probability.
- *contains* is a function that returns statements of first argument contained in the second.
- *consistent* performs consistency checking of two theories and return True if they are consistent.
- *enhance* adds statements that are not yet present in the considered ("self", "this") knowledge.
- *remove* deletes statements from argument present in the considered ("self", "this") object (could make use of *contains*).

We could have chosen to pick a combination of hypotheses that discovers more abducibles by penalizing the solutions including too few different abducibles with a scoring function inspired by the BIC (Schwarz, 1978): $score = -2\ln(error) + \lambda \cdot f(k,n)$ with k being the number of chosen hypotheses, n the number of abducibles, f a function that indicates the structural complexity of the combination of hypotheses (decreasing with the increase of n and increasing with the increase of k) and *error* the product of the

probabilities of chosen hypotheses. We assume here that we can use their relative significations in *error* by unbiasing the score with a λ parameter. So that the goal of such an algorithm would be to discover all abducibles while minimizing this score.

CONCLUSION

As we found that our results (for time T=0) agreed with existing background knowledge in biology and our ODEs-based simulator, this paper showed a method to deal with the kinetics of metabolic pathways with a symbolic model (i.e. Fig 1). We explained how to discretize biology experiments into relevant levels to be used with ILP and logic programs in the large. Moreover, based on these discretization of concentration into levels, we explained our process to transform Michaelis-Menten analytical kinetics equation into logic rules, the authors are not aware of any previous work in this direction. Therefore the originality of the work is given by the capacity of a logical model to find the dynamic response of microorganism when a pulse of glucose has been made. We think that this approach improves the accuracy of the metabolic flux analysis. Allowing for other kinds of kinetic modeling (two substrate and/or two products reactions) would enable us to work with more complete models.

As in (King et al., 2005), this approach tries to study the behaviour of many ordinary differential equations while considering a symbolic model with its advantages whereof the statistical evaluation of hypotheses. The process of statistically evaluating hypotheses, thanks to BDD-EM (Inoue et al., 2009), is seen as a good method to find relevant knowledge among the large quantity of processed data. The practical validity of this full process (including discretization) has been shown by the results of this paper while working in a well-known theoretical framework (Inoue, 2004; Mooney, 1997). We strongly believe that the use of time series discretization and a kinetic modeling to enable ILP to deal with ODE will yield great results. We also prefer to consider knowledge discovery as an iterative loop where one must review his knowledge base in the light of new findings (i.e. add "New KB" next turn in Fig. 2).

Still, our modeling can be improved, and time and concentration discretization could be finer. Experiments dealing with more than 3 levels and many time steps will be lead on the Glycolysis and Pentose Phosphate pathways of another bacteria, *Saccharomyces Cerevisiae* (yeast), with both real world data from experiments and simulated data. More experiments with enhancing and updating the knowledge base on this dataset is necessary to get more accurate results. A more global approach of discretizing experimental data and using it in conjunction with automatically generated symbolic pathways extracted from KEGG (Kanehisa and Goto, 2000; Kanehisa et al., 2008) can be applied regardless of the model chosen for infering new knowledge. This approach can be generically applied to turn quantitative results from systems biology into qualitative (symbolic) ones.

REFERENCES

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

Beal, M. (2003). *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Comp. Neurosc. Unit, University College London.

Benhamou, F. (1994). Interval constraint logic programming. *Lecture Notes in Computer Science*, 910.

Chassagnole, C., Rodrigues, J., Doncescu, A., and Yang, L. T. (2006). *Differential evolutionary algorithms for in vivo dynamic analysis of glycolysis and pentose phosphate pathway in Escherichia Coli*. A. Zomaya.

Cheeseman, P. and Stutz, J. (1995). Bayesian classification (autoclass): Theory and results. In *Advances in Knowledge Discovery and Data Mining*, pages 153–180. The MIT Press.

De Raedt, L. (2008). *Logical and Relational Learning*. Springer.

Doncescu, A., Yamamoto, Y., and Inoue, K. (2007). Biological systems analysis using Inductive Logic Programming. In *IEEE International Symp. on Bioinf. and Life Science Computing*.

Dworschak, S., Grell, S., Nikiforova, V., Schaub, T., and Selbig, J. (2008). Modeling biological networks by action languages via answer set programming. *Constraints*, 13(1/2):21–65.

Fages, F., Soliman, S., and France, I. R. (2008). Model revision from temporal logic properties in systems biology. In *In: Probabilistic Inductive Logic Programming. LNAI*, volume 4911, pages 287–304.

Gauvain, J.-L. and Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298.

Geurts, P. (2001). Pattern extraction for time-series classification. *Lecture Notes in Artificial Intelligence*, 2168:115–127.

Inoue, K. (1992). Linear resolution for consequence finding. *Artificial Intelligence*, 56:301–353.

Inoue, K. (2004). Induction as consequence finding. *Machine Learning*, 55:109–135.

Inoue, K., Sato, T., Ishihata, M., Kameya, Y., and Nabeshima, H. (2009). Evaluating abductive hypotheses using and EM algorithm on BDDs. In *Proc. of IJCAI-09*, pages 820–815. AAAI Press.

Ishihata, M., Kameya, Y., Sato, T., and Minato, S. (2008). Propositionalizing the EM algorith by BDDs. Technical report, TR08-0004, Dept. Comp. Sc., Tokyo Instute of Technology.

Ji, S., Krishnapuram, B., and Carin, L. (2006). Variational bayes for continuous hidden markov models and its application to active learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):522–532.

Juvan, P., Demsar, J., Shaulsky, G., and Zupan, B. (2005). Genepath: from mutations to genetic networks and back. *Nucleic Acids Res.*, 33.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, 36:480–484.

Kanehisa, M. and Goto, S. (2000). Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30.

Keogh, E., Lin, J., and Fu, A. (2005). HOT SAX: efficiently finding the most unusual time series subsequence. In *5th IEEE International Conference on Data Mining*.

King, R., Garrett, S., and Coghill, G. (2005). On the use of qualitative reasoning to simulate and identify metabolic pathways. *Bioinformatics*, 21(9):2017–2026.

King, R., Whelan, K., Jones, F., Reiser, P., Bryant, C., Muggleton, S., Kell, D., and Olivier, S. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427:247–252.

Kitano, H. (2002). Systems biology toward system-level understanding of biological systems. *Science*, 295(5560):1662–1664.

Mooney, R. (1997). Integrating abduction and induction in machine learning. In *Working Notes of the IJCAI97 Workshop on Abduction and Induction in AI*, pages 37–42.

Muggleton, S. (1995). Inverse entailment and progol. *New Generation Computing*, 13(3/4):245–286.

Nabeshima, H., Iwanuma, K., and Inoue, K. (2003). SOLAR: A consequence finding system for advanced reasoning. In *Proc. of the 11th International Conference TABLEAUX 2003, LNAI*, volume 2786, pages 257–263.

Peters-Wendisch, P., Schiel, B., Wendisch, V., and et *al.*, E. K. (2001). Pyruvate carboxylase is a major bottleneck for glutamate and lysine production by corynebacterium glutamicum. *Molecular Microbiol. Biotechnol.*, 3(2).

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–286.

Ray, O., Whelan, K., and King, R. (2009). A nonmonotonic logical approach for modelling and revising metabolic networks. *Complex, Intelligent and Software Intensive Systems, IEEE*.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Tamaddoni-Nezhad, A., Chaleil, R., Kakas, A., and Muggleton, S. (2006). Application of abductive ILP to learning metabolic network inhibition from temporal data. *Machine Learning*, 64:209–230.

Tiwari, A., Talcott, C., Knapp, M., Lincoln, P., and Laderoute, K. (2007). Analyzing pathways using SAT-based approaches. In *Proc of the 2nd Int. Conf. on Algebraic Biology*, pages 155–169.