Kinetic Models and Qualitative Abstraction for Relational Learning in Systems Biology
Gabriel Synnaeve, Katsumi Inoue, Andrei Donescu, Hidetomo Nabeshima, Yoshitaka Kameya, Masakazu Ishihata, Taisuke Sato

To cite this version:
Gabriel Synnaeve, Katsumi Inoue, Andrei Donescu, Hidetomo Nabeshima, Yoshitaka Kameya, et al.. Kinetic Models and Qualitative Abstraction for Relational Learning in Systems Biology. BIOSTEC Bioinformatics 2011, Jan 2011, Rome, Italy. <hal-00607296>

HAL Id: hal-00607296
https://hal.archives-ouvertes.fr/hal-00607296
Submitted on 8 Jul 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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 attempts 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.

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:

- KEgg2Sym, 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 $15 \times 3$ 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.

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 developing 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.

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 \textit{Escherichia coli} (Chassagnole et al., 2006). The Fig.4 shows the simplified pathway that we modeled logically with relations 

\[
\text{reaction(Substrate, Enzyme, Product, Km)}
\]

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-

\footnote{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 \((n \neq n'\text{ and } t \neq t')\).}
table 1.

<table>
<thead>
<tr>
<th>#</th>
<th>Metab.</th>
<th>Conc.</th>
<th>Lvl</th>
<th>#</th>
<th>Metab.</th>
<th>Conc.</th>
<th>Lvl</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>glucose</td>
<td>0.055</td>
<td>0</td>
<td>2</td>
<td>g6p</td>
<td>3.480</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>fructose-1,6-bisphosphate</td>
<td>(f6p)</td>
<td>0.595</td>
<td>0</td>
<td>3</td>
<td>fructose-6-phosphate</td>
<td>(f6p)</td>
</tr>
<tr>
<td>4</td>
<td>fructose-6-phosphate</td>
<td>(f6p)</td>
<td>0.218</td>
<td>0</td>
<td>5</td>
<td>pyruvate</td>
<td>(pyr)</td>
</tr>
<tr>
<td>6</td>
<td>pyruvate</td>
<td>(pyr)</td>
<td>2.670</td>
<td>0</td>
<td>7</td>
<td>glyceraldehyde-3-phosphate</td>
<td>(gap)</td>
</tr>
<tr>
<td>8</td>
<td>glyceraldehyde-3-phosphate</td>
<td>(gap)</td>
<td>1.955</td>
<td>0</td>
<td>9</td>
<td>g6p</td>
<td>0.653</td>
</tr>
<tr>
<td>11</td>
<td>nadp</td>
<td>0.195</td>
<td>0</td>
<td>12</td>
<td>nadh</td>
<td>0.100</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>nadph</td>
<td>0.195</td>
<td>0</td>
<td>14</td>
<td>nadh</td>
<td>0.100</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>nad</td>
<td>1.470</td>
<td>1</td>
<td>16</td>
<td>nadh</td>
<td>0.100</td>
<td>0</td>
</tr>
</tbody>
</table>
With only 3 levels, as we have in our discretization of E.Coli experiments, we will get the following simple rules:

- If $[S] < K_m \Rightarrow \frac{V_m}{V_m - [S]} \Rightarrow |P|_{T+1} = |P|_T \text{reaction}(S, P, Km) \land \text{conc}(S, 0, T) \land \text{conc}(Km, 2, T) \land \text{conc}(P, L, T)$

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

- If $[S] > K_m \Rightarrow \frac{V_m}{V_m - [S]} \Rightarrow |P|_{T+1} = V_m + |P|_T \text{reaction}(S, P, Km) \land \text{conc}(S, 2, T) \land \text{conc}(Km, 0, T) \land \text{conc}(P, L, T)$

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).

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 $\text{conc}(\cdot, \cdot, 0)$. Using SOLAR, we get 98 hypotheses as: $H76 = \text{conc}(\text{g6p}, 2, 0) \land \text{conc}(\text{adp}, 2, 0)$

$\land \text{conc}(\text{gap}, 0, 0) \land \text{conc}(\text{adh}, 0, 0) \land \text{conc}(\text{gp3}, 2, 0) \land \text{conc}(\text{pep}, 2, 0) \land \text{conc}(\text{atp}, 0, 0) \land \text{conc}(\text{pyr}, 2, 0)$

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 $\text{conc}(\text{Compound}, \text{Level}, \text{Time})$ and $\text{reaction}(\text{Substrate}, \text{Product}, \text{Reaction})$.

Each of the elements of $\mathcal{A}$ is a boolean variable. One of its subsets is the subset of abducibles $\Gamma$ composed of all the possible values of $\text{conc}(\text{Compounds}, \text{Level}, 0)$. 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 $B = (H_1 \lor \cdots \lor H_n) \land \text{ground}(B)$ to set the good $\Theta$ 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:

<table>
<thead>
<tr>
<th>Hyp. #</th>
<th>Probability</th>
<th>Abduced conc. levels at T=0</th>
</tr>
</thead>
<tbody>
<tr>
<td>H76</td>
<td>$\approx 1.000$</td>
<td>g6p: 2, adp: 2, fdp: 0, fdp: 0, dhap: 0, gap: 0, glucose: 2, pg3: 2, pep: 2, atp: 0, pyr: 2</td>
</tr>
<tr>
<td>H41</td>
<td>0.822</td>
<td>the same as H76 except pg3: 0</td>
</tr>
<tr>
<td>H56</td>
<td>0.625</td>
<td>the same as H76 except g6p: 0</td>
</tr>
<tr>
<td>H70</td>
<td>0.553</td>
<td>the same as H76 except adp: 2</td>
</tr>
<tr>
<td>H13</td>
<td>0.515</td>
<td>the same as H76 except ghap: 0</td>
</tr>
<tr>
<td>H90</td>
<td>0.455</td>
<td>the same as H76 except pg3: 0</td>
</tr>
<tr>
<td>H82</td>
<td>0.442</td>
<td>the same as H76 except dhap: 2</td>
</tr>
<tr>
<td>H43</td>
<td>0.369</td>
<td>the same as H76 except pyr: 1</td>
</tr>
<tr>
<td>H9</td>
<td>0.364</td>
<td>the same as H41 except adp: 2</td>
</tr>
<tr>
<td>H68</td>
<td>0.346</td>
<td>g6p: 0, adp: 0, fdp: 0, fdp: 0, dhap: 0, gap: 0, glucose: 2, pg3: 2, pep: 2, atp: 2</td>
</tr>
</tbody>
</table>

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.

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)
discourable.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) + λ · 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
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 microorganisms 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 a 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 inferring new knowledge. This approach can be generically applied to turn quantitative results from systems biology into qualitative (symbolic) ones.

REFERENCES


