Analyzing Phylogenetic Trees with Timed and Probabilistic Model Checking: The Lactose Persistence Case Study

José Ignacio Requeno[1,*] and José Manuel Colom[1]

[1]Department of Computer Science and Systems Engineering (DIIS), Universidad de Zaragoza, C/ María de Luna 1, 50018 Zaragoza, Spain

Summary

Model checking is a generic verification technique that allows the phylogeneticist to focus on models and specifications instead of on implementation issues. Phylogenetic trees are considered as transition systems over which we interrogate phylogenetic questions written as formulas of temporal logic. Nonetheless, standard logics become insufficient for certain practices of phylogenetic analysis since they do not allow the inclusion of explicit time and probabilities. The aim of this paper is to extend the application of model checking techniques beyond qualitative phylogenetic properties and adapt the existing logical extensions and tools to the field of phylogeny. The introduction of time and probabilities in phylogenetic specifications is motivated by the study of a real example: the analysis of the ratio of lactose intolerance in some populations and the date of appearance of this phenotype.

1 Introduction

A phylogenetic tree is a description of the evolution process which is discovered via molecular sequencing data and morphological data matrices [1]. Computer science tools have upgraded the capabilities of biologists for their construction as well as for extracting and analyzing the implicit biological messages embedded on them [2, 3]. Nowadays, more and more applications rely on the existence of a support phylogenetic tree for the confirmation of biological hypothesis that are valuable for the scientific community. In this sense, we use a phylogenetic tree for testing biological hypothesis. Indirectly, the evaluation results help to feedback the phylogeny and increase its quality.

The wide range of heterogeneous methods and tools used by biologists for the analysis of phylogenies and verification tasks recommends the possibility of researching in a generic framework for heterogeneous hypothesis testing over trees. One of the most relevant challenges is the introduction of a generic framework for heterogeneous hypothesis verification over trees.

To this end, we have explored the features of model checking, a paradigm stemming from computer science based on temporal logics which has been successfully applied in industry for system modeling and verification [4]. Model checking has been proposed as a generic unifying framework that allows the phylogeneticist to focus on tree structures, biological properties and symbolic manipulation of phylogenies described using temporal logic, instead of on implementation issues concerned with verification algorithms [5, 6]. The basic principles allowing

*To whom correspondence should be addressed. Email: nrequeno@unizar.es

the use of this framework in the context of phylogeny is the interpretation of the phylogenetic tree as a transition system (in the context of computer science) representing a computational model of the evolution process. Now, the next step is to formulate the property that we desire to investigate in a given phylogeny using a formal language as, for example, the *Computational Tree Logic* (CTL [7]). Besides, model checking allows us to uncouple software tools from the definition of properties and it hides the underlying implementation technology [7]. Even, phylogenetic properties can be exported and evaluated in other structures (i.e., trees or networks) with minimum effort.

Standard logics such as CTL allow to express biological properties referred to the structure of the tree or the composition of the DNA sequence [5]. Some of the phylogenetic hypothesis are extended with explicit time and probabilities in the specifications and models. A small but representative portion of these researches combine phylogenetic trees (constructed via the mentioned tools using the information of the genome) with geographical or phenotypical data in order to trace the epoch of human migrations or the distribution of endemic diseases [8, 9].

Although time and probabilities are not directly supported by qualitative logics, they can be considered doing several modifications to the traditional model checking framework. To do that, we need to complete the tree with quantitative labels that include extra information beyond the original propositional information of the states [10]. These numerical annotations of the tree changes with every particular study, but they are commonly divided in a) *timed* or distance information, b) *probabilistic* information and c) *raw quantitative* information. The first two ones are mainly related to the labeling of the tree branches [11]. The last one is more general and involves numerical values and comparisons in the atomic propositions. On the other hand, we must define a formal language that extends the CTL logic in order to capture the new numerical information.

Hence, the aim of this paper is to analyze the requirements of non-qualitative phylogenetic properties, adapt the existing probabilistic and timed extensions to temporal logics and use the associated model checking tools in the field of phylogeny. The introduction of time and probabilities in phylogenetic specifications is motivated by means of a real example: the study of the lactose intolerance. Here, we develop and extend the work presented in [12]. The paper is divided in seven sections. After this introduction, Section 2 explains a real example that motivates the definition of the discrete-time probabilistic logic and structures of Section 3. Next, Section 4 details an algorithm of model checking that computes the probabilities and verifies the logical specifications over the phylogenetic tree. Later, Section 5 shows the experimentation with a temporal and probabilistic model checking tool. Section 6 applies the previous concepts and techniques to the particular study of the lactose tolerance in Tibetan populations. Finally, Section 7 briefs the conclusions and draws the future work.

2 Motivation and Presentation of the Case of Study

A phylogenetic tree is a directed graph that offers a realistic model of aggregated evolution in which each vertex represents an inferred state of the evolution characterized by biological sequences (e.g., DNA) [5]. The phylogenies are occasionally enriched with time labels or weights

in the edges. This knowledge is useful for learning complex properties about the evolution, for instance, the estimation of the temporal point of divergence between species [13] or the diaspora of human populations [9]. The extension of the phylogenetic properties in [5, Table 1] with time and probabilities increases the expressivity of biological hypothesis.

We will study the following example. The lactose intolerance in adults is a chronic disease caused by the inhibition of the lactase gene after the breastfeeding and childhood. The inability for processing the milk and its derivations is not homogeneously distributed in the human population. While in some African pastoralist groups of North/East Africa and the northern cultures of Europe their stock breeding tradition and diet motivated an evolutionary adaptation to digest the milk ($> 70\%$ of tolerance), the percentage of acceptation decreases in the rest of areas and ethnic groups [14, 8]. In addition, the phenotype in Europe and Africa appeared at a different epoch and the point mutations that regulate the activation of the lactase persistence are disparate [15, 16]. Some illustrative questions that we desire to ask to the phylogeny, and that are expressed below, require the addition of time to the branches of a population tree. The time allows the estimation of the divergence points between individuals or mutations, while the probability of the lactose persistence in different zones is calculated through the study of the distribution of the point mutations that regulate the phenotype. The questions are:

- What is the rate of lactase persistence in a population? i.e., do their members define a characteristic haplogroup? and in that case,
- Which polymorphism, among the multiple activators and inhibitors of the lactase gene, is the most frequent over there? and finally,
- When did this phenotype approximately start to predominate? i.e., does this date mark a major event in the diet, culture or migration of that population?

These properties ask about the time (dates) and probabilities (frequencies/rates) over a phylogeny. They can be solved including time and probabilities in the model and specifications. In order to answer questions about the time, we need to scale the edges of the tree with weights representing the amount of nucleotide changes according to a mutation clock of reference. The notion of time introduced here embeds the concept of evolutionary or chronological clock. On the other hand, the probability expressing the ratio of states satisfying a property is obtained by traversing the tree and calculating the phenotypical density.

Thus, the deductive process that answers the queries also needs the manipulation of quantitative information. To this end, we must introduce a logic, a transition system and a model checking algorithm capable of expressing and managing these kind of questions. Firstly, we introduce the logic. The logic presented in Section 3 covers two aspects: the qualitative properties presented in our previous works [5], and the phylogenetic hypothesis studied in this paper.

3 Discrete-time Probabilistic Logic and Structure

In this section we are considering a phylogenetic tree enriched with numerical information that tells the probability of selecting a branch descending from an internal node. Therefore, we can

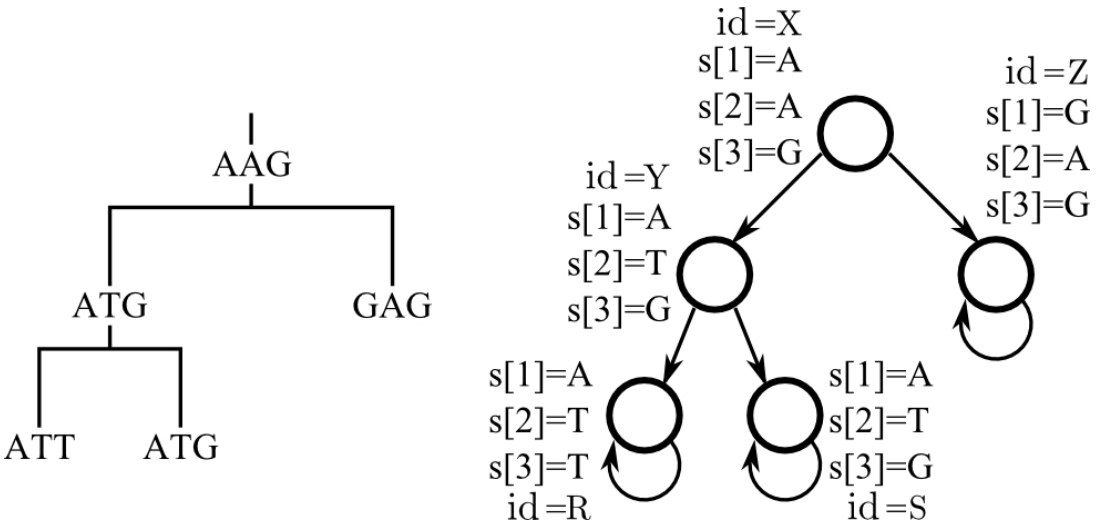

Figure 1: Translation from a phylogenetic tree to a Kripke structure.

analyze properties like: what is the probability of reaching a set of states of the tree from the root? The logic and data structure defined here settle the basis for future updates and extensions for continuous-time systems [17]. Stochastic systems generally use Markov chains as the underlying data structure that provides semantics to the verification process. Discrete-time Markov chains capture the essentials of probabilities between states of the tree and implicitly associates an unit time step to every transition of the system.

Definition 1 (Discrete-time Markov Chain) *A discrete-time Markov chain is a finite transition system represented by a tuple* $M = (S, S_0, \mathbf{P}, L)$*, where:*

- S *is a finite set of states,*
- $S_0 \subseteq S$ *is the set of initial states,*
- $\mathbf{P} : S \times S \to [0, 1]$ *is the transition probability matrix that indicates the probability of moving from a state* s_i *to a state* s_j *satisfying* $\Sigma_{s_j \in S} P(s_i, s_j) = 1$*, and*
- $L : S \to 2^{AP}$ *is the labeling function that associates each state with the subset of atomic propositions (AP) that are true of it.*

A phylogenetic tree is assimilated to a discrete-time Markov chain making the corresponding association of states to the definition of phylogeny ([5, Definition 3]). The leaves are labeled with the genome information of the population or species they represent, plus additional data when necessary. Each branch of the phylogeny is labelled with an element $\mathbf{P}(s_i, s_{i+1}) > 0$ of the transition probability matrix. This value gives the probability of moving from state s_i to state s_{i+1} in one time step. The self-loops of the terminal leaves in the Kripke structure of a branching-time phylogeny are represented by a single transition going back to the same state with probability 1 in the transition probability matrix (Figure 1).

For any set of infinite paths Π starting in the initial state s_0, the subset $\Pi(\pi_n)$ selects the paths $\pi \in \Pi$ whose prefix equals to the finite sequence $\pi_n = s_0 s_1 s_2 \dots s_n$ of length $n+1$ states. The set of infinite sequences sharing the prefix π_n has probability $Pr(\Pi(\pi_n)) = \mathbf{P}_\Pi(\pi_n)$. The probability $\mathbf{P}_\Pi(\pi_n)$ is calculated as the product of probabilities for each intermediate transition, except for paths with unitary length in which case $n = 0$, $\pi_0 = s_0$ and $\mathbf{P}_\Pi(\pi_0) = \mathbf{P}_\Pi(s_0) = 1$. That is, $\mathbf{P}_\Pi(\pi_n) = \mathbf{P}(s_0, s_1) \cdot \mathbf{P}(s_1, s_2) \cdot \ldots \cdot \mathbf{P}(s_{n-1}, s_n)$.

Bayesian model checking methods allow for analyzing stochastic systems [18]. Probabilistic CTL (PCTL) [19, 20, 7] helps to formulate conditions on a discrete-time Markov chain. The properties are referred to state formulas (ϕ) or path formulas (Φ). Besides, PCTL allows enriched queries such as $\mathbb{P}_{\sim\lambda}(\Phi)$. Given an initial state s and a comparison $\sim \in \{<, \leq, =, \geq, >\}$, the operator $\mathbb{P}_{\sim\lambda}(\Phi)$ returns true if the probability for a set of paths satisfying Φ is $\sim \lambda$, with $\lambda \in [0, 1]$.

Definition 2 (Probabilistic Computation Tree Logic) *A temporal logic formula ϕ is defined by the following grammar, where $p \in AP$, $k \in \mathbb{N} \cup \{\infty\}$:*

$$\begin{aligned} \phi &::= true \mid p \mid \neg\phi \mid \phi \vee \phi \mid \mathbb{P}_{\sim\lambda}[\Phi] \\ \Phi &::= \mathbf{X}\phi \mid [\phi \mathbf{U}_{\leq k} \phi] \end{aligned} \tag{1}$$

The formulas are checked against a structure M considering all infinite paths $\pi \in \Pi$ from a certain state s_0. Notice that $M, s_0 \vDash \phi$ means that s_0 satisfies ϕ. The semantics of well-formed formulas is as follows (let $\pi = s_0 s_1 s_2 \dots$):

- $M, s_0 \vDash p \Leftrightarrow p \in L(s_0)$,
- $M, s_0 \vDash \neg\phi \Leftrightarrow M, s_0 \nvDash \phi$,
- $M, s_0 \vDash \phi \vee \psi \Leftrightarrow M, s_0 \vDash \phi$ *or* $M, s_0 \vDash \psi$,
- $M, s_0 \vDash \mathbb{P}_{\sim\lambda}[\Phi] \Leftrightarrow Prob(M, s_0, \Phi) \sim \lambda$,

The calculation of the probability $Prob(M, s_0, \Phi)$ requires the identification of the infinite paths π satisfying the path formula $M, \pi \vDash \Phi$:

- $M, \pi \vDash \mathbf{X}\phi \Leftrightarrow M, s_1 \vDash \phi$
- $M, \pi \vDash [\phi \mathbf{U}_{\leq k} \psi] \Leftrightarrow \exists 0 \leq i \leq k, \forall 0 \leq j \leq i : (M, s_i \vDash \psi) \wedge (M, s_j \vDash \phi)$

This set, $\{\pi \in \Pi \mid M, \pi \vDash \Phi\}$, can be obtained by the union of finitely many pairwise disjoint subsets $\Pi(\pi_n)$ by [17, Definition 3], each one characterized by the finite prefix π_n of all infinite sequences of the set.

Therefore, $Prob(M, s_0, \Phi) = Pr\{\pi \in \Pi \mid M, \pi \vDash \Phi\} = \Sigma_{\pi_n} Pr(\Pi(\pi_n))$ computes the probability as the summation of probabilities in all possible prefixes π_n by [17, Theorem 1].

The logic supports timed transitions in the $\mathbf{U}$ operator. The notion of time in a Markov chain falls within the concept of state distances. Each state transition of the discrete-time Markov chain involves an unit time step. A mapping between the chronological time and state distances allows the inference of the evolutionary speed in the branches of the phylogenetic tree. The computation of time and probabilities are embedded in the model checking algorithm. Timed variants of the modal operators $\mathbf{F}$ and $\mathbf{G}$ are obtained via $\mathbf{U}$ as $\mathbf{F}_{\sim c}\phi = true\ \mathbf{U}_{\sim c}\phi$ and $\mathbf{G}_{\sim c}\phi = \neg\mathbf{F}_{\sim c}\neg\phi$. Instead of writing the intervals explicitly, sometimes they are abbreviated with comparisons. For example, $\mathbb{P}_{\leq 0.5}\left[\Phi\right]$ denotes $\mathbb{P}_{[0,0.5]}\left[\Phi\right]$. Almost every property expressed in CTL is compatible with PCTL, which guarantees that the same logic PCTL will be able to handle a great number of formulas.

By now, we can translate the questions presented in the motivation example of lactose into the PCTL syntax. In a phylogenetic tree, the tips correspond to individuals of disjoint populations whose states are tagged with their DNA and a boolean indicating if they are lactose (in)tolerant. The internal nodes of the inferred ancestors are labeled with their estimated DNA sequence and lactose phenotype as well. The following equation asks if there exists an ancestor ($\mathbb{P}_{>0}$) at distance 3 or above from the initial state ($\mathbf{F}_{\geq 3}$) that is the root of a population with lactase persistence over 70% ($\mathbb{P}_{\geq 0.7}\left[\mathbf{F}_{\geq 0}\, lactose\_tolerant\right]$). The members of a population, including the leaves and internal nodes, are reached by $\mathbf{F}_{\geq 0}$.

$$\mathbb{P}_{>0}\left[\mathbf{F}_{\geq 3}\ \left(\mathbb{P}_{\geq 0.7}\left[\mathbf{F}_{\geq 0}\, lactose\_tolerant\right]\right)\right] \tag{2}$$

The outer restriction $\mathbb{P}_{>0}\left[\mathbf{F}_{\geq 3}\right]$ corresponds to the question 3 of the motivation. It searches for an internal node from which the phenotype starts to be predominant after a certain date since the phylogenetic root. The inner formula $\mathbb{P}_{\geq 0.7}\left[\mathbf{F}_{\geq 0}\, lactose\_tolerant\right]$ answers the question 1 about the rate of lactase persistence in a population. Finally, the addition of a genetic marker in this place inside the $\mathbb{P}_{\geq 0.7}$ equation helps to investigate the relation between a polymorphism and phenotype (question 2). The evaluation of the formulas needs the algorithm introduced in the next section.

4 Algorithm for PCTL Model Checking

The algorithm for managing and solving PCTL formulas in stochastic systems is mainly identical to that of classic model checking except for the resolution of $\mathbb{P}_{\sim\lambda}[\Phi]$, i.e., the $\mathbf{X}$ and $\mathbf{U}$ operators with probability thresholds. In short, the recursive algorithm of model checking incorporates the new sentence [7]:

$$Sat(\mathbb{P}_{\sim\lambda}\left[\Phi\right]) = \{s \in S \mid Prob(M, s, \Phi) \sim \lambda\}$$

$\mathbb{P}_{\sim\lambda}[\mathbf{X}\phi]$ **formula.** In PCTL, the probability of satisfying the next operator requires the probabilities of the immediate transitions from s. It is resolved by $Prob(M, s, \mathbf{X}\phi) = \Sigma_{s' \in Sat(\phi)}\mathbf{P}(s, s')$.

$\mathbb{P}_{\sim\lambda}[\psi\mathbf{U}_{\leq k}\phi]$ **formula.** The computation of the probability for the until operator depends on the value of k. For $k \in \mathbb{N}$, then $Prob(M, s, \psi\mathbf{U}_{\leq k}\phi)$ is:

$$\begin{cases} 1 & \text{if } s \in Sat(\phi) \\ 0 & \text{if } k = 0 \text{ or } s \in Sat(\neg\phi \wedge \neg\psi) \\ \Sigma_{s' \in S} \mathbf{P}(s, s') \cdot Prob(M, s', \psi \mathbf{U}_{\leq k-1} \phi) & \text{otherwise} \end{cases}$$

When $k = \infty$, the until operator is analogous to the original until operator of CTL with semantics of infinite paths. That is, $Prob(M, s, \psi \mathbf{U}_{\leq \infty} \phi)$ can be rewritten as $Prob(M, s, \psi \mathbf{U} \phi)$ and it equals to:

$$\begin{cases} 1 & \text{if } s \in Sat(\phi) \\ 0 & \text{if } k = 0 \text{ or } s \in Sat(\neg\phi \wedge \neg\psi) \\ \Sigma_{s' \in S} \mathbf{P}(s, s') \cdot Prob(M, s', \psi \mathbf{U} \phi) & \text{otherwise} \end{cases}$$

The time complexity of verifying a PCTL formula ϕ against a discrete-time Markov chain is linear in $|\phi|$ and polynomial in the size of S, with $|\phi|$ the number of logical connectives and temporal operators of the formula. In short, $\Theta(poly|S|) * k_{max} * |\phi|)$ where k_{max} is the maximal step bound of a path subformula $\psi_1 \mathbf{U}_{\leq k} \psi_2$ of ϕ, with $k_{max} = 1$ if it doesn't contain any $\mathbf{U}_{\leq k}$ subformula.

In sum, most of the investigations related with quantitative information are a prolongation of the phylogenetic properties analyzed with qualitative logics. Consequently, the inconveniences presented in [5] would appear even more dramatically now. Some of the techniques introduced for the optimization and scalability in classic model checking environments also work for this context [21]. For example, the incorporation of external data bases for storing the labeling of the states is compatible with any kind of atomic propositions, quantitative or not. Besides, the inclusion of databases allows the storage of previous results that can be reused in the future without reevaluation. The vertical and horizontal partitioning of the database table adds an extra dimension of parallelism. Furthermore, the summations in the computation of probability paths, specially with the $\mathbf{U}_{\leq k}$ operator, could be executed in parallel.

5 Model Checking Tools and Experimentation

There exists a considerable diversity of model checking tools with different performances [22] and qualities (Table 1). PRISM [23] is a generic model checking tool capable of handling probabilistic and timed specifications over Markov chains. Among its basic conceptions, PRISM checks if the probability of reaching a set of satisfiable states is up or below a predefined threshold. Although the real performance depends on the particular structure of the model and specifications, PRISM offers Java portability, a powerful syntax for handling time and probabilities in models and specifications, and a good scientific community support. Besides, it is open source, which allows the modification and optimization of its code. Hence, PRISM becomes one of the best options for evaluating temporal and probabilistic properties.

The model checking tool requires two input files for the verification process: a first file with the description of the model, and a second file with the specification of the properties. A description of the phylogenetic tree in PRISM syntax musts be provided by the user as first

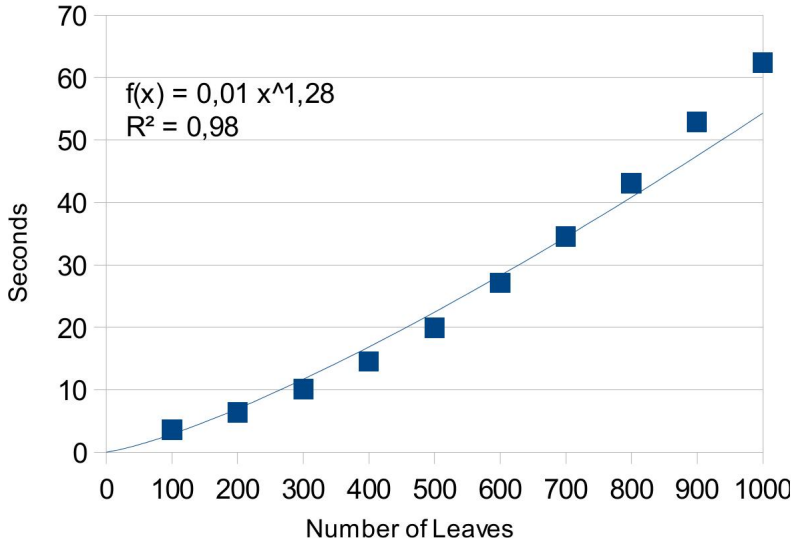

Figure 2: Time required for the verification of a set of probabilistic formulas with respect to the number of tips in the phylogeny.

input for the model checker. The description details the Kripke structure with the genome as atomic proposition. To this end, we precompute a sequence alignment and a phylogenetic tree.

The PRISM codification of the phylogeny follows the same idea presented for NuSMV in [5]. Figure 3 shows the implementation of the branching-time phylogenetic tree of Figure 1 in PRISM code. The main module describes the topology of the evolutionary tree, where the names of the tree nodes (taxa) are defined numerically $(1, \ldots, 5)$. PRISM distinguishes between the current state and the next state using the quotation mark (’). The second part of the description consists of a function returning the DNA string associated to each node, which evolves in synchronization with the tree skeleton ([id1] tags in PRISM). The translation of the phylogenetic tree to the PRISM syntax has been performed automatically by a BioPerl script [32]. The script can be upgraded in order to include extra features such as the generation of multiple instances of phylogenetic trees, bootstrapping and so on.

The data set used for this experimentation is synthetic. With this data set, we try to cover the spectrum of small phylogenies and analyze the cost of the evaluation of the lactose property over there. We have created random phylogenetic trees of up to 1000 tips with a Yule speciation model [33]. For each tree size (number of tips), we have generated ten random trees and calculated the harmonic mean time. The DNA sequences have 50 bases with an homogeneous distribution of nucleotides.

We have evaluated the lactose formula introduced in the motivation but enriched for the detection of polymorphisms. The underlying objective consists of the identification of a correlated evolution between lactose tolerance and genomic patterns. Other studies such as [8] use cultural information for discovering this coevolution and the influence of a milk-based diet. The utilization of phylogenetic comparative methods and regression techniques establishes the essentials of this approach.

The probability threshold of the internal $\mathbb{P}_{\geq x}\left[\mathbf{F}_{\geq 0}\, seq[i] = j\right]$ ranges from $x \in [0.1, 0.9]$, with $i \in [1, 50]$ the position where we search for the polymorphism and j a certain nucleotide. The

Table 1: List of some available model checkers.

Name	Main feature	Properties Language	Platform
NuSMV [24]	OpenSource	LTL, CTL, RTCTL, PSL	Windows, Unix, MacOS
PROD [25]	On-the-fly Model Checker	CTL	Linux
SPIN [26]	Generic Model Checker	LTL	Windows, Unix
DiViNe Tool [27]	Distributed Multicore Model Checker	LTL	Unix
Eddy Murphi [28]	Distributed Multicore Model Checker	Assertions	Unix
PVeSta [29]	Statistical Parallel & Multicore Model Checker	PCTL	Windows, Linux, MacOS
PRISM [23]	Probabilistic & Quantitative Logics	PCTL, CSL, LTL, PCTL*	Windows, Linux, MacOS
UPPAAL [30]	Commercial Software for Real Time Systems	TCTL	Windows, Linux
TLQSolver [31]	Temporal Logic Query Checker for Mining Properties	Query CTL	Linux

Figure 2 plots the time required for the computation of 50×9 formulas corresponding to the expansion of i and x for all the columns of the alignment and probability bounds. All tests have been run on a Intel Core 2 Duo E6750 @ 2.66 GHz with 8 GB RAM and Linux.

PRISM performs well for the verification of the lactose formulas in small phylogenies as it follows a polynomial trend in time with respect to the number of tips. However, it requires the integration of new technologies and solutions to scale for larger phylogenetic trees and specifications. In fact, we desire to find the values of x, i and j for which the verification of the equation returns true. The definition of patterns is a common procedure, which intuitively leads to parametric model checking [34]. Nonetheless, mining for knowledge without prior information requires a more or less thorough exploration of the structure, which can be combinatorial in some or all of its dimensions. Although inherently parallel, the exhaustive inspection of potential solutions involves the test of large sets of formulas and an intensive use of the topology and information of each state. The application of parametric model checking for model exploration is a future extension that will increase the potential of our framework.

```
// Markow decision process
mdp

// Nucleotids
const int a = 1; ... const int t = 4;

module TREE
    // Initial state
    id: [1..5] init 1;

    // Root successors
    [id2] id=1 -> (id'=2); [id3] id=1 -> (id'=3);

    // Left clade successors
    [id4] id=2 -> (id'=4); [id5] id=2 -> (id'=5);

    // Self loops
    [] id=3 -> (id'=3); ...

    // In case of continious/discrete time markov chains
    // (ctmc/dtmc) or probabilistic timed automata (pta), 'x'
    // and 'y' will indicate probabilities or branch lengths
    // [] id=1 -> x:(id'=2)+y:(id'=3);
endmodule

module SEQUENCES
    // Root sequence
    x1 : [a..t] init a; x2 : [a..t] init a; x3 : [a..t] init g;

    // Sequences
    [id2] true -> (x1'= a)&(x2'= t)&(x3'= g);
    [id3] true -> (x1'= g)&(x2'= a)&(x3'= g);
    ...
endmodule
```

Figure 3: Mapping of the phylogenetic tree of Figure 1 in PRISM.

6 Case of Study: Polymorphisms Regulating the Lactose Tolerance in Tibetan Populations

In previous sections, we presented the extensions of the model checking framework for manipulating time and probability that provides the conceptual base for this study. Additionally, the experimentation with a synthetic data set shows the feasibility of our approach. Now, in this section we apply the framework introduced in this paper to a particular example: the analysis of single nucleotide polymorphisms (SNPs) regulating the lactose tolerance in Tibetan populations [35].

Here, we replicate the experimentation of [35] in order to emphasize that our proposal can be applied to a real case of study with similar results and performance, plus the inherent advantages of model checking. The DNA examined correspond to 495 Tibetan individuals. The sequences cover a region of 321 bp (position -14044 to -13724 upstream LCT) including the four SNPs associated to the lactose regulation in European and some Central Asia populations (-13910C/T), and in some African and Middle Eastern populations (-13907C/G, -13915T/G and -14010G/C). We have inferred a phylogenetic tree using the accelerated bootstrapping option of RAxML [36] and 100 bootstraps. It is supposed to organize the individuals in subtrees by prefecture and country. This phylogeny is the model over which we will evaluate the formulas in temporal logic.

More in detail, we evaluate formulas with the pattern $\mathbb{P}_{\geq 0.8}\,[\mathbf{F}_{\leq 3}\, seq[i] = j]$ in every internal node of the tree and for each one of the three SNPs associated to the lactose persistence in Tibet. We search for a set of individuals clustered in the same subtree and sharing a nucleotide j in position i with frequency $\geq 80\%$. The internal nodes of the tree are easily identified symbolically with a proper labeling of the states. The execution of the complementary formula returns a counterexample with the common ancestors satisfying the equation.

Table 2 shows the formulas evaluated in PRISM and their verification results. The third equation returns `false` because it is the rarest SNP according to [35] (only one person has that polymorphism). The results for the rest of equations match with the previous study of [35]. All tests have been run on a Intel Core 2 Duo E6750 @ 2.66 GHz with 8 GB RAM and Linux. The time required for the initialization of PRISM and the verification of the lactose formulas in every internal node of the phylogeny raises up to 19min 10s. The use of `filters` extends easily the evaluation to every internal node.

Table 2: Formulas evaluated over the phylogenetic tree using PRISM and the output result of the verification.

`filter(exists, P>=0.8 [ F<=3 (x206 = A) ], internal);`	`true`
`filter(exists, P>=0.8 [ F<=3 (x138 = A) ], internal);`	`true`
`filter(exists, P>=0.8 [ F<=3 (x136 = T) ], internal);`	`false`

7 Conclusions

Model checking represents a generic framework to support the analysis of qualitative properties over phylogenies [5, 6]. The model checking approach that we have proposed in our previous works considers a phylogenetic tree as a *model* of a particular system endowing an evolutionary process in which biological properties expressed with formal logics can be *verified*. A property is considered as a claim over the tree and, then, the verification determines the truth value for this claim. The evaluation of biological specifications may return counterexamples that validate, nullify or refine the phylogeny as a plausible model of the evolution during the tree construction phase. For example, we can detect and discard trees having an internal node with a lethal mutation in the DNA (according to certain biochemical constraints) because it should not generate offspring.

Nevertheless, some properties are beyond the expressiveness of qualitative logics. In this paper, we suggest to complete our framework including a new kind of information. Mainly, we have motivated the extension of model checking for phylogenetic analysis using quantitative data. We have proposed the inclusion of time and probabilities in the branches of the tree because of its natural interpretation in the phylogeny. In particular, we have presented a phylogenetic example based on the lactose (in)tolerance that needs these kind of information. To this end, we have introduced an extended logic and data structure adapted for probabilities and time together with the algorithms and computations for managing them. Our first goal has been the increase of the logical capabilities for querying about the date of appearance and degree of distribution of mutations and phenotypes.

Next, we have experimented with synthetic data in order to prove the feasibility of our approach with existing probabilistic model checking tools. Besides, we have used PRISM for the evaluation of the lactose persistence in a real case of study with Tibetan populations. PRISM is a generic model checking tool that performs well for small phylogenies in polynomial time with respect to the number of tips. The tool is independent of the application domain: it automatically verifies any proposition expressed with temporal logic over a model of the system.

However, it requires the integration of new technologies and solutions to scale for larger phylogenies and specifications due to the particularities of the phylogenetic analysis. Some of the ideas introduced for improving the performance of standard CTL model checking can be applied here. For example, the integration of PRISM with an external database constitutes a natural extension for storing atomic propositions and partial evaluation results that can be reused in the future analysis without recomputing them. The distribution of the Markov chain structure and the paralellization of the formula verification are stronger problems that should be studied in the future.

This work opens the door for the review of bigger phylogenies with properties similar to the lactose persistence. The modularity of our framework allows the evaluation of hypothesis and the comparison of results for a set of phylogenetic trees by only changing the tree file (the specification of the property remains constant). Finally, the search for the valuations that verify a certain specification leads to an intensive exploration of the formula space or the solution of linear systems. The introduction of parametric model checking for the automatic discovery and mining of phylogenetic information outlines our future work. Further studies may also focus on the evaluation of (quantitative) functions over the phylogeny, taking into account the temporal labels in the tree for determining the truth value of a formula. For instance, the computation of MLE scores falls in this category.

Acknowledgements

This work was supported by the Spanish Ministry of Science and Innovation (MICINN) [Project TIN2011-27479-C04-01] and the Government of Aragon [B117/10].

References

[1] J. Felsenstein. *Inferring phylogenies*. Sinauer, 2003.

[2] Z. Yang and B. Rannala. Molecular phylogenetics: Principles and practice. *Nature Reviews Genetics*, 13(5):303–314, 2012.

[3] W. M. Fitch. Uses for evolutionary trees. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 349(1327):93–102, 1995.

[4] O. Grumberg and H. Veith. *25 years of model checking: History, achievements, perspectives*. Springer, 2008.

[5] J. I. Requeno, G. de Miguel Casado, R. Blanco and J. M. Colom. Temporal logics for phylogenetic analysis via model checking. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(4):1058–1070, 2013.

[6] J. I. Requeno. *Formal methods applied to the analysis of phylogenies: Phylogenetic Model Checking*. Ph.D. thesis, School of Engineering and Architecture, University of Zaragoza, 2014.

[7] C. Baier and J.-P. Katoen. *Principles of model checking*. The MIT Press, 2008.

[8] C. Holden and R. Mace. Phylogenetic analysis of the evolution of lactose digestion in adults. *Human Biology*, 81(5/6):597–619, 2009.

[9] L. L. Cavalli Sforza and M. W. Feldman. The application of molecular genetic approaches to the study of human evolution. *Nature Genetics*, 33:266–275, 2003.

[10] C. M. Zmasek and S. R. Eddy. ATV: Display and manipulation of annotated phylogenetic trees. *Bioinformatics*, 17(4):383–384, 2001.

[11] A. Rambaut. How to read a phylogenetic tree, 2013. URL `http://epidemic.bio.ed.ac.uk/how_to_read_a_phylogeny`.

[12] J. I. Requeno and J. M. Colom. Timed and probabilistic model checking over phylogenetic trees. In M. P. Rocha et al. (editors), *Proceedings 8th International Conference on Practical Applications of Computational Biology and Bioinformatics*, volume 294 of *Advances in Intelligent and Soft Computing*, pages 105–112. Springer, Berlin, 2014.

[13] T. G. Barraclough and S. Nee. Phylogenetics and speciation. *Trends in Ecology and Evolution*, 16(7):391–399, 2001.

[14] D. M. Swallow. Genetics of lactase persistence and lactose intolerance. *Annual Review of Genetics*, 37(1):197–219, 2003.

[15] S. A. Tishkoff, F. A. Reed, A. Ranciaro et al. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics*, 39(1):31–40, 2006.

[16] C. J. E. Ingram, C. A. Mulcare, Y. Itan, M. G. Thomas and D. M. Swallow. Lactose digestion and the evolutionary genetics of lactase persistence. *Human Genetics*, 124(6):579–591, 2009.

[17] M. Kwiatkowska, G. Norman and D. Parker. Stochastic model checking. In M. Bernardo and J. Hillston (editors), *7th International School on Formal Methods for Performance Evaluation*, volume 4486 of *LNCS*, pages 220–270. Springer, Berlin, 2007.

[18] S. K. Jha, E. M. Clarke, C. J. Langmead, A. Legay, A. Platzer and P. Zuliani. A Bayesian approach to model checking biological systems. In P. Degano and R. Gorrieri (editors), *Proceedings 7th International Conference on Computational Methods in Systems Biology*, volume 5688 of *LNCS*, pages 218–234. Springer, Berlin, 2009.

[19] R. Segala and N. Lynch. Probabilistic simulations for probabilistic processes. In B. Jonsson and J. Parrow (editors), *Proceedings 5th International Conference on Concurrency Theory*, volume 836 of *LNCS*, pages 481–496. Springer, Berlin, 1994.

[20] H. Hansson and B. Jonsson. A logic for reasoning about time and reliability. *Formal Aspects of Computing*, 6(5):512–535, 1994.

[21] J. I. Requeno and J. M. Colom. Model checking software for phylogenetic trees using distribution and database methods. *Journal of Integrative Bioinformatics*, 10(3):229–233, 2013.

[22] D. N. Jansen, J.-P. Katoen, M. Oldenkamp, M. Stoelinga and I. Zapreev. How fast and fat is your probabilistic model checker? an experimental performance comparison. In K. Yorav (editor), *Proceedings 3rd International Haifa Verification Conference on Hardware and Software, Verification and Testing*, volume 4899 of *LNCS*, pages 69–85. Springer, Berlin, 2008.

[23] Marta Kwiatkowska, Gethin Norman and David Parker. PRISM 4.0: Verification of probabilistic real-time systems. In G. Gopalakrishnan and S. Qadeer (editors), *Proceedings 23rd International Conference on Computer Aided Verification*, volume 6806 of *LNCS*, pages 585–591. Springer, Berlin, 2011.

[24] A. Cimatti, E. M. Clarke, E. Giunchiglia, F. Giunchiglia, M. Pistore, M. Roveri, R. Sebastiani and A. Tacchella. NuSMV 2: An opensource tool for symbolic model checking. In E. Brinksma and K. Larsen (editors), *Proceedings 14th International Conference on Computer Aided Verification*, volume 2404 of *LNCS*, pages 241–268. Springer, Berlin, 2002.

[25] K. Varpaaniemi, J. Halme, K. Hiekkanen and T. Pyssysalo. PROD reference manual. Technical report, Helsinki University of Technology, Digital Systems Laboratory, Espoo, Finland, 1995.

[26] G. J. Holzmann. The model checker SPIN. *IEEE Transactions on Software Engineering*, 23(5):279–295, 1997.

[27] J. Barnat, L. Brim, M. Ceska and P. Rockai. DiVinE: Parallel distributed model checker. In *Proceedings 9th International Workshop on Parallel and Distributed Methods in Verification and 2nd International Workshop on High Performance Computational Systems Biology*, pages 4–7. IEEE, 2010.

[28] I. Melatti, R. Palmer, G. Sawaya, Y. Yang, R. M. Kirby and G. Gopalakrishnan. Parallel and distributed model checking in Eddy. *International Journal on Software Tools for Technology Transfer*, 11(1):13–25, 2009. Springer.

[29] M. AlTurki and J. Meseguer. PVeStA;: A parallel statistical model checking and quantitative analysis tool. In A. Corradini, B. Klin and C. Cîrstea (editors), *Proceedings 4th International Conference on Algebra and Coalgebra in Computer Science*, volume 6859 of *LNCS*, pages 386–392. Springer, Berlin, 2011.

[30] G. Behrmann, A. David and K. Larsen. A tutorial on Uppaal. In M. Bernardo and F. Corradini (editors), *Formal Methods for the Design of Real-Time Systems*, volume 3185 of *LNCS*, pages 33–35. Springer, Berlin, 2004.

[31] M. Chechik and A. Gurfinkel. TLQSolver: A temporal logic query checker. In W. A. Hunt Jr. and F. Somenzi (editors), *Proceedings 15th International Conference on Computer Aided Verification*, volume 2725 of *LNCS*, pages 210–214. Springer, Berlin, 2003.

[32] J. E. Stajich, D. Block, K. Boulez et al. The Bioperl toolkit: Perl modules for the life sciences. *Genome Research*, 12(10):1611–1618, 2002.

[33] M. Steel and A. McKenzie. Properties of phylogenetic trees generated by Yule-type speciation models. *Mathematical Biosciences*, 170(1):91–112, 2001.

[34] V. Bruyere and J.-F. Raskin. Real-time model-checking: Parameters everywhere. In P. K. Pandya and J. Radhakrishnan (editors), *Proceedings 23rd Conference on Foundations of Software Technology and Theoretical Computer Science*, volume 2914 of *LNCS*, pages 100–111. Springer, Berlin, 2003.

[35] M.-S. Peng, J.-D. He, C.-L. Zhu, S.-F. Wu, J.-Q. Jin and Y.-P. Zhang. Lactase persistence may have an independent origin in Tibetan populations from Tibet, China. *Journal of Human Genetics*, 57(6):394–397, 2012.

[36] A. Stamatakis. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.