Data-Driven Statistical Learning of Temporal Logic Properties*

Ezio Bartocci[1], Luca Bortolussi[2,3], and Guido Sanguinetti[4,5]

[1] Faculty of Informatics, Vienna University of Technology, Austria
[2] DMG, University of Trieste, Italy
[3] CNR/ISTI, Pisa, Italy
[4] School of Informatics, University of Edinburgh, UK
[5] SynthSys, Centre for Synthetic and Systems Biology, University of Edinburgh, UK

Abstract. We present a novel approach to learn logical formulae characterising the emergent behaviour of a dynamical system from system observations. At a high level, the approach starts by devising a data-driven statistical abstraction of the system. We then propose general optimisation strategies for selecting formulae with high satisfaction probability, either within a discrete set of formulae of bounded complexity, or a parametric family of formulae. We illustrate and apply the methodology on two real world case studies: characterising the dynamics of a biological circadian oscillator, and discriminating different types of cardiac malfunction from electro-cardiogram data. Our results demonstrate that this approach provides a statistically principled and generally usable tool to logically characterise dynamical systems in terms of temporal logic formulae.

1 Introduction

Dynamical systems are among the most widely used modelling frameworks, with important applications in all domains of science and engineering. Much of the attraction of dynamical systems modelling lies in the availability of effective simulation tools, enabling predictive modelling, and in the possibility of encoding complex behaviours through the interaction of multiple, simple components. This leads naturally to the notion of *emergent properties*, i.e. properties of the system trajectories which are a non-trivial consequence of the local interaction rules of the system components. Emergent properties of deterministic dynamical systems can often be easily verified through simulations. Quantitatively identifying the emergent properties of a stochastic system, instead, is a much harder problem.

In the simplest scenario, one assumes that a mathematical model of the system of interest is already available (e.g. as a continuous time Markov chain, or a stochastic differential equation), generally thanks to the availability of domain expertise. This problem is often termed *mining requirements*: this is an active field of research, with many recent contributions extending its scalability and applicability [18,27]. This approach

* L.B. acknowledges partial support from the EU-FET project QUANTICOL (nr. 600708) and by FRA-UniTS. G.S. acknowledges support from the ERC under grant MLCS306999. E.B. acknowledges the support of the Austrian FFG project HARMONIA (nr. 845631).

A. Legay and M. Bozga (Eds.): FORMATS 2014, LNCS 8711, pp. 23–37, 2014.

is predicated on two premises: first, a trustworthy model of the system must be available, and, secondly, efficient model checking algorithms must be available for the class of properties/models under consideration. These two conditions are often onerous in many scientific applications, where models can be both complex and highly uncertain. However, data generation is becoming increasingly cheap for many complex systems, raising the possibility that emergent properties may be formally identified from data.

This problem, although clearly of considerable practical relevance, has received comparatively little attention in the literature. Early work by [10] proposed a greedy algorithm to identify formulae with high support directly from data, with the ultimate aim of unravelling the logical structure underpinning observed dynamics in systems biology. More recently, Asarin et al. in [4] proposed a geometric construction to identify the formula (within a specified parametric family) which fitted observations best. In both cases, the methods work directly with the raw data, and are hence potentially vulnerable to noise in the data. Furthermore, both sets of authors remark that the identifiability of formulae is severely limited by the quantity of data available, which hampers the applicability of the methods in many practical circumstances.

Here, we aim to address both identifiability and robustness problems by taking an alternative, statistical approach, which brings back a model-based perspective to the data-driven approach. We consider a variation of the property learning problem, where we observe trajectories from two distinct processes and the aim is to identify properties that best discriminate between two observed processes, i.e. are satisfied with high probability by trajectories from one process and with low probability by trajectories from the second. At the core of our method is a *statistical abstraction*, a flexible, data driven statistical model which provides a compact representation of the dynamics of the system. The choice of the statistical model is performed using statistical model selection techniques, combining domain expertise with data driven methods; in this paper, we will illustrate our approach on two contrasting applications: a systems biology application where considerable prior knowledge permits the use of a rather restricted and complex family of candidate models, and a biomedical application where such knowledge is unavailable, and hence we use a more black box model. Once a suitable model is selected, the satisfaction probability of a formula can be evaluated quantitatively (using a model checking tool), enabling rational selection of formulae with high support or that best discriminate two models obtained from two datasets. This *property learning* problem can be further broken down into two subproblems: learning the structure of the formula, and learning parameters involved in the formula. These optimisation problems can be tackled in many ways: here, we use a local search algorithm for structure learning, and a recently proposed, provably convergent algorithm [25] for learning the parameters of the formula. Figure 1 illustrates schematically the modular structure of our approach.

The rest of the paper is organised as it follows: in the next section we give an overview of the proposed approach, reviewing the relevant statistical and logical concepts. We then present results on the two case studies, briefly describing the procedure through which the statistical model was devised in each case, and illustrating the capabilities of the approach to infer non-trivial properties from the data. We conclude the paper by discussing the implications of our contribution, both from the practical and the methodological aspect.

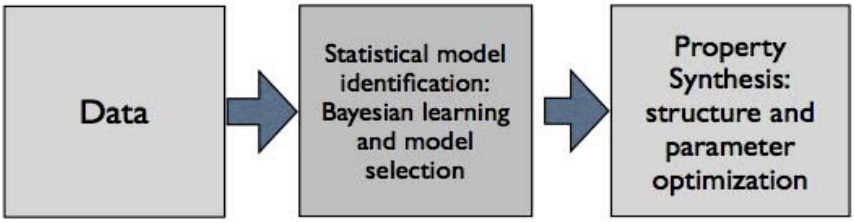


Fig. 1. Schematic workflow of our approach: starting from data, a suitable statistical model is chosen (within a family of models) via Bayesian learning methods. That enables us to evaluate the probability of formulae in a suitable logic, which can then be maximised as a function of the formula structure or formula parameters

2 Problem Statement and Methodology

The property synthesis problem can have many flavours. One can be interested in finding the properties that best characterise a single set of observations, or find properties that discriminate between a good and a bad set of observed scenarios. The examples discussed in this paper fall into this second class, but a similar machinery can be used for finding properties with high support. The discrimination problem is the following:

Given two sets of signals/time traces (the good and the bad set), find a temporal logic formula that best discriminates between them, i.e. such that it is satisfied with high probability by the good set and with low probability by the bad one.

Essentially, this problem can be seen as a temporal logic version of a classification problem, in which we look for temporal patterns separating two sets of observed signals.

At a high level, our approach is made up of two distinct modules: a model selection step, where a suitable statistical model is learnt from the data, and a property synthesis step, where we perform learning of formulae with high discriminating power. The advantage of the statistical generalisation performed in the first phase, which distinguishes our approach from other related work (see also Section 4) is that it offers a statistically sound treatment of noise and the ability of generating simulated data, avoiding the data shortage problem in the second phase.

In this section we describe the methods we use for performing these steps in this paper, as well as providing some background on the specific logic we will use to encode emergent properties. We emphasize however that, while we believe the choices we made in performing the two steps are state-of-the-art, the concept of our approach is entirely modular, so that any other model selection/optimisation method could be employed towards the same goal.

2.1 Statistical Modelling of Data: Learning and Model Selection

Our statistical methodologies will be embedded within the *probabilistic machine learning* framework [8]. Let $\mathbf{x}$ denote the state variables associated with our system, and let

$\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N$ denote observations of the system at times $0 \leq t_1, \ldots, t_N \leq T$. Our statistical models will then take the form of *joint probability distributions*

$$p(\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N, \mathbf{x}_{0:T}, \mu_{0:T}|\Theta) = p(\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N|\mathbf{x}_{0:T}, \Theta)\, p(\mathbf{x}_{0:T}, \mu_{0:T}|\Theta)$$

where μ represent a set of auxiliary *latent variables* and the index $0:T$ denotes the *whole trajectory* of the respective stochastic process within the bounded time interval $[0,T]$. In general, we will assume that the prior dynamics of the system (specified by $p(\mathbf{x}_{0:T}, \mu_{0:T}|\Theta)$) are Markovian, and that the observation noise is independent and identically distributed at different time points. Additionally, the models are parametrised by a family of parameters Θ which may enter both the noise model (probability of the observations given the true state of the system $\mathbf{x}$) and the prior dynamics $p(\mathbf{x}_{0:T}, \mu_{0:T}|\Theta)$. The introduction of the latent variables can be justified in several ways: in some cases, the latent variables represent physically relevant unobserved quantities (e.g., promoter occupancy state as in Section 3.1); otherwise, they may be a convenient device to represent a more rich dynamics in a compact way (as in the heart modelling example in Section 3.2). We stress that Hidden Markov Models, Continuous-Time Markov Chains, (Stochastic) Differential Equations and Hybrid Systems all fall into the class of models considered here.

The general principle for learning in probabilistic models is based on the concept of *evidence maximisation*, whereby one seeks to determine the value of parameters Θ that maximises the evidence or marginal likelihood

$$p(\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N|\Theta) = \int d\mu_{0:T} d\mathbf{x}_{0:T}\, p(\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N, \mathbf{x}_{0:T}, \mu_{0:T}|\Theta)$$

where the integral sign is used generically to denote marginalisation (it is replaced by a sum in the case of discrete variables). In general, the marginalisation procedure is computationally problematic, and much research in machine learning is devoted to find efficient marginalisation algorithms for specific classes of models.

The evidence at the optimal value of the parameters provides a measure of the goodness of fit of a model to a data set. However, models with different numbers of parameters will not necessarily be comparable in terms of evidence: richer models with more parameters tend to have higher evidence. One therefore needs to penalise the complexity of the model. There exist several information criteria which combine the maximum value of the likelihood with a penalty on the number of parameters. Here we use the *Akaike Information Criterion (AIC)* [1], which penalises the likelihood by subtracting a term containing the logarithm of the number of parameters. Explicitly, the AIC score is defined as

$$AIC = 2k - 2\log L$$

where k is the number of parameters of the model, and L is the optimised value of the marginal likelihood. This simple score can be shown to approach asymptotically, in the large sample limit, the information lost by using the model as a proxy for the (unknown) data generating process. Therefore minimisation of the AIC score across a finite number of models is often used as a criterion for model selection.

2.2 Learning Properties

The second module of our approach consists of algorithms for optimising the probability of a formula being true within a discrete set of parametric formulae. This difficult hybrid optimisation problem is naturally broken down in a discrete and a continuous optimisation problem, which can be interleaved. Before describing the algorithms we use, we briefly review the logic we consider, the Metric Interval Temporal Logic (MITL).

Metric Interval Temporal Logic. Temporal logic [22] provides a very elegant framework to specify in a compact and formal way an emergent behaviour in terms of *time-dependent* events. Among the myriads of temporal logic extensions available, Metric Interval Temporal Logic [3] (MITL) is very suitable to characterise properties of (real-valued) signals evolving in continuous time. The syntax of MITL is as follows.

Definition 1 (MITL syntax). *The syntax of MITL is given by*

$$\varphi := \top \mid q \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \mathbf{X}_{[a,b]}\varphi \mid \varphi_1 \, \mathbf{U}_{[a,b]} \, \varphi_2,$$

where $\top$ is a true formula ($\bot = \neg\top$ is false), q is an atomic proposition which is either true or false in each state $\mathbf{x}$ (we denote with $\mathcal{L}(\mathbf{x})$ the set of atomic propositions true in $\mathbf{x}$), conjunction and negation are the standard boolean connectives, $[a,b]$ is a dense-time interval with $0 \leq a < b$, $\mathbf{X}_{[a,b]}$ is the next *operator and $\mathbf{U}_{[a,b]}$ is the until operator.*

The (bounded) until operator $\varphi_1 \, \mathbf{U}_{[a,b]} \, \varphi_2$ requires φ_1 to hold from now until, in a time between a and b time units, φ_2 becomes true, while the (bounded) next operator $\mathbf{X}_{[a,b]}\varphi$ requires φ to hold in the next state, to be reached between a and b units of time. The *eventually* operator $\mathbf{F}_{[a,b]}$ and the *always* operator $\mathbf{G}_{[a,b]}$ can be defined as usual: $\mathbf{F}_{[a,b]}\varphi := \top \mathbf{U}_{[a,b)}\varphi$, $\mathbf{G}_{[a,b]}\varphi := \neg\mathbf{F}_{[a,b]}\neg\varphi$. More precisely, MITL can be given a semantics based on boolean signals, which are functions of time to $\{\top, \bot\}$. Boolean signals corresponding to atomic propositions are obtained from a (real-valued) input signal $\mathbf{x}(t)$ by point-wise lifting: $q(t) := q \in \mathcal{L}(\mathbf{x}(t))$. The extension of MITL that deals with real-valued signals is known as Signal Temporal Logic, see [19] for further details on the logic and the monitoring algorithm.

MITL is a logic that is interpreted over traces, and a formula φ identifies the subset of traces that satisfy it, $\{\mathbf{x} \models \varphi\}$. A stochastic model $\mathcal{M}$, however, is a probability distribution on the space of traces, and as such we can measure how much $\mathcal{M}$ satisfies φ by computing the probability $p(\varphi|\mathcal{M}) = Prob_{\mathcal{M}}\{\mathbf{x} \models \varphi\}$. This probability is notoriously difficult to calculate analytically even for simple models [11], hence, we resort to Monte Carlo methods, applying statistical model checking (SMC) [15] to estimate the probability of a MITL formula in a generative model.

Discrimination Function. In order to set up a proper learning problem, we need to consider a score function to optimise, encoding the criterion to discriminate between two models. Here we choose a simple score function, namely the log odds ratio between the satisfaction probabilities. More precisely, let $\mathcal{M}_1$ and $\mathcal{M}_2$ be the two models learnt

from the two datasets and φ a candidate MITL formula. The log odds ratio score R_φ is defined as

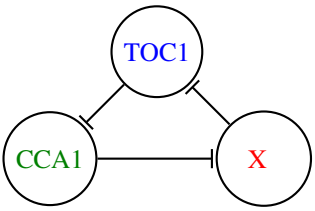
$$R_\varphi = \log \frac{p(\varphi_1 \mid \mathcal{M}_1)}{p(\varphi_2 \mid \mathcal{M}_2)}, \tag{2.1}$$

and it is maximised when the probability of the first model is close to one and the probability in the second model is close to zero.

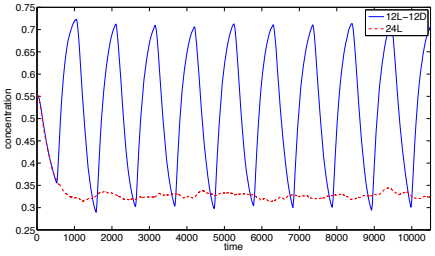
Structure Learning. Identifying the structure of a MITL formula which is satisfied with high probability by the model is a difficult combinatorial optimisation problem. Combinatorial optimisation algorithms exist but we are aware of few theoretical convergence guarantees. In this paper we do not tackle the problem in its full generality, but we set up a greedy search scheme which requires some basic knowledge of the domain at hand.

More specifically, we assume to have a fixed set of *basic* template formulae $\mathcal{T}$. First, we search exhaustively in $\mathcal{T}$ by optimising the continuous parameters of each $\varphi \in \mathcal{T}$, and thus computing its best score (i.e the log odds ratio). Then, we rank the formulae in $\mathcal{T}$ and select the subset of higher score. If in this way we find a few good candidate formulae, we proceed to the second phase, otherwise we enlarge the set $\mathcal{T}$, and try again. The choice of the thresholds to select good candidates is delicate and problem dependent. In the second phase, we take the best formulae $\mathcal{T}_{best}$ and combine them using some predefined combination rules (for instance, boolean combinations), and run again the continuous optimisation on the parameters, ranking again the formulae and selecting those with highest score. As the set $\mathcal{T}_{best}$ is expected to be small, we will be searching exhaustively a reasonably small set of formulae. At this stage, we expect this greedy optimisation to have found some good formula. If not, we can proceed to combine together the best formulae of this second round, possibly with another set of combinators, or reconsider the choice of the basic templates $\mathcal{T}$.

Parameter Learning. We now turn to the issue of tuning the parameters of (a set of) formulae to maximise their satisfaction probability. More specifically, we assume that we have a MITL formula φ_θ which depends on some continuous parameters θ. We aim to maximise its discriminative power $R_\varphi(\theta)$ defined in equation (2.1). Naturally, this quantity is an intractable function of the formula parameters; its value at a finite set of parameters can be noisily estimated using an SMC procedure. The problem is therefore to identify the maximum of an intractable function with as few (approximate) function evaluations as possible. This problem is closely related to the central problem of reinforcement learning of determining the optimal policy of an agent with as little exploration of the space of actions as possible. We therefore adopt a provably convergent stochastic optimisation algorithm, the GP-UCB algorithm [25], to solve the problem of continuous optimisation of formula parameters. Intuitively, the algorithm interpolates the noisy observations using a stochastic process (a procedure called emulation in statistics) and uses the uncertainty in this fit to determine regions where the true maximum can lie. This algorithm has already been used in a formal modelling scenario in [9].



(a) Repressilator-like gene network of the O. Tauri circadian clock [21]



(b) O. Tauri circadian clock: average.

Fig. 2. The repressilator-like model of the O.Tauri circadian clock (left) is a cyclic negative-feedback loop composed of three repressor genes: TOC_1, CCA_1, and an unknown gene X. The comparison of the average evolution (right) of the circadian clock for the 12h light/12h dark model (blue solid line) and the 24h light model (red dashed line) shows that light plays a crucial role in stabilising the oscillatory period. Parameters of the simulation are as in [21].

3 Results

3.1 Logical Characterisation of a Biological Oscillator

Our first case study is the circadian clock in *Ostreococcus Tauri*, a simple unicellular alga often used as a minimal plant model organism [21]. The circadian clock is an important regulator of the metabolism of the plant and is controlled by the mutually repressive interaction of three genes, TOC1, CCA1 and one expressing a not yet characterized protein (denoted as X here), see Fig. 2 left for a scheme of the genetic circuit. Gene repression of the TOC1 gene is further modulated by light, which plays the role of an external input and acts as a stabiliser of the oscillatory pattern.

In this example, a parametrised statistical model was already learned from data in [21]. The stochastic hybrid models we consider couple Stochastic Differential Equations (SDEs) for protein dynamics with a two-state model of gene promoter, which can be either free or bound to the repressor. In the latter case, the protein expression is reduced. More precisely, the protein dynamics is given by the SDE

$$dX_i = (A_i\mu_i + b_i - \lambda_i X_i)dt + \sigma dW,$$

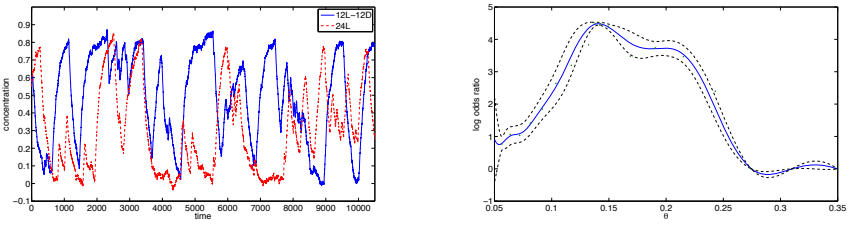
where μ_i denotes the state of the promoter gene i (with $\mu_i = 1$ denoting the repressed state and $\mu_i = 0$ the active state), b_i is the basal production rate, $A_i < 0$ reduces it in case of repression, and λ_i is the degradation rate. The dynamics of the promoter is a two-state Markov chain with switch rates given by

$$f_{bind,i}(\mathbf{X}) = k_{p_i}\exp(k_{e_i}X_j), \quad f_{unbind,i}(\mathbf{X}) = k_u,$$

i.e. with constant unbinding rate and with binding rate depending on the repressor concentration. To model the influence of light on the protein TOC1, we modify the binding and unbinding rates of its regulatory protein X as follows:

Table 1. Statistics for six runs of the optimization of parameter $\theta \in [0.05, 0.35]$. The algorithm was initialised by sampling the function (2.1) 16 times from 100 simulation runs of each mode, and terminated with less than 4 additional samples on average. The variability of the results is due to the noisy nature of the function evaluation.

Av. θ	Av. log odds ratio	Av. sat. prob. 12L-12D	Av. sat. prob. 24L
0.148	4.295	0.83	0.008
Range θ	Range log odds ratio	Range sat. prob. 12L-12D	Range sat. prob. 24L
[0.138, 0.157]	[3.689, 4.522]	[0.77, 0.87]	[0.004, 0.012]



(a) O. Tauri circadian clock: single trace. (b) Log odd ratio: emulation function

Fig. 3. Left: Single trace of TOC1 protein evolution for the O. Tauri model, with parameters as in [21]. Right: Emulated log odds ratio as a function of the threshold θ of MITL Formula 3.2 (blue solid line) and 95% error bounds (black dashed lines). The green points are the estimated values of the function.

$$f_{bind,TOC1}(\mathbf{X}) = (1-\gamma)\left(k_{p_i}\exp(k_{e_i}X_j)\right) + \gamma \cdot k^b_{light}(t), \quad f_{unbind,i}(\mathbf{X}) = (1-\gamma)k_u + \gamma \cdot k^u_{light}(t),$$

where γ is set to 0.20 and the values of $k^{b/u}_{light}(t)$ depend on the light conditions.

As an example of our property learning procedure, we seek a temporal logic formula which discriminates system trajectories between the following two conditions: the system is *entrained*, i.e. is receiving a 12h light/12h dark input signal (12L-12D), or it is being kept in constant light (24L). These conditions are encountered by *O. tauri* at high latitudes, and it is a scientifically important question how clock regulation can withstand such extreme environmental changes. In this example, we will *fix a template* and limit ourselves to learn parameters with the Bayesian continuous optimisation scheme discussed in Section 2. The key difference between the 12L-12D model and the 24L model is that oscillations in the 12L - 12D regime should maintain phase coherence with the input. This is indeed true, as can be seen from Figure 2(b), where we show the average of 500 trajectories. Detecting phase coherence on single trajectories is however a much tougher proposition, as subtle phase shifts can be easily masked by irregularities due to the intrinsic stochasticity of the processes, see Figure 3(a).

We therefore use the approach of [12], converting the signal from the time domain to the frequency domain using the Short Time Fourier Transform (STFT) [2]. This technique is generally employed to analyse non-stationary signals, whose statistic characteristics vary with time. STFT consists of reading the samples of the signals using fixed

window of time where to apply the standard Fourier transform. The result is a spectrogram where is possible to observe for each region of time the characteristic frequencies of the signal. Since we know the oscillation frequency ω_o of the 12L-12D model, by fixing a time window of $T_o = 10/\omega_o$, we expect to find a peak in the STFT at frequency ω_o in the 12L-12D model, but not in the 24L model. Using the STFT (with rectangular window), we can require this peak to persist for a certain amount of time T, leading to the formula

$$\varphi_{FFT,\theta} = G_{[0,T]}(f(\omega_0,\cdot) \geq \theta), \tag{3.2}$$

where $f(\omega_0,t)$ is the absolute value of the STFT at frequency ω_o for the window of length T_o starting at time t, and T is fixed to 1000. The goal therefore becomes to find the best discriminating θ. In Table 1, we report the results of 6 runs of the optimization algorithm, searching for the best $\theta \in [0.05, 0.35]$, while the functional dependency of log odds ratio on θ, as emulated by the Bayesian optimisation procedure, is shown in Figure 3(b). We find an optimal value (the median from the 6 runs) of 0.1492, corresponding to a satisfaction probability in the 12L-12D model of approximatively 0.84 and a satisfaction probability in the 24L model of approximatively 0.01, confirming that this formula has a good discriminatory power.

3.2 Logical Discrimination of Cardiac Arrhythmias

Basic cardiac physiology - Arrhythmias are electrophysiological cardiac malfunctions which cause significant mortality and morbidity. The most common, non-invasive diagnostic tool to monitor the heart's electrophysiological function is the electrocardiogram (ECG). An ECG machine is able to record the electrical activity of the heart through a set of electrodes (called ECG leads) placed by the physician on the chest wall and limbs of the patient. As Figure 4 b) illustrates, in a healthy patient the ECG signal consists of three main consecutive waves: **the P wave** corresponding to the depolarization and the consequent contraction of the atria, **the QRS complex** representing the rapid depolarization and contraction of the ventricles and **the T wave** identifying the recovery or depolarization of the ventricles.

ECG signals are interpreted by physicians through a hierarchy of annotations. The fundamental unit in the ECG is the heartbeat (or, simply, beat) defined as the interval between two consecutive **R peaks**. The beats are annotated using a symbol characterizing the type of beat observed (some of them shown in Figure 4 a-d). Beats are usually machine annotated through pattern recognition algorithms such as support vector machines. In this work, we will use directly an annotated version of the ECG signals as a sequence of beat symbols with associated beat durations.

A higher level annotation of ECG data is given by the *rhythms*, sequences of beats exhibiting a coherent pattern. Figure 5 a) shows an example of an ECG pattern for a normal sinus rhythm. Even in this case some abnormal heartbeats (such as a premature ventricular contraction in Figure 5 d)) can sporadically occur without medical significance. We present here initial results on annotated ECG data from the MIT-BIH Arrhythmia Database [20]. We restricted our attention to a subset of possible rhythms which were more prevalent in the data: bigeminy, trigeminy, ventricular tachycardia and the normal rhythm. These signals are predominantly composed of **V** and **N** symbols, often with a similar frequency, hence discrimination is more challenging.

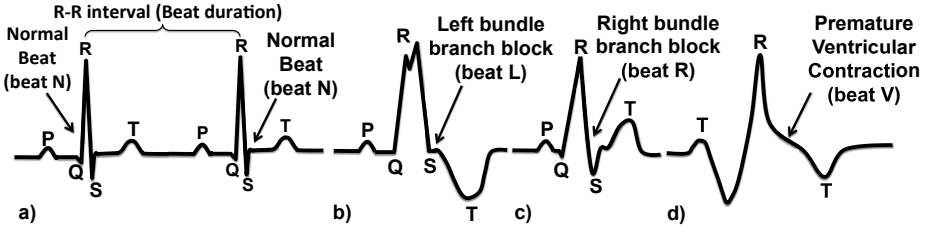


Fig. 4. a) ECG pattern for two normal beats; b-c) Left /Right bundle branch block is an abnormal beat where one ventricle is delayed and contracts later than the other; d) Premature ventricular contraction is characterized by a premature wider QRS complex, not preceded by a P wave and followed by an usually large T wave with an opposite concavity than in the normal beat.

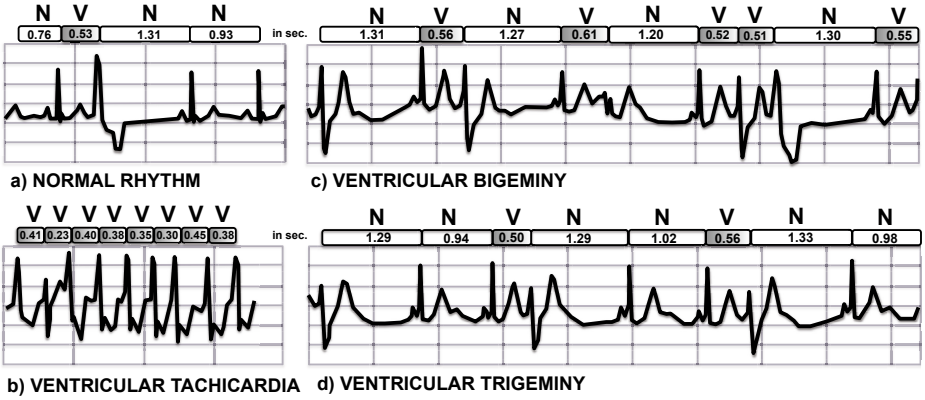


Fig. 5. Some ECG patterns: a) normal sinus rhythm; b) ventricular tachycardia, b) ventricular bigeminy, c) ventricular trigeminy. On the top of each signal is reported the annotation for each beat and its duration in seconds, while on the bottom is reproduced the electrical signal. The ECG data was obtained from the MIT-BIH Arrhythmia Database [20]

Statistical modelling - Due to the discrete time nature of the signal, we selected Hidden Markov Models (HMMs) as a class of statistical models that could provide a suitable statistical abstraction of the data. HMMs [23] are a workhorse of statistics and signal processing and have been previously employed in the context of formal modelling of heart function in [6]. Briefly, an HMM is a tuple $H = \langle S, A, O, B, \pi \rangle$ containing a finite set S of states, a transition probability matrix A, a set O of observation symbols, an observation probability distribution B, and an initial state distribution π. In our case, we have hybrid observations consisting of pairs o_s, o_t where o_s is the emitted symbol (type of beat) and o_t the beat duration (in seconds). We therefore assume the observation probability to factorise as a product of a discrete probability on the beat types and a Gaussian on the observation times. HMM models were learnt using the Baum-Welch algorithm [23] and selected using the AIC score defined in Section 2.

Summary of experimental procedure - Due to space restrictions, we present results on a single patient (patient 233); other patients yielded similar results. Code to recreate the experimental results is available for academic use from the authors upon request. The experimental procedure can be summarised as follows

- For each rhythm, we learn HMM models with 2 to 6 states, and select the one with best AIC score. We learn the model simultaneously on all segments annotated as a certain rhythm (e.g. bigeminy).
- For each pair of abnormal/ normal rhythm, we learn template formulae starting from the basic set of formulae $\mathcal{T}_2$, corresponding to possible patterns of length 2 of symbols V and N: $\mathcal{T}_2 = \{\mathbf{FG}_{\leq T}\varphi_{NN}, \mathbf{FG}_{\leq T}\varphi_{NV}, \mathbf{FG}_{\leq T}\varphi_{VN}, \mathbf{FG}_{\leq T}\varphi_{VV}\}$, where $\varphi_{YZ} = Y \wedge (\mathbf{X}_{[0,b_Y]}(Z \wedge \mathbf{X}_{[0,b_Z]}(\top)))$ and we optimise the continuous parameters (T, b_N, b_V) to obtain the maximum discriminative power [1].
- If after the optimisation phase no highly supported formula was found, i.e. a formula with high log-odd ratio of abnormal versus normal signal and high satisfaction probability, we rerun the procedure increasing the pattern length of one (hence, first for $\mathcal{T}_3$, then $\mathcal{T}_4$, and so on).
- We then selected the most supported formulae of $\mathcal{T}_k$ to further combine them, as discussed in the previous section. We run the continuous optimisation also for these formulae, and chose the ones having both high log-odd ratio and satisfaction probability for the abnormal rhythm.

We now present results on discrimination of the three abnormal rhythms in more detail.

Bigeminy - Learning formula templates for the discrimination of bigeminy against normal heart behaviour proceeded as follows: in the first optimisation run, the two formulae with highest log-odd ratio where $\mathbf{FG}_{\leq T}\varphi_{NV}$ and $\mathbf{FG}_{\leq T}\varphi_{VN}$, scoring more than 5, with a satisfaction probability in bigeminy of about 0.8.

The other two formulae, instead, have a log-odd ratio zero or less. Hence, we selected these two formulae for the second phase of the discrete search, obtaining $\mathbf{FG}_{\leq T}(\varphi_{NV} \vee \varphi_{VN})$ as the only candidate for the second round. This formula clearly codes for the pattern VN repeated many times (for as long as T units of time). Running the continuous optimisation, we find a log-odd ratio of 4.08, which is lower than in the previous case, but it corresponds to a satisfaction probability of 0.9994 in the abnormal rhythm, and a probability of 0.016 in the normal one, corresponding to a sensitivity of $> 99\%$ and a specificity of approximately 98%. Hence, this formula turns to have a good discriminative power, and its relatively low log-odd ratio depends on its high sensitivity to small values of the denominator. The upper bound of time T is optimally set to 3.8, close to the maximum of 4. Upper bounds on beat duration are also close to their maximum. Note that the alternation of V and N is precisely what characterises bigeminy: our method learned the correct pattern used by physicians, and additionally quantitated the time such a pattern persists for.

[1] We search in the following space: maximal duration of symbols is constrained between 0 and 2.5 seconds, while the lower bound was set to zero. The total duration T varies between 0 and an upper bound depending on the signal, equal to 4 for bigeminy, 7 for trigeminy, 2 for tachycardia. We generate signals of fifteen seconds. The choice of bounds for T is consistent with the duration of raw signals in the training set.

Table 2. Average log-odd ratio and satisfaction probability of abnormal and normal signals for the best discriminating formulae learned from patient 223, as tested on other three patients (per type of abnormality). The fourth row reports the number of signals in the training set to learn each model of the abnormal signals in patient 223. The formulae learnt were tested on patients: 119, 213, and 233 for bigeminy; 119, 201,208 for trigeminy; 213, 215, 233 for ventricular tachycardia. The fifth row shows the number of signals considered in the testing set per type of abnormality.

	Bigeminy	Trigeminy	V. Tachycardia
Av. log-odd ratio	3.32	2.99	7.68
Av. prob. abnormal	0.99	0.99	0.99
Av. prob. normal	0.06	0.08	0.0005
Learning Set (num. of signals)	7	3	7
Testing Set (num. of signals)	84	75	10

Trigeminy - To discriminate trigeminy vs normal rhythm, we proceeded analogously as for bigeminy, starting with the same set $\mathcal{T}_2$ of formulae. In this case, however, no formula of length 2 was found to have a high support in discriminating trigeminy (less than 3.5), hence we considered basic templates corresponding to patterns of length 3. The analysis in this case gave high log-odd ratio (4 or greater) to three formulae: $\mathbf{FG}_{\leq T}\varphi_{VNN}$, $\mathbf{FG}_{\leq T}\varphi_{NVN}$, and $\mathbf{FG}_{\leq T}\varphi_{NNV}$, with for a small duration T for all three cases. We then took all possible combination of at least two of those formula using disjunction, and found the most discriminating formula (log-odd ratio 7.8, satisfaction probability for trigeminy 0.9968, and for normal signal of 0.004) to be $\mathbf{FG}_{\leq T}(\varphi_{VNN} \vee \varphi_{NVN} \vee \varphi_{NNV})$, corresponding to the pattern VNN repeating in time for approximatively $T = 4.25$ seconds. Again in this case, the method found the hallmark pattern of trigeminy and additionally quantified its persistent behaviour. We also tested that this formula works well in discriminating trigeminy versus bigeminy (log-odd ratio of 8.5).

Ventricular tachycardia - This case turned out to be the simplest one. A good discriminating formula was found already in the set $\mathcal{T}_2$, corresponding to the pattern VV. In particular, the continuous optimisation returned a log-odd ratio of 2.9, corresponding to a satisfaction probability in the abnormal rhythm of 0.9998 and of 0.05 in the normal rhythm, with the global validity time T set approximatively to 1.25 seconds. This corresponds to tachycardia being characterised by a stretch of about 3 to 4 V beats.

Discrimination on other patients - So far, we considered discriminative power as applied to the same patient on which the models were learnt. We now consider the much harder task of assessing whether estimated formulae remain discriminative when also applied to other patients' ECG data. We considered three other patients for each arrhythmia, and obtained an high discriminative power, as reported in Table 2. We also tested the formulae on raw signals taken from the database[2], obtaining the following results, in terms of satisfaction probability: 0.954 for bigeminy versus 0.038 for normal rhythms (on the same patients); and 0.918 for trigeminy versus 0.287 for normal rhythms. The high satisfaction probability on normal rhythms for this last case is almost entirely

[2] 22 signals for bigeminy of length at least 4.5, 49 for trigeminy of length at least 5, and about 80 for normal rhythms. We did not treat tachycardia because there were too few signals.

explained by the presence, in the extended data set, of several patients with slow heart beats. The relatively low duration of the pattern learnt on the training patient, 4.25s, can be matched in the slow beat patients by patterns consisting of a single V preceded and followed by two Ns, which are very common in normal rhythm. Increasing the time bound to 7s reduced the satisfaction probability in normal rhythms to 0.014, while the satisfaction probability for trigeminy remained stable to 0.906.

4 Related Work

Mining temporal logic specifications from data is an emerging field of computer aided verification [4,10,18,14,27,28]. Generally, this task is predicated on the availability of a fully specified and deterministic model, enabling a quantitative evaluation of the probability that a certain formula will hold. This enables the deployment of optimisation based machine learning techniques, such as decision trees [14] or stochastic optimisation methods [28,27]. Learning temporal logic specifications directly from observed traces of the system is considerably more challenging. In general, solving the full structure and parameter learning problem is infeasible, due to the intractability resulting from a hybrid combinatorial/continuous optimisation problem. Heuristic search approaches have been proposed in [10]; while these may prove effective in specific modelling problems, they generally do not offer theoretical guarantees, and can be prone to over-fitting/vulnerable to noise. Geometric approaches such as the one proposed in [4] rest on solid mathematical foundations but can also be vulnerable to noise, and require potentially very large amounts of data to permit identification. While preparing this manuscript, we became aware of a work of imminent publication [17] which employs a notion of robustness of satisfiability of a formula to guide an optimisation based mining procedure. While this approach can be applied also in a model-free scenario, empirical estimation of the robustness of a formula may require the observation of a large number of traces of the system; for example, one of the case studies in [17] used 600 independent realisations of the system, a number that far exceeds the experimental capabilities in many applications such as systems biology.

Our strategy of constructing a statistical model of data ameliorates this issue, at the price of an increased complexity of the mining problem, which we tackle by combining statistical modelling ideas from machine learning with formal verification methods. In this respect, our work is related to a number of other recent attempts to deploy machine learning tools within a verification context [7,26,16]. Similar ideas to the ones used in this paper have been deployed on the parameter synthesis problem in [9,5], where the GP-UCB algorithm was used to identify the parameters of a model which maximised the satisfaction/robustness of a formula. Statistical abstractions draw their roots in the *emulation* field in statistics: within the context of dynamical systems, emulation has been recently used in [13] to model compactly the interface between subsystems of complex gene regulatory networks.

5 Conclusions

Modern science is increasingly becoming data intensive, with vast amounts of data being produced across disciplines as diverse as economics, physics and biology. Marrying

formal computational modelling with statistical, data-modelling techniques is therefore a pressing priority to advance the applicability of computational thinking to real world problems. In this paper, we exploit concepts from formal modelling and machine learning to develop methodologies which can identify temporal logic formulae which discriminate different stochastic processes based on observations. While we aim to be guided by the data, our approach is not entirely data driven: approaches which rely directly on induction from data, such as [10,4], often need very long time series, which are not available in many applications such as systems biology. Rather, we use a statistical abstraction, i.e. a family of stochastic models, to represent the data, and use machine learning methods to select an optimal model based on the data. This procedure brings back a model based perspective, with considerable advantages in terms of interpretability of the underlying dynamics. Furthermore, it enables us to deploy advanced machine learning methods to statistically optimise the temporal logic formulae we are seeking.

While we believe our machine learning perspective brings some distinctive novel ideas to the problem, several major avenues remain open for further research. Our approach focussed primarily on parametrising temporal logic formulae, rather than determining also a template for the formula structure. This combinatorial optimisation problem is intrinsically computationally hard, and may require directly imposing restrictions on the logic as in [17]. Scaling our approach to high dimensional spaces of parameters could also be problematic, as Bayesian optimisation methods severely suffer from the curse of dimensionality. In this respect, sparse approximation may be beneficial [24] but are so far untested in a Bayesian optimisation context. Finally, our experimental section demonstrated the applicability of our approach to a potentially wide class of problems. We hope this may lead to more focussed interdisciplinary studies in emerging application fields such as synthetic biology.

References

1. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Aut. Control 19(6), 716–723 (1974)
2. Allen, J.B.: Short term spectral analysis, synthesis, and modification by discrete fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing 3, 235–238
3. Alur, R., Feder, T., Henzinger, T.A.: The benefits of relaxing punctuality. J. ACM 43(1), 116–146 (1996)
4. Asarin, E., Donzé, A., Maler, O., Nickovic, D.: Parametric Identification of Temporal Properties. In: Khurshid, S., Sen, K. (eds.) RV 2011. LNCS, vol. 7186, pp. 147–160. Springer, Heidelberg (2012)
5. Bartocci, E., Bortolussi, L., Nenzi, L., Sanguinetti, G.: On the robustness of temporal properties for stochastic models. In: Proc. of HSB 2013, pp. 3–19 (2013)
6. Bartocci, E., Corradini, F., Di Berardini, M.R., Smolka, S.A., Grosu, R.: Modeling and simulation of cardiac tissue using hybrid I/O automata. Theor. Comput. Sci. 410(33-34), 3149–3165 (2009)
7. Bartocci, E., Grosu, R., Karmarkar, A., Smolka, S.A., Stoller, S.D., Zadok, E., Seyster, J.: Adaptive runtime verification. In: Qadeer, S., Tasiran, S. (eds.) RV 2012. LNCS, vol. 7687, pp. 168–182. Springer, Heidelberg (2013)
8. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
9. Bortolussi, L., Sanguinetti, G.: Learning and Designing Stochastic Processes from Logical Constraints. In: Joshi, K., Siegle, M., Stoelinga, M., D'Argenio, P.R. (eds.) QEST 2013. LNCS, vol. 8054, pp. 89–105. Springer, Heidelberg (2013)

10. Calzone, L., Chabrier-Rivier, N., Fages, F., Soliman, S.: Machine learning biochemical networks from temporal logic properties. In: Priami, C., Plotkin, G. (eds.) Trans. on Comp. Sys. Bio. VI. LNCS (LNBI), vol. 4220, pp. 68–94. Springer, Heidelberg (2006)
11. Chen, T., Diciolla, M., Kwiatkowska, M., Mereacre, A.: Time-bounded verification of CTMCs against real-time specifications. In: Fahrenberg, U., Tripakis, S. (eds.) FORMATS 2011. LNCS, vol. 6919, pp. 26–42. Springer, Heidelberg (2011)
12. Donzé, A., Maler, O., Bartocci, E., Nickovic, D., Grosu, R., Smolka, S.A.: On temporal logic and signal processing. In: Chakraborty, S., Mukund, M. (eds.) ATVA 2012. LNCS, vol. 7561, pp. 92–106. Springer, Heidelberg (2012)
13. Georgoulas, A., Clark, A., Ocone, A., Gilmore, S., Sanguinetti, G.: A subsystems approach for parameter estimation of ode models of hybrid systems. In: Proc. of HSB 2012. EPTCS, vol. 92 (2012)
14. Grosu, R., Smolka, S.A., Corradini, F., Wasilewska, A., Entcheva, E., Bartocci, E.: Learning and detecting emergent behavior in networks of cardiac myocytes. Commun. ACM 52(3), 97–105 (2009)
15. Jha, S.K., Clarke, E.M., Langmead, C.J., Legay, A., Platzer, A., Zuliani, P.: A Bayesian approach to model checking biological systems. In: Degano, P., Gorrieri, R. (eds.) CMSB 2009. LNCS, vol. 5688, pp. 218–234. Springer, Heidelberg (2009)
16. Kalajdzic, K., Bartocci, E., Smolka, S.A., Stoller, S.D., Grosu, R.: Runtime Verification with Particle Filtering. In: Legay, A., Bensalem, S. (eds.) RV 2013. LNCS, vol. 8174, pp. 149–166. Springer, Heidelberg (2013)
17. Kong, Z., Jones, A., Ayala, A.M., Gol, E.A., Belta, C.: Temporal Logic Inference for Classification and Prediction from Data. In: Proc. of HSCC 2014, pp. 273–282. ACM (2014)
18. Lee, C., Chen, F., Roşu, G.: Mining parametric specifications. In: Proc. of ICSE 2011, pp. 591–600. ACM (2011)
19. Maler, O., Nickovic, D.: Monitoring temporal properties of continuous signals. In: Lakhnech, Y., Yovine, S. (eds.) FORMATS/FTRTFT 2004. LNCS, vol. 3253, pp. 152–166. Springer, Heidelberg (2004)
20. Moody, G.B., Mark, R.G.: The impact of the MIT-BIH arrhythmia database. IEEE Eng. Med. Biol. Mag. 20(3), 45–50 (2001)
21. Ocone, A., Millar, A.J., Sanguinetti, G.: Hybrid regulatory models: a statistically tractable approach to model regulatory network dynamics. Bioinformatics 29(7), 910–916 (2013)
22. Pnueli, A.: The temporal logic of programs. In: IEEE Annual Symposium on Foundations of Computer Science, pp. 46–57 (1977)
23. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
24. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press (2006)
25. Srinivas, N., Krause, A., Kakade, S.M., Seeger, M.W.: Information-theoretic regret bounds for gaussian process optimization in the bandit setting. IEEE Transactions on Information Theory 58(5), 3250–3265 (2012)
26. Stoller, S.D., Bartocci, E., Seyster, J., Grosu, R., Havelund, K., Smolka, S.A., Zadok, E.: Runtime Verification with State Estimation. In: Khurshid, S., Sen, K. (eds.) RV 2011. LNCS, vol. 7186, pp. 193–207. Springer, Heidelberg (2012)
27. Xiaoqing, J., Donzé, A., Deshmukh, J.V., Seshia, S.A.: Mining Requirements from Closed-loop Control Models. In: Proc. of HSCC 2013, pp. 43–52. ACM (2013)
28. Yang, H., Hoxha, B., Fainekos, G.: Querying Parametric Temporal Logic Properties on Embedded Systems. In: Nielsen, B., Weise, C. (eds.) ICTSS 2012. LNCS, vol. 7641, pp. 136–151. Springer, Heidelberg (2012)