Kinetic Models for Logic-Based Hypothesis Finding in Metabolic Pathways

Gabriel Synnaeve[1,2], Andrei Doncescu[1,3], and Katsumi Inoue[1]

[1] National Institute of Informatics, Japan
[2] Grenoble University (IMAG), France
[3] LAAS-CNRS 31007 Toulouse, France

Abstract. In this paper we propose a logical model of Glycolysis and Pentose phosphate pathways of *E.Coli* that allows the analysis of the dynamical response of a biological system perturbed by a pulse of glucose. Our goal is to give a better comprehension of the physiological state of the cell from a better interpretation of the interactions between metabolic and signaling networks. Starting with the discretization of concentrations of some metabolites considered in steady state, we fully explain our approach to build a symbolic model applied to kinetics. Finally, hypothesis finding produces logical formulas on the metabolites through abduction which cannot be measured during dynamical state.

1 Introduction

Nowadays, bioinformatics 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 [1]. The majority of kinetic models in biology is described by coupled differential equations and simulators are implemented with the appropriate methods to solve these systems. However, for most nonlinear dynamical systems it is difficult to find an analytical solution, even its existence had still to be proved [2]. The understanding of the phenomenon described by a complex system is carried out by a qualitative study of its behaviour such as stability or forking. Our goal is to elaborate symbolic models of these systems in order to discover the mechanisms that govern them. For that, we clusterize continuous concentrations of metabolites over time into discrete levels and discrete timesteps. Then, we worked on an inverse problem: given the measured concentrations of some metabolites in steady state, we compute the concentrations of metabolites before the dynamic transitionto this steady state thanks to our kinetic model.

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

2 Modeling of E-coli Central Metabolism

To obtain an understanding of the central metabolism, a logical model has been developed according with the kinetic model including the glycolysis and the pentose phosphate pathway for *Escherichia coli* [4]. The structure of such networks is commonly displayed on metabolic maps, where each reaction is described in terms of the participating enzyme, metabolites, cofactors and the reaction stoichiometry. These chemical reactions and transport steps can be thought of as the primary connections between metabolite pools that affect each other by mass action. For the purpose of analysing the dynamic behaviour of metabolic systems, we consider the same chemical compound in different pools as separate metabolites. The metabolic networks dynamics are in their enzymatic part ruled by the combination of classical kinetics: essentially Michaelis-Menten, Hill and allosteric ones. The choice of Michaelis-Menten kinetic model have been made, because it is the most general representation for a non-linear allosteric regulation system. It assumes that the two binding equilibria are rapid when compared to the interconversion of ES and EP.

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

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

If both S and 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*. 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 choiced to work with a constant timestep :

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

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: glucose6phosphate, fructose6phosphate, fructose1-6bisphosphate, glyceraldehyde3phosphate, phospho-enolpyruvate, pyruvate, 6phosphate-gluconate,

glucose1phosphate as well as the cometabolites: atp, adp, amp, nad, nadh, nadp, nadph were measured using enzymatic methods or *High Performance Liquid Chromatography*. All the measured steady-state concentrations and their corresponding discrete levels are summarized in Table 1.

Metabolite	Concentration	Level	Metabolite	Concentration	Level
glucose	0.0556	0	g6p	3.480	2
f6p	0.600	0	fdp	0.272	0
gap	0.218	0	pep	2.670	2
pyr	2.670	2	6pg	0.808	1
g1p	0.653	0	amp	0.955	1
adp	0.595	0	atp	4.270	2
nadp	0.195	0	nadph	0.062	0
nad	1.470	1	nadh	0.100	0

Table 1. Concentrations (mM) and their discretized levels for steady states

3 Discretization of Continous Values

Discretizing time series is a research domain on its own and many works [8, 9] have been conducted recently. Our practical problem is that we want to have a statistically relevant (unsuppervised) discretization for N chemical compounds concentrations over time. For that purpose, we compute an appropriate number of levels (that was 3 for *E.Coli*) in regard to a Bayesian score such as Bayesian Information Criterion [10]. We use continuous (Gaussian) hidden Markov models with parameter tying, which means that each chemical compound has a corresponding HMM but all the N Gaussian HMM share the same parameters (means and covariances), to share the same discrete outputed levels between the different compounds of one experiment. This relevant discretized levels of concentration are computed through expectation maximisation with maximum a posteriori [11, 12] or through variational Bayes EM [13]. Then, we use a very simple round-mean aggregation of them for time-sampling. We intend to do future different works in the direction of discretization of our time-series from molecular biology experiments but current results are already useable (see Table 1. and Fig.2).

4 Model Analysis by ILP

Inductive Logic Programming, used for induction or abduction [7], is able to deal with discrete levels and qualitative rules [6]. Its goal is to find Hypothesis such as Background $\wedge$ Hypothesis $\models$ Examples. In [5], Inoue proposed a simple, yet powerful method to handle inverse entailment for computing inductive hypotheses. The resulting method called *CF-induction* does not restrict the bridge formula U as the set of literals entailed by $B \wedge \neg E$, but consider the *characteristic clauses* of $B \wedge \neg E$, which obviously generalizes the method of the bottom

clause. CF-induction then realizes sound and complete hypothesis finding from *full clausal theories*, and not only definite clauses but also non-Horn clauses and integrity constraints can be constructed as H.

The logical model used by CF-induction is based on the simplified Michaelis-Menten equation **(3)** which has been here represented by 3 background clauses using the `Conc(Compound, Level, Time)` predicate (for concentration):

$$[S] \ll K_m \Rightarrow \frac{\Delta[P]}{\Delta T} = \frac{V_m}{K_M} \Rightarrow [P]_{T+1} = [P]_T$$

`Conc(s, 0, 0)` $\wedge$ `Conc (Km, 2, 0)` $\wedge$ `Conc(p, L, 0)` $\rightarrow$ `Conc(p, L, 1)`

$$[S] \simeq K_m \Rightarrow \frac{\Delta[P]}{\Delta T} = \frac{V_m}{2} \Rightarrow [P]_{T+1} = V_m/2 + [P]_T$$

`Conc(S, 1, 0)` $\wedge$ `Conc (Km, 1, 0)` $\wedge$ `Conc(p, L, 0)` $\rightarrow$ `Conc(P, L, 1)`

$$[S] \gg K_m \Rightarrow \frac{\Delta[P]}{\Delta T} = V_m \Rightarrow [P]_{T+1} = V_m + [P]_T$$

`Conc(S, 2, 0)` $\wedge$ `Conc (Km, 0, 0)` $\rightarrow$ `Conc(P, 2, 1)`

5 Experiments and Results

We developed the beginning of an automated framework to deal with different real world pathways and experiments. It is currently composed of two tools:

- `kegg2symb`, written in Python and using KEGG API, that transform pathways from KEGG [14, 15] into symbolic models.
- The combination of HMM Utility Program (that computes continuous HMM [12, 13]) with `py-tsdisc`, a Python automating wrapper.

Here, using CF-induction with 3 levels and simplified Glycolysis and Pentose Phosphate pathway, we obtain many hypothesis including this one:

$concentration(glucose, 0, 1) \wedge concentration(glucose, 2, 0) \wedge concentration(pyr, 2, 0)$

This hypothesis is corresponding to our biological knowledge that pyruvate is a bottleneck [16] and that the glucose that is totally consummed (see Fig.1 from simulation) was in high concentration at the beginning of the experiment (pulse).

6 Perspectives and Conclusion

Experiments dealing with more than 3 levels through a `compute` predicate implemented in SOLAR [17] are being lead on the Glycolysis and Pentose Phosphate

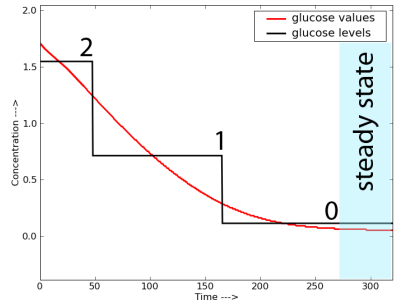

Fig. 1. Pulse of glucose in the Glycolysis Pathway of *E.Coli* with 3 levels

pathways of *Saccharomyces Cerevisiae* (yeast) with both real world data from experiments and simulated data.
This paper showed one mean to discretize biology experiments into relevant levels to be used with ILP and logic programs in the large, the authors are not aware of any previous work in this direction. Also, we explained our processus to transform Michaelis-Menten analytical kinetics equation into logic rules. This processus can be generically applied to turn quantative results into qualitative (symbolic) ones.

Acknowledgments

Many thanks go to Yoshitaka Kameya for his implementation of continuous HHM-based clustering (HMM Utility Program) and his kind explanations. Thanks to Taisuke Sato for his supervising. Thanks also to Yoshitaka Yamamoto and Takehide Soh for their kind support at NII.

References

1. Kitano H.: Systems Biology Toward System-level Understanding of Biological Systems Kitano. In Science Vol. 295. no. 5560, pp. 1662-1664 (2002).
2. Mirrahimi E., Rouchon P., Turinici G.: Lyapunov control of bilinear Schrdinger equations. Automatica 41(11):1987-1994 (2005)
3. Tamaddoni-Nezhad A., Chaleil R., Kakas A., and Muggleton S.H.: Application of abductive ILP to learning metabolic network inhibition from temporal data. Machine Learning, 64:209-230 (2006).
4. Chassagnole C., Rodrigues J. C., Doncescu A. and Tianruo Yang L.: Differential evolutionary algorithms for in vivo dynamic analysis of glycolysis and pentose phosphate pathway in Escherichia Coli. In: A. Zomaya (ed.), *Parallel Computing in Bioinformatics and Computational Biology*, ISBN 0-471-71848-3.

5. Inoue K.: Induction as consequence finding. Machine Learning, 55:109–135 (2004).
6. Doncescu A., Yamamoto Y., Inoue K.: Biological Systems Analysis using Inductive Logic Programming. The 2007 IEEE International Symposium on Bioinformatics and Life, Science Computing BLSC07 (2007).
7. R. J. Mooney: Integrating abduction and induction in machine learning. In Working Notes of the IJCAI97 Workshop on Abduction and Induction in AI, 37–42 (1997).
8. Geurts P.: Pattern extraction for time-series classification. Proceedings of PKDD 2001, 5th European Conference on Principles of Data Mining and Knowledge Discovery, LNAI 2168, 115-127 (2001).
9. Keogh, E. and Lin, J. and Fu, A.: HOT SAX: efficiently finding the most unusual time series subsequence. 5th IEEE International Conference on Data Mining (2005).
10. Schwarz G.: Estimating the dimension of a model. Annals of Statistics, Vol.6, No.2, pp.461-464 (1978).
11. Beal M. J.: Variational Algorithms for Approximate Bayesian Inference. PhD. Thesis, Gatsby Computational Neuroscience Unit, University College London (2003).
12. Gauvain J.-L. and Lee C.-H.: Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Transactions on Speech and Audio Processing, Vol.2, Issue 2, pp.291-298 (1994).
13. Ji S., Krishnapuram B., and Carin L.: Variational Bayes for continuous hidden Markov models and its application to active learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.28, Issue 4, pp.522-532 (2006).
14. Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000).
15. Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y.; KEGG for linking genomes to life and the environment. Nucleic Acids Res. 36, D480-D484 (2008).
16. Peters-Wendisch P. G., Schiel B., Wendisch V. F., Efstratios Katsoulidis E., Mckel B.: Pyruvate Carboxylase is a Major Bottleneck for Glutamate and Lysine Production by Corynebacterium glutamicum. J. Mol. Microbiol. Biotechnol. 3(2) (2001).
17. Nabeshima H., Iwanuma K., and Inoue K.: SOLAR: A Consequence Finding System for Advanced Reasoning. Proceedings of the 11th International Conference TABLEAUX 2003, LNAI, Vol. 2796, pp. 257-263, Springer (2003).

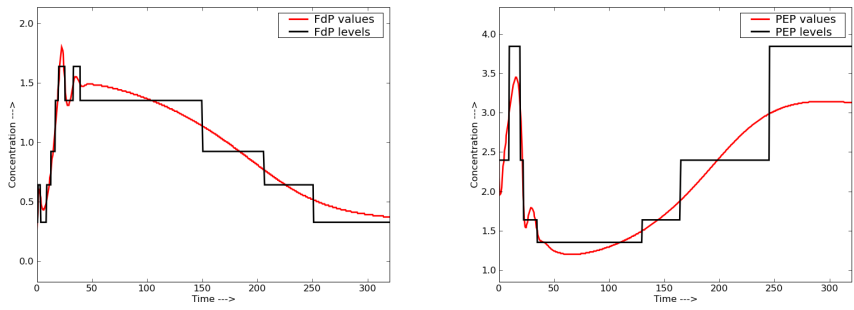

Fig. 2. Pulse of glucose in the Glycolysis Pathway of *S.Ce* with 7 levels