2011 International Conference on Complex, Intelligent, and Software Intensive Systems

Using Machine Learning Techniques for Modelling and Simulation of Metabolic Networks

Marenglen Biba
Department of Computer Science
University of New York in Tirana
Tirana, Albania
Email: marenglenbiba@unyt.edu.al

Fatos Xhafa
Department of Languages and Informatic Systems
Techical University of Catalonia
Barcelona, Spain
Email: fatos.xhafa@gmail.com

Floriana Esposito
Department of Computer Science
University of Bari
Bari, Italy
Email: esposito@di.uniba.it

Stefano Ferilli
Department of Computer Science
University of Bari
Bari, Italy
Email: ferilli@di.uniba.it

***Abstract*—Metabolomics is increasingly becoming an important field. The fundamental task in this area is to measure and interpret complex time and condition dependent parameters such as the activity or flux of metabolites in cells, their concentration, tissues elements and other biosamples. The careful study of all these elements has led to important insights in the functioning of metabolism. Recently, however, there is a growing interest towards an intagrated approach to studying biological systems. This is the main goal in Systems Biology where a combined investigation of several components of a biological system is thought to produce a thorough understanding of such systems. Metabolic networks are not only structurally complex but behave also in a stochastic fashion. Therefore, it is necessary to express structure and handle uncertainty to construct complete dynamics of these networks. In this paper we describe how stochastic modeling and simulation can be performed in a symbolic-statistical machine learning (ML) framework. We show that symbolic ML deals with structural and relational complexity while statistical ML provides principled approaches to uncertainty modeling. Learning is used to analyze traces of biochemical reactions and model the dynamicity through parameter learning, while inference is used to produce stochastic simulation of the network.**

I. Introduction

Metabolomics has rapidly developed into an important field [1], due also to the evolution of techniques and instruments for gathering, storing and analyzing voluminous metabolic data such as Mass Spectrometry (MS) or Nuclear Magnetic Resonance (NMR). With the current data capturing technologies MS is able to detect molecules at very low concentrations such as 10^{-18} molar, and high-field NMR can efficiently differentiate between molecules that are structurally very similar. The fundamental task in metabolomics is the study of the metabolome which represents the collection of all the metabolites in a biological organism [2]. This set of molecules consists of many metabolic intermediates, hormones and other signalling molecules, and secondary metabolites. Together, all these form the chemical fingerprints that all specific cellular processes leave behind during their happening. Therefore, in order to understand how cells work it is important to thoroughly explore and understand the metabolome in a principled and robust manner. Nevertheless, the classical approach has been a separate study of the metabolome from the remaining part of the system, which would not give a deep comprehension of the organism, since biological systems' behavior is determined by complex interactions between their building components. As a consequence, an integrated approach to studying biological systems has become necessary. This has led to the birth of the outstanding and challenging research area of Systems Biology (SB) [3]. In SB, the central problem is to uncover and model how function and behavior of the biological machinery are implemented through complex interactions among its building blocks. Metabolomic data provide precious traces of the cell's circuits functioning, hence it has become essential for the SB approach to integrate metabolomics techniques for a deeper understanding of the overall functioning of biological systems [4].

Biological circuits are extremely hard to model and simulate and important efforts are being made to develop computational models that can truly represent them and handle their intrinsic complexity [5]. In this work we focus on a particular problem of SB that concerns the modeling of metabolic pathways, their stochastic simulation and the possibility to discover biologically active paths through simulation after parameters have been learned from the data. A metabolic pathway is a number of sequential chemical reactions occurring within the cell. These reactions are catalyzed by enzymes which are particular proteins that convert metabolites (input molecules) into other molecules that represent the products of the reaction. The products can be stored in the cell under certain forms or can cause the initiation of another metabolic pathway. The metabolic network of a cell is formed by all the metabolic pathways occurring in the cell. It is through the metabolic networks that every single living organism carries out all its activities. Therefore, analyzing pathways is crucial in order to understand cell's behavior. ML methods have the power not only to learn parameters or structure from the data but also to

978-0-7695-4373-4/11 $26.00 © 2011 IEEE
DOI 10.1109/CISIS.2011.22

simulate stochastic biological networks through inference procedures. Learning, meant as reconstruction of structure and the related probabilistic parameters, is essential to infer knowledge from exponentially growing observation data gathered by high-throughput instruments. On the other side, once the model has been reconstructed through learning from data, simulation is fundamental to run the model and experiment variations in parameters in order to understand how the dynamics and the behavior of the model changes. This requires efficient and effective inference ML procedures that given a model (the structure and the parameters) and a query (the values of input metabolites), is able to produce a stochastic simelation and produce the most probable reaction path in the network.

Reactions can happen if the input molecules are available to the catalytic enzyme, thus a modeling framework must be able to model relations among entities. ML methods based on symbolic approaches such as logic-based techniques have the potential to model relations in structural complex domains. In particular, first-order logic representations have also the advantage that models are easily comprehensible to humans. On the other side, most part of biological systems performs its activity remaining hidden to the human modeler and here ML techniques can play an important role in discovering latent phenomena. However, symbolic-only approaches suffer the incapability of dealing with uncertainty or probabilistic representations. In models built with symbolic-only approaches, the learned rules are deterministic and do not incorporate any kind of mechanism for uncertainty modeling or simulation. Thus only relying on structural representation is not enough since we need to represent stochastic worlds. Biological systems intrinsically behave in a stochastic fashion with many interactions probable to happen. Since cell's life is determined at any instanct by the most probable interactions, handling uncertainty is crucial when the cell's machinery must be modeled and simulated. Statistical ML approaches based on probability theory represent a valuable mechanism to govern uncertainty. Moreover, observations of biological systems rarely reflect outside exactly what happens inside them and many variables remain hidden. For this reason, estimation techniques are precious in order to model what cannot be observed. Statistical ML has shown outstanding results in the ability to learn probability distributions from observations and hence is a suitable approach to modeling biological systems. Statistical inference algorithms developed in the ML area have also the power to produce reliable stochastic runs of the model for simulation purposes. On the other side, statistical-only approaches rarely are able to reason about relations and/or capture interactions among biological circuits as symbolic approaches do. Hence, there is strong motivation on developing and applying hybrid approaches to modeling biological systems. On one side, symbolic methods represent the structure of the model and on the other side statistical methods represent the stochastic nature of the model. A fixed structure without any probabilistic representation would only give a static view of the possible reactions, but no information can be deduced on the real behaviour of the biological system.

Machine learning, Pattern Recognition and Data Mining communities have long focused their attention on vector data which is mainly independent and identically distributed. However, since in the real-world, data is usually stored in relational databases and almost always involves interactions among entities and their attributes, relational worlds pose for ML the serious problem of learning from relational and non i.i.d data. Moreover, a critical problem in data analysis tasks is that most relational real-world databases are noisy and present a lot of missing data. This characteristic of real-world data has a strong impact on the performance of standard data analysis algorithms making them very unsuccessful for real tasks.

Recently, to deal with both aspects, rich relational structure and almost always noisy data, statistical relational models [6] are being developed in order to analyze data from noisy relational databases. These powerful models exploit the power of statistics to properly handle uncertainty of stochastic worlds and logic-based formalisms to represent relations among involved entities. Building rich relational-stochastic models has a long history in artificial intelligence and ML and starts with the works in [7], [8], [9]. Then several authors began using logic programs to define compact Bayesian networks for complex application scenarios. This approach was known as knowledge-based model construction [10]. Recently, several approaches for combining logic and statistics have been proposed in [11], [12], [13], [14]. The advantage of these models is that they are able to represent probabilistic dependencies between attributes of related different objects in a certain domain.

The contribution of this paper is at the intersection of Systems Biology, Metabolomics and Machine Learning. We show how in a ML framework, PRISM (PRogramming In Statistical Modelling) [15]),which is a symbolic statistical modeling language that integrates logic programming with algorithms for probabilistic learning and inference, we can first model metabolic networks, then learn parameters of the dynamic model from traces of biochemical reactions and finally perform stochastic simulations of the model. We show through experiments the feasibility of discovering significant active paths from metabolomics data in the form of traces of sequences of reactions, by running and simulating the reconstructed model.

The paper is organized as follows. Section 2 describes the problem of modeling metabolic pathways and the necessity for symbolic-statistical ML. Section 3 describes the hybrid learning and inference framework PRISM. Section 4 describes modeling in PRISM of the Bisphenol A Degradation pathway of Dechloromonas aromatica. Section 5 presents experiments on learning parameters from generated sequences of reactions. Section 6 presents experiment on simulation of the model with the learned and changed parameters. Section 7 describes related work and Section 8 concludes discussing future work.

II. Metabolic Networks

Metabolic networks of a cell are formed by all the metabolic pathways occurring in the cell. It is through the metabolic

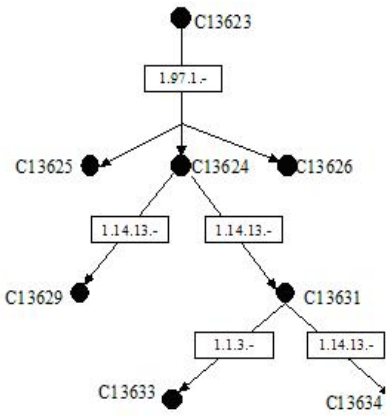

Fig. 1. Part of the pathway of Bisphenol A Degradation.

networks that every single living organism carries out all its tasks. Formally, metabolic pathways can be represented as graphs where each node represents a chemical compound and a chemical reaction corresponds to a directed edge labeled by a protein that catalyzes the reaction. Therefore, there is an edge from one compound (metabolite) to another compound (product) if there is an enzyme that transforms the metabolite into product.

Figure 1. shows one part of the pathway for the Bisphenol A Degradation in Dechloromonas aromatica extracted from the KEGG database. We present this pathway from the KEGG since, as we can see from the figure, starting from one point in the pathway there are multiple paths that can be explored. The path that is actually taken depends on the stochastic behavior of the system, therefore the task of modeling a metabolic network is not just of a mere structural repsresentation of the graph but we need to represent also its dynamic behaviour.

In order to model metabolic pathways, a suitable framework for their simulation must be able to handle relations. First-order logic representations have the expressive power to model structural and relational problems. The metabolic pathway in Figure 1 would easily be represented in a first-order logic formalism as follows:

$enzyme(1.97.1.-, reaction\_1\_97\_1\_, [c13623], [c13625, c13624, c13626]).$
$enzyme(1.14.13.-, reaction\_1\_14\_13\_a, [c13624], [c13629]).$
$enzyme(1.14.13.-, reaction\_1\_14\_13\_b, [c13624], [c13631]).$
$enzyme(1.1.3.-, reaction\_1\_1\_3, [c13631], [c13633]).$
$enzyme(1.14.13.-, reaction\_1\_14\_13\_c, [c13631], [c13634]).$

However, the above representation does not incorporate any further information regarding the reaction process. For instance, we can find that there are two competing reactions since enzyme 1.14.13.- catalyzes two different reactions with the same chemical compound c13624 in input. Subsequently, two enzymes, 1.14.13.- and 1.1.3.-, can elaborate the same input metabolites and thus two reactions compete among them. The happening of any of the reactions determines a certain sequence of successive reactions instead of another. Thus, at a certain moment, it is fundamental to know which reaction among the two is more probable to happen. The most probable reaction determines the path that is followed under certain conditions. This means with certain parameters, a biological pathway becomes inactive or useless and another pathway may become active and yield different overall products in the whole pathway. The conditions under which the reactions happen in cascade, change due to the stochastic behavior of the biological environment. For example, some input metabolites can suddenly be not available. This absence can inhibit a certain reaction and allow another to happen leading thus to a different sequence in the chain of reactions. Therefore, it is crucial to represent in the model how probable a certain reaction is. From a modeling point of view, this situation can be modeled by attaching to each reaction the probability that it happens. From a representation point of view, this requires a first-order framework that can represent and use for each predicate (that expresses a reaction) the probability that the predicate is true.

Simply incorporating probabilities in the static graph is not enough to model complex large metabolic networks. The conditions for the reactions to happen depend on many external and/or internal factors, such as the initial quantity of input metabolites, changes in the physical-chemical environment surrounding the cell and many more. Due to these reasons, it is extremely hard to observe all the states of the biological machinery under all the possible conditions and try to assign probabilities to reactions. Therefore we strongly need statistical ML methods that given certain conditions can learn probability distributions from observations (the conditions here are meant as physical-chemical entities such as temperature, concentration of metabolites, entropy etc).

Modeling and simulation of metabolic networks requires two tasks must be performed. First, a relational stochastic model that describes the structure of the metabolic network must be build. There is now available a large amount of accumulated knowledge about the structure of metabolic networks such as in KEGG and we can use all this background knowledge to avoid the structure building process and concentrate on reconstructing the dynamics of the model by analyzing observed raw wet experimental data. In fact, graph structures have become indeed abundant but their disadvantage in modeling cell's life is that they are static and do not show any dynamic feature. For instance, the pathway in Figure 1. does not express the stochastic dynamics in metabolic reactions. Currently available graphs in databases such as KEGG, can be cast as useful static templates to interpret what can happen in the cell, but to faithfully model and simulate the cell's activity we must build a dynamic model that represents at a certain moment and under certain conditions what happens inside the cell.

As a consequence, to model and simulate metabolic networks, we must first learn a stochastic model from sequences of reactions that have been observed under certain conditions. Once the model has been recontructed with a certain reliability, we may perform simulations of the network by changing

the parameters of the model and performing stochastic runs in order to observe the outcome of the chain or reactions. The final outcome of the stochastic execution of the model, is usually a set of metabolites. We also perform stochastic simulations to see which is the most probable path in the network by computing the probability of the related predicate through the Viterbi algorithm.

III. The Machine Learning Framework PRISM

PRISM (PRogramming In Statistical Modelling) [15] is a symbolic statistical modeling language that integrates logic programming with algorithms for probabilistic learning and inference. It is a combination of declarative programming with statistical procedures for parameter estimation and inference. PRISM programs represent not only just a probabilistic extension of logic programs in order to handle probabilities, but are also able to learn these probabilities from examples through the EM (Expectation-Maximization) algorithm which is built-in in the language. PRISM represents a formal knowledge representation language for modeling scientific hypotheses about phenomena which are governed by rules and probabilities. Rules help express the structure and relationships of the world, while probabilities help express stochastic phenomena. The parameter learning algorithm [16], provided in the language as a built-in procedure, is a new EM algorithm called graphical EM algorithm that when combined with the tabulated search has the same time complexity as existing EM algorithms, i.e. the Baum-Welch algorithm for HMMs (Hidden Markov Models), the Inside-Outside algorithm for Probabilistic Context-Free Grammars (PCFGs), and the one for singly connected Bayesian Networks (BNs) that have been developed independently in each research field.

PRISM programs can be arbitrarily complex and no restriction on the form or size of programs is posed. This is a fundamental property of this knowledge representation formalism which is at the same time a powerful ML framework. Due to this important property of the framework, the most popular probabilistic modeling formalisms such as HMMs, PCFGs and BNs can be described by general PRISM programs without any loss of expressiveness or learnability.

PRISM programs can be defined as logic programs with a probability distribution over the world facts that is called basic distribution. From a formal point of view a PRISM program can be written as $P = F \cup R$ where R is a set of logical rules working behind the observations and F is a set of facts that models observations' uncertainty with a probability distribution.

Based on the built-in graphical EM algorithm the parameters (probabilities) of F are learned and this learned probability distribution over the facts. through the rules induces a joint probability distribution over the set of least models of P, i.e. over the observations. This is known as distributional semantics [17]. As an example, we illustrate this induction process with a hidden markov model with two states slightly modified from that in [16]:

$values(init, [s0, s1]).$ % State initialization
$values(out(\_), [a, b]).$ % Symbol emission
$values(tr(\_), [s0, s1]).$ % State transition

$hmm(L) :-$ % To observe a string L:
$str\_length(N),$ % Get the string length as N
$msw(init, S),$ % Choose an initial state randomly
$hmm(1, N, S, L).$ % Start stochastic transition (loop)

$hmm(T, N, \_, []) :- T > N, !.$ % Stop the loop
$hmm(T, N, S, [Ob|Y]) :-$ % Loop: current state and time are S and T
$msw(out(S), Ob),$ % Output Ob at the state S
$msw(tr(S), Next),$ % Transit from S to Next.
$T1 \; is \; T + 1,$ % Count up time
$hmm(T1, N, Next, Y).$ % Go next (recursion)

$str\_length(10).$ % String length is 10

set_params :- $set\_sw(init, [0.9, 0.1]), set\_sw(tr(s0), [0.2, 0.8]),$
$set\_sw(tr(s1), [0.8, 0.2]), set\_sw(out(s0), [0.5, 0.5]),$
$set\_sw(out(s1), [0.6, 0.4]).$

The most important characteristic of PRISM is that it allows the users to use random switches to have probabilistic choices in the model representation and running. Every random switch has a name, a space of possible outcomes, and a probability distribution associated. In the program above that illustrates an HMM, $msw(init, S)$ probabilistically determines the initial state from which to start by tossing a coin. The predicate $set\_sw(init, [0.9, 0.1])$, states that the probability of starting from state $s0$ is 0.9 and from state $s1$ is 0.1.

The predicate *learn* in PRISM is used to learn from examples (a set of strings) the parameters (probabilities of init, out and tr) so that ML (Maximum-Likelihood) is achieved. For example, the learned parameters from a set of examples can be:

$switch \; init : s0(0.6570), s1(0.3429);$
$switch \; out(s0) : a(0.3257), b(0.6742);$
$switch \; out(s1) : a(0.7048), b(0.2951);$
$switch \; tr(s0) : s0(0.2844), s1(0.7155);$
$switch \; tr(s1) : s0(0.5703), s1(0.4296).$

After learning all these ML parameters, we can calculate the probability of a certain observation using the predicate *prob*:

$prob(hmm([a, a, a, a, a, b, b, b, b, b])) = 0.000117528.$

This way, we are able to define a probability distribution over the strings that we observe. Therefore from the basic distribution we have induced a joint probability distribution over the observations.

In simulating the constructed model, we are interested in the joint probability distribution induced over the observations by running the model. These runs are stochastic which means the execution of the program is defined by the probability distribution of the switches. By simulating the model we can observe what is the effect of changing parameters of the basic distribution, on the output of the model. This gives further insight into the structure of the model and how can we optimize it and its output by changing the parameters of the model.

IV. Modeling Metabolic Networks in PRISM

Since PRISM is a logic-based language that has the capability of expressing relations and structure, we can easily represent the metabolic pathway presented in the previous section. Predicates that describe reactions remain unchanged from a model representation point of view.

What we need is the addition of the stochastic view of the metabolic pathway. We perform this extension with the introduction of random switches in the logic program that describe the pathway. We may define this way for every reaction a random switch with its relative space outcome. For example, in the following we describe the random switches for the reactions in Figure 1.

$values(switch\_rea\_1\_97\_1, [rea\_1\_97\_1(yes, yes, yes, yes),$
$rea\_1\_97\_1(yes, no, no, no)]).$
$values(switch\_rea\_1\_14\_13\_a, [rea\_1\_14\_13\_a(yes, yes),$
$rea\_1\_14\_13\_a(yes, no)]).$
$values(switch\_rea\_1\_14\_13\_b, [rea\_1\_14\_13\_b(yes, yes),$
$rea\_1\_14\_13\_b(yes, no)]).$
$values(switch\_rea\_1\_1\_3, [rea\_1\_1\_3(yes, yes),$
$rea\_1\_1\_3(yes, no)]).$

$values(switch\_rea\_1\_14\_13\_b, [rea\_1\_14\_13\_c(yes, yes),$

$rea\_1\_14\_13\_c(yes, no)]).$

For each of the three reactions there is a random switch that can take one of the predefined values at every moment in time. For instance, the expression $rea\_1\_97\_1(yes, yes, yes, yes)$ means that at a certain moment the metabolite $c13623$ is available and the reaction occurs producing the compounds $c13623, c13624$ and $c13625$. The stochastic execution of the model is handled by PRISM which during the execution of the logic program makes random choices in the selection of the logical goals to prove according to the parameters of the models.

While the other expression $rea\_1\_97\_1(yes, no, no, no)$ indicates that the input metabolite is present but the reaction did not occur during the stochastic execution of the model, thus the final metabolites of the reaction are not produced. Below we report the remaining part of the PRISM program for modeling the pathway in Figure 1.

Together with the declarations in Section 2 for the possible reactions and those of the previous paragraph for the values of the random switches, the following logic program forms a model for stochastic modeling and execution of the pathway in Figure 1.

$produces(Metabolites, Products) :-$
$produces(Metabolites, [\,], Products).$
$produces(Metabolites, Delayed, Products) :-$
$(reaction(Metabolites, Reaction, Inputs, Outputs, Rest) ->$
$call\_reaction(Reaction, Inputs, Outputs, Call),$
$rand\_sw(Call, Value),$
$((Value == rea\_1\_97\_1(yes, yes, yes, yes);$
$Value == rea\_1\_14\_13\_a(yes, yes,);$
$Value == rea\_1\_14\_13\_b(yes, yes,);$
$Value == rea\_1\_14\_13\_c(yes, yes,);$
$Value == rea\_1\_1\_3(yes, yes)) ->$
$produces(Rest, Delayed, Products)$
;
$produces(Metabolites, [Reaction|Delayed], Products)$
)
;
$Products = Metabolites$
).
$rand\_sw(ReactAndArgs, Value) :-$
$ReactAndArgs = ..[Predicate|Arguments],$
$(Predicate == rea\_1\_97\_1- > msw(switch\_rea\_1\_97\_1, Value);$
$(Predicate == rea\_1\_14\_13\_a- > msw(switch\_rea\_1\_14\_13\_a, Value);$
$(Predicate == rea\_1\_14\_13\_b- > msw(switch\_rea\_1\_14\_13\_b, Value);$
$(Predicate == rea\_1\_14\_13\_c- > msw(switch\_rea\_1\_14\_13\_c, Value);$
$(Predicate == rea\_1\_1\_3- > msw(switch\_rea\_1\_1\_3, Value)$
;
$true))))).$ % do nothing

In the following, we trace the stochastic execution of the above logic program. This corresponds to a stochastic simulation of the model with the available parameters. A different simulation would change in the parameters of the model that are used in the execution. It is important to note that various simulations can be performed by only choosing a different set of parameters. This way, a thorough analysis of the behaviour of the model can be observed depending on the conditions which are usually expressed with the proabilities of the basic distribution.

In our stochastic model, the top goal to prove for PRISM, that represents in this case the observations (sequences of reactions vastly produced by high-throughput technologies) is $produces(Metabolites, Products)$. This goal will succeed if there is a pathway in the stochastic execution of the model that will lead from Metabolites to Products, in other words if there is a sequence of random choices (according to a probability distribution) that makes possible to prove the top goal.

The predicate *reaction* controls among the first clauses of the program, if there is a possible reaction with Metabolites in input. Suppose that at a certain moment $Metabolites = [c13624]$ and thus two reactions can happen and start being in competition with each other. Suppose one of the reaction is randomly chosen and the variables Inputs and Outputs are bounded respectively to $[c13624]$ and $[c13629]$. The predicate $call\_reaction$ constructs the body of the reaction that is the predicate $Call$ which is in the form: $rea\_1\_14\_13\_a(\_, \_, \_, \_)$. This means that the next predicate $rand\_sw$ will perform a random choice for the switch $switch\_rea\_1\_14\_13\_a$. This random choice which is made by the $msw(switch\_rea\_1\_14\_13\_a, Value)$ built-in predicate of PRISM, determines the next step in the stochastic execution, since Value can be either $rea\_1\_14\_13\_a(yes, yes)$ or $rea\_1\_14\_13\_a(yes, no)$. In the first case it means the reaction has been randomly chosen to happen and the next step in the execution of the program which corresponds to the next reaction in the metabolic pathway is the call $produces(Rest, Delayed, Products)$. In the second case, the random choice $rea\_1\_14\_13\_a(yes, no)$ means that probabilistically the reaction did not occur and the sequence of the execution will be another, given by $produces(Metabolites, [Reaction|Delayed], Products)$ which will try randomly to choose the competing reaction catalyzed by the same enzyme $1.14.13.-$ that given the same input $c13624$ produces the compound $c13631$. If this reaction occurs, then the next reaction in the sequence will be one of the competing reactions with $c13631$ as input. The randomness stands in the stochastic execution of the reactions which

determines the random order of their execution for every possible possible set of available metabolites.

In order to learn the probabilities of the reactions we need a set of observations of the form $produces(Metabolites, Products)$. These observations that represent metabolomic data, are being intensively collected through available high throughput instruments and stored in particular metabolomics databases. In the next section, we show that from these observations, PRISM is able to accurately learn reaction probabilities through the built-in graphical EM algorithm. This represents the modeling task which consists in constructing a true model of the metabolic network. The reliability of the re-construction of the model from metabolomic data depends on the quality of the learning procedure of PRISM. The simulation task requires performing statistical inference in order to run the model and compute the probabilities of observed atoms.

V. Reconstructing the Stochastic Model

In large metabolic networks, not all metabolic paths are equally probable to happen. Some reactions happen more frequently than others under certain conditions. In many cases a metabolic path becomes inactive or useless under certain conditions if a certain intermediate reaction in the path cannot occur under those conditions. The focus of this paper is not on the conditions themselves (usually these are stoichiometric constraints) but on the way of simulating this stochastic variables. What is important in our context is that the conditions evolve in a stochastic fashion and this requires stochastic modeling and simulation.

Dealing with probabilities in the metabolic pathway means that through simulating various conditions we make possible a set of reactions instead of another, i.e. each set of conditions gives rise to a set of possible reactions that make some paths in the metabolic pathway biologically active and others biologically inactive under those conditions. The stochastic behavior is natural for metabolic networks where random variations in the physical and chemical variables are normal.

In order to perform simulation under various conditions, for each experiment we may randomly assign probabilities to reactions. These probabilities represent the switches probabilities in PRISM. Therefore, we have made possible for each single experiment a set of conditions under the form of assigned probabilities to reactions (as probabilities are randomly generated and some of them may be equal to zero or in the range $[0.9 - 0.999]$, among competing reactions one of them may not occur and this will cause some paths in the metabolic pathway to be inactive).

The model constructed this way represents the state of the biochemical environment under the given conditions at a certain moment in time. In laboratory conditions, when the reactions happen, what is caught by a high-throughput instrument is a set of metabolites concentrations and their changes. For instance, if a certain reaction happens then the concentration of the input metabolites decreases and that of the product compounds increases.

The change in time of the metabolites concentration is registered as a reaction, therefore catching all the sequential changes in concentration (this is actually performed intensively and accurately by current high-throughput technologies), means registering a sequence of reactions. These represent the available data to be used to re-construct the model dynamics from the data by learning the probabilistic parameters from observations.

VI. Stochastic Simulation of Metabolic Networks

Once the model has been learned from the data (the structure of the model is fixed, we only learn the model parameters) we can simulate the model by running the PRISM program through the call of the following goal $produces(InputMetabolites, Products)$ where $InputMetabolites$ is a bounded list with the input compounds and $Products$ is a logic variable that will be bounded to the list of product compounds yielded by the series of reactions.

In order to evaluate the validity of our approach we have proceeded as follows to simulate the model. For each experiment (each has a different set of conditions, i.e. probabilities of random switches that are randomly assigned) we randomly generate sequences of reactions by sampling from the previously defined model. This is made possible by the predicate *sample* of PRISM. Once the sequences have been generated, we launch the predicate *learn* of PRISM to learn the probability of each random switch from the sequences.

We then simulate the model over the sequences and query it for active paths with the predicate *hindsight(Goal)* where Goal is bounded to the top-goal $[InputMetabolites, Products]$. With this predicate we get the probabilities of all the sub-goals for the top-goal Goal. The insight of the simulation is that if any of these probabilities is equal to zero then the relative path of the sub-goal is biologically inactive under the given conditions. The relative path can be extracted by the predicate $prob f(SubGoal, ExplGraph)$ where $ExplGraph$ (explanation graph in PRISM) represents the explanation paths for $SubGoal$.

The accuracy of model reconstruction and simulation depends on the ability to faithfully recontruct the model from the sequences and on the accuracy of the statistical inference algorithm. In order to assess the accuracy of learning the probabilities of the reactions and of simulating the model, we adopt the following method:

We call the initial probability distribution (that represents the conditions) assigned to the clauses of the logic program the *true probability distribution* and call the M parameters the true parameters. Once the sequences have been randomly generated with this model, we forget the true parameters and replace their probabilities by uniformly distributed ones. When learning starts, PRISM learns M new parameters, that represent the learned reaction probabilities from the sequences. In order to assess the accuracy of the learned $P_i^{'}$ towards P_i we use the RMSE (Root Mean Square Error) for each single experiment with S sequences.

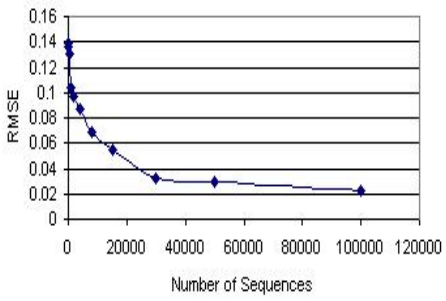

Fig. 2. Results for mean squared error.

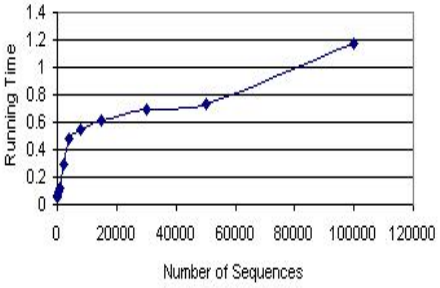

Fig. 3. Results for running time.

$$RMSE = \sqrt{\sum_{i=1}^{M} \frac{(P_i - P_i')^2}{M}} \quad (1)$$

With this measure we can evaluate the difference between the actual observations and the response predicted by the model. We perform learning and simulation with different experiments using a growing number *S* of sequences in order to evaluate how the number of sequences affects the accuracy and the learning time. A growing number of sequences means also a larger simulation process. This is a fundamental issue in simulating large metabolic networks where memory and time constraints are usually critical. Moreover, we also want to investigate large datasets of sequences in order to provide a robust methodology for real dimensions of data. Since real metabolomics datasets are in general highly voluminous, there is strong motivation to provide scalable solutions to modeling and simulation of large networks.

For each number of sequences *S*, we have performed 100 experiments where for each experiment the set of conditions is randomly generated as presented above.

As the results show, the accuracy of the model reconstruction and simulation increases as more data are available. PRISM has some very desirable features in terms of scalability. Due to tabulation techniques in PRISM, learning times increases reasonably as data dimension grows significantly. This is an essential advantage of PRISM as a modeling and simulation tool for application domains like metabolic networks where many parameters should be learned and then handled during the stochastic simulation of the model.

The accuracy of model learning can be evaluated as good for a number of sequences between 1000 and 15000 and excellent for a number of sequences greater than 15.000 considering that the range where probabilities fall is $[0, ..., 1]$ and the RMSE is under $0,05$. This means that the dynamics in the paths has been faithfully reconstructed from the sequences and thus the predicates $hindsight$ and $probf$ in PRISM faithfully reproduce and simulate the stochastic behavior of the pathway.

Moreover, from empirical observations, we noted that all the stochastic execution of the model performed by these two predicates reflected real biological paths that are supposed to have produced the sequences. For instance, we observed that in all cases where the probability of the reaction catalyzed by the enzyme $1.14.13.-$ (with input the compound $c13624$ and output $c13631$) was randomly assigned to be too low (from 0 to 0.05) by the conditions generation phase, then the path that involves one of the two next reactions, the one catalyzed by the enzyme $1.1.3.-$ and producing in output $c13633$, was not part of the stochastic simulation of the active path in the network.

Finally, we observed in all the experiments that by slightly changing the conditions, inactive paths became suddenly active and vice versa during simulation. This is important for two reasons: first it means that we can learn from sequences of reactions (data gethered in laboratory for real organisms) how conditions evolve in order to understand what changes them and what governs their randomness. This is made possibe by the ML capabilities of PRISM. Second, if we have the model, we can run stochastic simulations with different parameters and investigate the output in order to understand what are the final metabolites produced under the given conditions.

VII. Related Work

In this section we describe important related work regarding modeling and simulation of metabolic networks in general, and then in ML environments.

Important related work are presented in [18], [19] where graph-theory based approaches are used to find common or unique sub-graphs in different pathway graphs to understand better why and how pathways differ or are similar. However, these works do not deal with automatic reconstruction or simulation of the model, but try to understand similarities between different metabolic networks.

Other successful approaches focus on text mining for metabolic pathways [20]. These methods have been applied to the voluminous literature on metabolic pathways to discover knowledge about the structure of the pathways. Text mining techniques focus on the structure building process trying to identify, in the accumulated experience about metabolic pathways, significant structural properties. However, what these methods usually produce is just the structure of the network which gives only a static view of the real stochastic biological environment.

Other approaches attempt to only randomly simulate biochemical processes such [21] or [22]. These are powerful tools to model the dynamic nature of cells for simulation purposes but lack ML abilities to infer dynamics from observations. Only simulating the model is not sufficient for a thorough understanding of metabolic networks. Most of their nature and behavior remains hidden or unobserved. Simulation helps to understand the behavior of the observed part of the model, while ML helps to uncover the unobserved part.

From a learning point of view, the most important related work is that in [23] where a probabilistic relational formalism is used for modeling metabolic networks. The PRISM program we have presented here is syntactically quite similar to the logic program in [23], but is semantically different in the way probability distributions are defined. Stochastic Logic Programs (SLPs) [24], used in [23], assign probabilities to clauses and define probability distributions on Prolog proof trees, while PRISM programs are based on the distributional semantics [17] and assign probabilities to atoms as we explained in Section 3. Interesting contribution in the field is also represented by the work in [25].

VIII. Conclusion and Future Work

We have applied the symbolic-statistical machine learning framework PRISM to the problem of modeling and simulation of metabolic pathways. We have shown through experiments the feasibility of learning reaction probabilities from metabolomics data and performing stochastic simulations of the model. The power of the proposed approach stands in the description language that allows to model relations and in the ability to model stochastic dynamics in a robust manner. Simulating the model helps to understand the behavior of the observed part of the model, while machine learning helps to uncover the unobserved part. We have shown how machine learning can be used to reconstruct the metabolic network dynamics from the final outcomes observed, and how statistical inference methods in a joint framework can be used for stochastic simulations of the model.

Although we have been able to reconstruct the model from the sequences of reactions, there is still much to do to complete the real picture of a biochemical network. First, we have not taken into consideration stoichiometric constraints which express quantitative relationships of the reactants and products in chemical reactions. Adding these constraints to our approach will help reproduce better models.

Another direction for future work regards using other sources of data for a better reconstruction of the model. Considering multiple sources of data can lead to better models in modeling metabolic pathways [26]. Using PRISM to achieve this is straightforward because relational problems can be easily modeled due to the logic-based nature of PRISM. Another important challenge is to construct the model from incomplete raw metabolomic data. EM algorithms [27] are the state-of-the art for learning in the presence of missing data and since the graphical EM algorithm [16] used in PRISM, is a version of this class of learning algorithms, we believe that this will help in dealing with incomplete real datasets. In addition, in this paper we have considered a medium-sized metabolic pathway. For future work we intend to model very large metabolic pathways.

References

[1] G. G. Harrigan and R. e. Goodacre, *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis.* Kluwer Academic Publishers, Boston, 2003.

[2] S. G. Oliver, M. K. Winson, D. B. Kell, and F. Baganz, "Systematic functional analysis of the yeast genome." *Trends Biotechnol.*, vol. 16(10), pp. 373–378, 1998.

[3] H. Kitano, *Foundations of Systems Biology.* MIT Press, 2001.

[4] W. Weckwerth, "Metabolomics in systems biology." *Annu. Rev. Plant Biol.*, vol. 54, pp. 669–689, 2003.

[5] A. Kriete and R. Eils, *Computational Systems Biology.* Elsevier - Academic Press, 2005.

[6] L. Getoor and B. Taskar, *Introduction to Statistical Relational Learning*. MIT, 2007.

[7] F. Bacchus, *Representing and Reasoning with Probabilistic Knowledge*. Cambridge, MA: MIT Press, 1990.

[8] J. Halpern, "An analysis of first-order logics of probability," *Artificial Intelligence*, vol. 46, pp. 311–350, 1990.

[9] N. Nilsson, "Probabilistic logic," *Artificial Intelligence*, vol. 28, pp. 71–87, 1986.

[10] J. S. Wellman, M. Breese and R. P. Goldman, "From knowledge bases to decision models," *Knowledge Engineering Review*, vol. 7, 1992.

[11] L. Getoor, N. Friedman, D. Koller, and B. Taskar, "Learning probabilistic models of link structure," *JMLR*, vol. 3, pp. 679–707, 2002.

[12] S. Natarajan, P. Tadepalli, E. Altendorf, T. G. Dietterich, A. Fern, and A. C. Restificar, "Learning first-order probabilistic models with combining rules," in *ICML*, 2005, pp. 609–616.

[13] J. Neville and D. Jensen, "Dependency networks for relational data," in *Proc. 4th ICDM*. IEEE Computer Society, 2004, pp. 170–177.

[14] M. Jaeger, "Parameter learning for relational bayesian networks," in *ICML*, 2007, pp. 369–376.

[15] T. Sato and Y. Kameya, "Prism: A symbolic-statistical modeling language," in *Proc. of the 15th IJCAI*. Nagoya, Japan, 1997, pp. 1330–1335.

[16] ——, "Parameter learning of logic programs for symbolic-statistical modeling," *JAIR*, vol. 15, pp. 391–454, 2001.

[17] T. Sato, "A statistical learning method for logic programs with distribution semantics." in *In Proc. 12th ICLP*, 1995, pp. 715–729.

[18] M. Koyuturk, A. Grama, and W. Szpankowski, "An efficient algorithm for detecting frequent subgraphs in biological networks," in *In Proc. 12th ISMB*, 2004, pp. 200–207.

[19] C. You, L. Holder, and J. Cook, "Application of graph-based data mining to metabolic pathways." in *Workshop on Data Mining in Bioinformatics, ICDM,*, 2006.

[20] R. Hoffmann, M. Krallinger, E. Andres, J. Tamames, C. Blaschke, and A. Valencia, "Text mining for metabolic pathways, signaling cascades, and protein networks." *Sci STKE 283*, vol. 21, 2005.

[21] N. Le Novre and T. S. Shimizu, "Stochsim: modelling of stochastic biomolecular processes." *Bioinformatics*, vol. 17, pp. 575–576, 2001.

[22] G. M. Klamt S, Stelling J and G. ED., "Fluxanalyzer: exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps." *Bioinformatics*, vol. 19(2), pp. 261–269, 2003.

[23] N. Angelopoulos and S. Muggleton, "Machine learning metabolic pathway descriptions using a probabilistic relational representation." *ETAI*, vol. 6, 2002.

[24] S. Muggleton, "Stochastic logic programs," in *In L. De Raedt (Ed.), Advances in inductive logic programming*. IOS Press, Amsterdam, 1996.

[25] A. Tamaddoni-Nezhad, R. Chaleil, A. Kakas, and S. Muggleton, "Abduction and induction for learning models of inhibition in metabolic networks," in *In Proc. of the 4th ICMLA*, 2005.

[26] O. Fiehn, "Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks." *Comp. Funct. Genomics*, vol. 2(3), pp. 155–168, 2001.

[27] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm." *Royal Statistical Society*, vol. B39(1), pp. 1–38, 1977.