Artificial Intelligence in Medicine 57 (2013) 185–195

Contents lists available at SciVerse ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Discovering human immunodeficiency virus mutational pathways using temporal Bayesian networks

Pablo Hernandez-Leal[a,*], Alma Rios-Flores[a], Santiago Ávila-Rios[b], Gustavo Reyes-Terán[b], Jesus A. Gonzalez[a], Lindsey Fiedler-Cameras[a], Felipe Orihuela-Espina[a], Eduardo F. Morales[a], L. Enrique Sucar[a]

[a] Coordinación de Ciencias Computacionales, Instituto Nacional de Astrofísica, Óptica y Electrónica, Luis Enrique Erro #1, Sta. María Tonantzintla, Puebla, Mexico
[b] Centro de Investigación en Enfermedades Infecciosas, Instituto Nacional de Enfermedades Respiratorias, Calzada de Tlalpan #4502, Ciudad de México, Mexico

ARTICLE INFO

Article history:
Received 12 December 2011
Received in revised form 12 January 2013
Accepted 18 January 2013

Keywords:
Probabilistic graphical models
Probabilistic learning
Human immunodeficiency virus mutations

ABSTRACT

Objective: The human immunodeficiency virus (HIV) is one of the fastest evolving organisms in the planet. Its remarkable variation capability makes HIV able to escape from multiple evolutionary forces naturally or artificially acting on it, through the development and selection of adaptive mutations. Although most drug resistance mutations have been well identified, the dynamics and temporal patterns of appearance of these mutations can still be further explored. The use of models to predict mutational pathways as well as temporal patterns of appearance of adaptive mutations could greatly benefit clinical management of individuals under antiretroviral therapy.
Methods and material: We apply a temporal nodes Bayesian network (TNBN) model to data extracted from the Stanford HIV drug resistance database in order to explore the probabilistic relationships between drug resistance mutations and antiretroviral drugs unveiling possible mutational pathways and establishing their probabilistic-temporal sequence of appearance.
Results: In a first experiment, we compared the TNBN approach with other models such as static Bayesian networks, dynamic Bayesian networks and association rules. TNBN achieved a 64.2% sparser structure over the static network. In a second experiment, the TNBN model was applied to a dataset associating antiretroviral drugs with mutations developed under different antiretroviral regimes. The learned models captured previously described mutational pathways and associations between antiretroviral drugs and drug resistance mutations. Predictive accuracy reached 90.5%.
Conclusion: Our results suggest possible applications of TNBN for studying drug-mutation and mutation–mutation networks in the context of antiretroviral therapy, with direct impact on the clinical management of patients under antiretroviral therapy. This opens new horizons for predicting HIV mutational pathways in immune selection with relevance for antiretroviral drug development and therapy plan.

1. Introduction

Viral evolution is an important aspect of the epidemiology of viral diseases such as influenza, hepatitis and the acquired immunodeficiency syndrome (AIDS). This evolution greatly impacts the development of successful vaccines and antiviral drugs, as mutations conferring immune escape or drug resistance often develop early after the virus is placed under selective pressure. This is particularly relevant for human immunodeficiency virus (HIV), a virus ranking among the fastest evolving organisms in the planet [1]. The remarkable viral replication capability of HIV is coupled with a high mutation rate and a high probability of recombination in the viral genome during its replication cycle. These features allow HIV to boast a wide genetic variability even considering only the viral population within a given host. This variation capability gives the virus a remarkable ability to adapt to multiple selective pressures, including the immune response and antiretroviral therapy. Several questions remain open regarding HIV intra-host genetic variability, for example: To what extent do selective pressures such as immune responses and antiretroviral treatment shape viral evolution compared to genetic drift? What is the relationship between genetic diversity and clinical outcome? Is it feasible to anticipate HIV evolution in order to reduce drug resistance and viral adaptation to immune responses? In the case of this last question, being able

* Corresponding author. Tel.: +52 2222663100.
E-mail addresses: pablohl@ccc.inaoep.mx, hlealpablo@gmail.com (P. Hernandez-Leal), santiago.avila@cieni.org.mx (S. Ávila-Rios), esucar@inaoep.mx (L.E. Sucar).

0933-3657/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.artmed.2013.01.005

to predict mutational pathways in HIV and their timely patterns of appearance could help to anticipate viral adaptation and plan timely interventions with the proper monitoring of infected individuals. In the present study, we addressed the problem of finding mutation–mutation and drug–mutation associations in individuals receiving antiretroviral therapy, using temporal nodes Bayesian networks (TNBNs). We have focused on protease inhibitors (PIs), a family of antiretroviral drugs widely used in modern antiretroviral therapy. PIs block the activity of the viral protease, an enzyme that processes viral precursor polyproteins in order to promote viral maturation and infectivity. It is well described that resistance to PIs results from the accumulation of mutations that together modify the conformation of the enzyme.

Our goal was to test whether the model could predict previously described mutational pathways for specific drugs. Our probabilistic graphical model correctly identified antiretroviral drug-associated mutational patterns in the protease gene, revealing the co-occurrence of mutations and its temporal relationships. This work[1] is organized as follows. Section 2 highlights some basic notions regarding HIV and how it develops drug resistance. Section 3 justifies the use of TNBNs over other existing graphical probabilistic approaches. Section 4 describes the TNBN model. Section 5 presents the first experiments and compares TNBNs with other three approaches. Section 6 presents the results of a learned TNBN with specific drugs and important mutations; this section highlights the clinical relevance of the results. Finally, Section 7 summarizes the findings and indicates future lines of research.

2. HIV and its defense against antiretroviral therapy

2.1. Motivation

HIV is the causing agent of AIDS, a condition in which progressive failure of the immune system allows opportunistic life-threatening infections to occur. HIV is a virus with relatively recent introduction to human populations [2] representing a huge global burden to human health [3].

The HIV replication cycle is characterized by a reverse-transcription step of the viral ribonucleic acid genome to a double-stranded deoxyribonucleic acid molecule, which is then inserted into the host cell genome. To combat HIV infection several antiretroviral (ARV) drugs belonging to different drug classes that affect specific steps in the viral replication cycle have been developed. Antiretroviral therapy (ART) generally consists of well-defined combinations of three or four ARV drugs. Due to its remarkable variation capabilities, HIV can rapidly adapt to the selective pressure imposed by ART through the development of drug resistance mutations, that are fixed in the viral population within the host in known mutational pathways. The development of drug resistant viruses compromises HIV control, with a consequent further deterioration of the patient's immune system. Hence, there is interest in having a profound understanding of the dynamics of appearance of drug resistance mutations. We focus our analyses on one of the most common target proteins of antiretroviral drugs: the viral protease, an enzyme that cleaves viral polyprotein precursors into mature proteins. The protease acts late in HIV replication cycle and is essential for viral infectivity. Drugs designed to interfere with protease are known as protease inhibitors. Resistance to PIs occurs via the accumulation of primary and secondary mutations that are located within and outside of the enzyme active site, respectively [4]. The patterns of mutations selected by currently available PIs have been characterized [5,6]. More than 20% of the 99 amino acids comprising the HIV protease have been shown to mutate in response to drug pressure. Consequently, the genotypic patterns in patients with PI-resistant HIV are complex. The primary mutations selected by different PIs may be drug-specific, but secondary mutations tend to be common to several drugs in the PI class, potentially limiting the success of subsequent PI therapy following failure on any PI-containing regimen [7]. Moreover, the appearance of some drug resistance mutations implies high costs for viral replication capacity. These costs in viral replication capacity are often compensated by the appearance of additional mutations known as compensatory mutations. Due to their polymorphic nature the frequency of compensatory mutations can vary between viruses circulating in different geographic areas, making it relevant to study HIV mutational pathways in the context of different infected populations.

2.2. Related work

Abundant literature exists describing computational models aimed to better understand HIV evolution and immunopathogenesis. A good part of these models is devoted to predicting phenotypic HIV resistance to antiretroviral drugs using different approaches such as decision trees [8] or neural networks [9]. Other models have tried to identify relevant associations between clinical variables and the HIV disease [10]. Surprisingly, among this wealth of literature, works aimed towards the identification of temporal relationships among mutations and drugs in HIV are scarce.

In [11] association rules between clinical variables and ART failure were assessed. The authors used 15 clinical variables from 8000 patients from data collected since 1981. The results obtained were temporal rules that have as antecedent the increase of a subset of clinical variables and as consequent ART failure. None of the clinical variables considered were HIV mutations. Some other works using probabilistic models have studied the development of resistance to PIs, mainly Nelfinavir, Indinavir and Saquinavir, through learned Bayesian networks [12,13]. However, none of these learned models yielded any temporal information.

Finally, in [14], the order of appearance of resistance mutations in reverse transcriptase was assessed. In order to overcome the scarce amount of longitudinal data available for patients under the same antiretroviral regimen, the authors used large sets of cross-sectional data. By using a variant of the expectation-maximization algorithm they learned a mixture model of directed trees that accurately captured the order in which mutations of the HIV-1 reverse transcriptase accumulate.

In this work, we proposed a novel approach to study antiretroviral drug resistance mutation dependencies and temporal relations. We use temporal nodes Bayesian networks because of their capacity to obtain a global representation of the temporal dynamics of HIV mutations.

3. Bayesian networks

Information in clinical databases is often imprecise, incomplete, and with errors (noisy). Bayesian Networks (BNs) [15] are probabilistic graphical models particularly well suited to deal with uncertainty. BNs represent probabilistic dependencies among domain entities. BNs have a visual representation, a graph consisting of nodes and edges facilitating their analysis and interpretation. Nodes represent random variables and edges represent a set of conditional independence assumptions [16]. This kind of graphical representation can be easily understood. An additional advantage is the availability of several methods to learn BNs from data [17].

[1] This paper is an expanded version of "Unveiling HIV mutational networks associated to pharmacological selective pressure: a temporal Bayesian approach" presented in: A. Hommersom and P. Lucas, editors. Probabilistic Problem Solving in Biomedicine Workshop (ProBioMed-11), Bled Slovenia.

BNs have proven to be successful in various domains such as medicine [18,19] and bioinformatics [20–23]. However, *classical* BNs are not well equipped to deal with temporal information. Dynamic Bayesian networks (DBNs) evolved to tackle this shortcoming [24]. A DBN can be seen as multiple slices of a *static* BN over time, with temporal relations captured as links between adjacent slices. In a DBN, a base model is cloned for each time stage. These copies are linked via the so-called transition network. In this transition network is common that only links between consecutive stages are allowed. Whenever temporal changes occur infrequently, the DBN representation becomes unnecessarily over expressive. One alternative are temporal nodes Bayesian networks [25].

4. Temporal nodes Bayesian networks

In a TNBN, each node, known as *temporal node* (TN), represents a random variable that may be in a given state, i.e. value interval, throughout the different temporal intervals associated to it. An arc between two temporal nodes describes a temporal-probabilistic relation. In TNBNs, each variable (node) represents an event or state change. So, only one (or a few) instance(s) of each variable is (are) required, assuming there is one (or a few) change(s) of a variable state in the temporal range of interest. No copies of the model are needed, thus compacting the representation without losing expressiveness.

A TNBN [25,26] is composed by a set of TNs connected by arcs representing a probabilistic relationship between TNs. A TN, v_i, is a random variable characterized by a set of states **S**. Each state is defined by an ordered pair $S = (\lambda, \tau)$, where λ is the particular value taken by v_i during its associated interval $\tau = [a, b]$, corresponding to the time interval in which the state changes, i.e. change in value occurs. In addition, each TN contains an extra default state $s =$ ('Not', $\emptyset$) with no associated interval. Time is discretized in a finite number of intervals, allowing a different number and duration of intervals for each node (multiple granularity). Each interval, defined for a child node, represents the possible delays between the occurrence of one of its parent events and the corresponding child event. If a node lacks defined intervals for all its states then it is referred to as *instantaneous node*. There is at most one state change for each variable (TN) in the temporal range of interest.

Formally, let **V** be a set of temporal and instantaneous nodes and **E** a set of arcs between nodes, a TNBN is defined as:

Definition 1. A TNBN is a pair $B = (G, \Theta)$ where G is a directed acyclic graph, $G = (\mathbf{V}, \mathbf{E})$, and Θ is a set of parameters quantifying the network. Θ contains the values $\Theta_{v_i} = P(v_i|Pa(v_i))$ for each $v_i \in \mathbf{V}$; where $Pa(v_i)$ represents the set of parents of v_i in G.

The following is an example of a TNBN based on [25], its corresponding graphical representation is shown in Fig. 1. Each node in the TNBN has a conditional probability table. For the sake of visualization, we decided to present only the prior probabilities for each state of the node. They are calculated with the learned parameters and then propagating probabilities over the network.

Example 1. Assume that at time $t = 0$ an accident occurs, a *collision*. This kind of accident can be classified as severe, moderate and mild. For the sake of simplicity let us consider only two immediate consequences for the person involved in the collision, *head injury* and *internal bleeding*. *Head Injury* can take two values true or false, *internal bleeding* can be gross, slight or false. These events will generate subsequent changes that are not immediate: *dilated pupils* and unstable *vital signs*, that depend on the severity of the accident. These events (*dilated pupils* and *vital signs*) have temporal intervals associated. *Dilated pupils* has three intervals {[0–15] [15–30] [30–60]} and *vital signs* has two intervals {[0–10], [10–45]}. For this example, the intervals represent minutes and a physician would appreciate not only the information of the occurrence of the events but also the time at which they appear, in order to obtain a better evaluation/diagnosis of the person. For instance, if a collision is severe, the person involved has high probability of having a head injury and therefore to have dilated pupils in the first 15 min after the accident. On the other side, if a collision is mild, then the probability of having a head injury is low, thus the probability of having dilated pupils is extremely low.

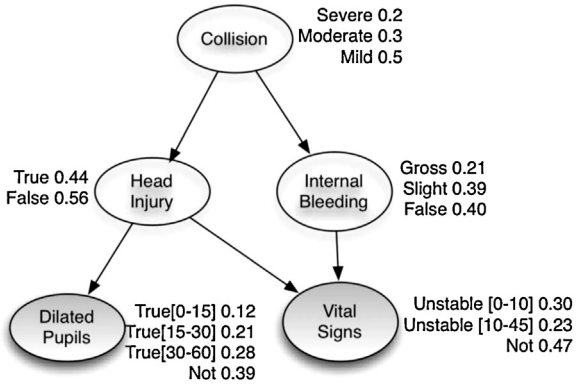

Fig. 1. An example of a TNBN. Each oval represents a random variable. All the nodes have prior probabilities associated with each state. The three upper nodes (`collision, head injury` and `internal bleeding`) are instantaneous nodes, so they do not have temporal intervals. The `dilated pupils` and `vital signs` are temporal nodes with intervals associated with them.

As mentioned, there are several methods to learn BNs from data [17]. Unfortunately, none of the algorithms used to learn BNs deal with learning temporal intervals. Therefore, these cannot be applied directly to learn TNBNs. The learning algorithm for TNBN used in this work has been presented in [27]. Briefly, the learning algorithm proceeds as follows:

1. First, it performs an initial discretization of the temporal variables, for example using an equal-width discretization. With this process it obtains an initial approximation of the intervals for all the temporal nodes.
2. Then it performs a standard BN structural learning, the K2 learning algorithm [28] is used to obtain an initial structure.
3. The interval learning algorithm refines the intervals for each TN by means of clustering. For this, it uses the information of the configurations of the parent nodes. To obtain the intervals a Gaussian mixture model is used as a clustering algorithm for the temporal data. Each cluster corresponds, in principle, to a temporal interval. The intervals are defined in terms of the mean and the standard deviation of the clusters. The algorithm obtains different sets of intervals that are merged and combined, this process generates different interval sets that will be evaluated in terms of the predictive accuracy (relative Brier score). The algorithm applies two pruning techniques in order to remove some sets of intervals that may not be useful and also to keep a low complexity of the TNBN. The best set of intervals (that may not be those obtained in the first step) for each TN is selected based on predictive accuracy. When a TN has as parents other temporal nodes (an example of this situation is illustrated in Fig. 4), the configurations of the parent nodes are not initially known. In order to solve this problem, the intervals are sequentially selected in a top-down fashion according to the TNBN structure.

The algorithm then iterates between structure learning and interval learning. However, for the experiments presented in this work, we show the results of only one iteration.

Table 1
An example of the data. Patient Pat_1 with 3 temporal studies, and patient Pat_2 with two temporal studies.

Patient	Initial treatment	List of mutations	Time (weeks)
Pat_1		L63P, L10I	15
	LPV, FPV, RTV	V77I	25
		I62V	50
Pat_2			
	NFV, RTV, SQV	L10I	25
		V77I	45

5. Finding pathways – a simple approach

This section presents a first experiment in which the objective is to compare the TNBN model with different approaches. Therefore, we used a simple approach for selecting the mutations and drugs to be included in the models. We present the learned TNBN and compare it to *static* Bayesian networks, dynamic Bayesian networks and association rules.

5.1. Data and preprocessing

Clinical data from 2373 patients with HIV subtype B was retrieved from the HIV Stanford database (HIVDB) [5,6]. We chose to work with this subtype because it is the most common in the Americas [29], our geographical region of interest. The isolates in the HIVDB were obtained from longitudinal treatment profiles reporting the evolution of mutations in individual sequences.

For each patient, data consisted of an initial treatment (a combination of drugs) administered to the patient and a list of laboratory resistance tests at different times (in weeks). Each test included a list of the most frequent mutations in the viral population within the host at a specific time after the initiation of treatment. An example of the data is presented in Table 1. The number of studies available varied from 1 to 10 studies per patient history. Since we are interested in the temporal evolution of the mutational networks, we filtered out those patients having only one study.

In order to apply the TNBN learning algorithm, the data exemplified in Table 1 was transformed into another table similar to the one presented in Table 2, where each column represents a drug or mutation, and each row represents a patient case. For the drugs, the values were: `used` or `not used`, and for the mutations the values were: `appear`, with the number of weeks elapsed before the mutation appeared for the first time; or `not`, when the mutation did not appear in that case.

Antiretrovirals are usually classified according to the enzyme that they target. We focused on the viral protease, as this is the smallest of the viral enzymes in terms of number of amino acids. Nine protease inhibitors are currently available, namely: Amprenavir (APV), Atazanavir (ATV), Darunavir (DRV), Lopinavir (LPV), Indinavir (IDV), Nelfinavir (NFV), Ritonavir (RTV), Tripanavir (TPV) and Saquinavir (SQV). All 9 PIs were considered during this experiment. Fig. 2 presents a histogram of the frequency of administration of each PI in the individuals included in the dataset. Data retrieved belong to a period of 10 years, from 1995 to 2005. Since data from HIVDB originates from different studies, it was not rare to find some fields with missing data. Fig. 2 evidences a small portion of cases reporting the administration of a PI, without describing the specific drug. Also, there was a small proportion of cases reporting an *unknown* drug. To handle this missing data, if the patient case only contained *unknown* or *none* in the drug fields that case was removed. However, if the case contained other drug (apart from *unknown*), that information was included in the model. After removing the incomplete cases and the cases without temporal information we ended up with a final set of 973 patients.

Table 2
An example of the data used to learn the TNBN model. Each row represents a patient. Each column represents whether a drug is used or not, or whether a mutation appeared or not.

Drug-1	Drug-2	Drug-3	...	Mutation-1	Mutation-2	...
LPV	**FPV**	**NFV**	...	**L63P**	**M36I**	...
Used	Used	Not used	...	(Appear) *15*	Not	

In order to define a target set of relevant mutations, we used a simple frequency approach. Fig. 3 shows a frequency histogram of mutations appearing in the original dataset. A total of 733 different mutations appeared at least once in the data; however, most of the mutations were rare. Thus, we only considered mutations appearing more than 1500 times in the dataset: **L63P, I93L, V77I, I62V, L10I, E35D, L90M, M36I, A71V and R41K** (shown in black columns in Fig. 3).

The order of variables provided to the K2 algorithm was also determined by frequency: first the antiretrovirals sorted by frequency, then the selected mutations sorted by frequency.

5.2. Model evaluation and results

Since a gold standard or a reference TNBN does not exist, three indirect measurements were used for the evaluation of the model: the relative Brier score (RBS), the relative time error and the total number of intervals in the model. For learning, 80% of the total 973 patients was used. The remaining 20% was used for evaluation.

The Brier score is a measure of the predictive accuracy of the network, and it is defined as:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(1 - P_i)^2$$

where P_i is the marginal posterior probability of the correct value of each node given the evidence, this applies for all the selected nodes, n, of the TNBN. The RBS is defined as:

$$\text{RBS}(in\%) = (1 - BS) \times 100$$

For each case of the data (a row in Table 1), the RBS is obtained by instantiating a random subset of variables in the model, predicting the unseen variables, and obtaining the RBS for these predictions.

The relative temporal error (RTE) evaluates the temporal part of the model. First, we define the expected time as:

$$t_e = \frac{t_{end} + t_{ini}}{2}$$

where t_{ini} and t_{end} are the initial and final values of the interval, so the expected time is the average of them. The range of a temporal node T is the difference between the maximum and the minimum value of all the intervals in the node. The relative temporal error for a Temporal node T with respect to the original value t_{orig} is defined as:

$$\text{RTE} = \frac{|t_e - t_{orig}|}{range(T)}$$

this is the difference between the real event (original data) and the expected mean of the interval, normalized by the range of T.

Finally, the number of intervals is defined as the total number of intervals learned across all variables. This is a rough estimate of the complexity of the network and a low number of intervals is a desirable property for simplicity of the model.

The best model would afford a high RBS, a low time error and a low complexity (low number of intervals). The technical performance of the model reflects its predictive accuracy and complexity, but it should not be confused with the biological/physiological plausibility of the model.

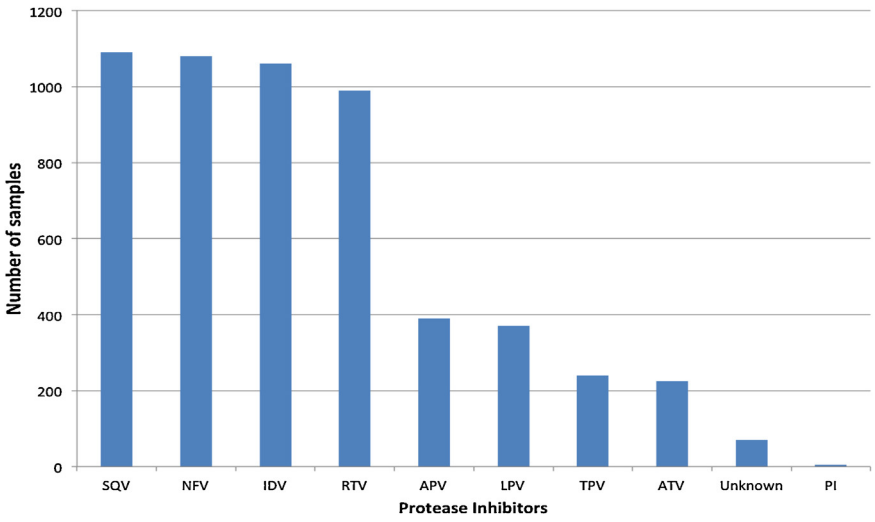

Fig. 2. Histogram of the protease inhibitors administration in the full dataset including all 2373 patients.

The learning algorithm finds a local maximum, and thus is influenced by initial parametrization. Thus, for our experiments we explored different initializations; the number of initial intervals was allowed to vary from 2 to 4 and equal-width discretization [30] was used to initialize those intervals.

Table 3 summarizes the results of the experiments and Fig. 4 illustrates the best TNBN model instantiation in terms of higher RBS for the experiment. The figure represents the network, the intervals and the prior probabilities obtained for each TN.

Results show that for different initial intervals there is a small variation in the three measures used. Note that all models obtained competitive predictive scores of nearly 90%.

Statistical significance of edge strengths can be evaluated using bootstrapping [23]. In non-parametric bootstrapping, a re-shuffled dataset is generated from the original (re-sampling with replacement), the graph is built from this new dataset and the procedure is repeated a number of times. Confidence in a particular edge is measured as a percentage of the number of times that edge actually appears in the set of reconstructed graphs. We performed 20-fold

Table 3
Evaluation of the models in terms of RBS, RTE (in percentage), and number of intervals. The best results are shown in boldface type.

Initial intervals	RBS	RTE	Number of total intervals
2	87.3	15.0	**30**
3	**88.5**	**14.7**	31
4	87.5	15.9	35

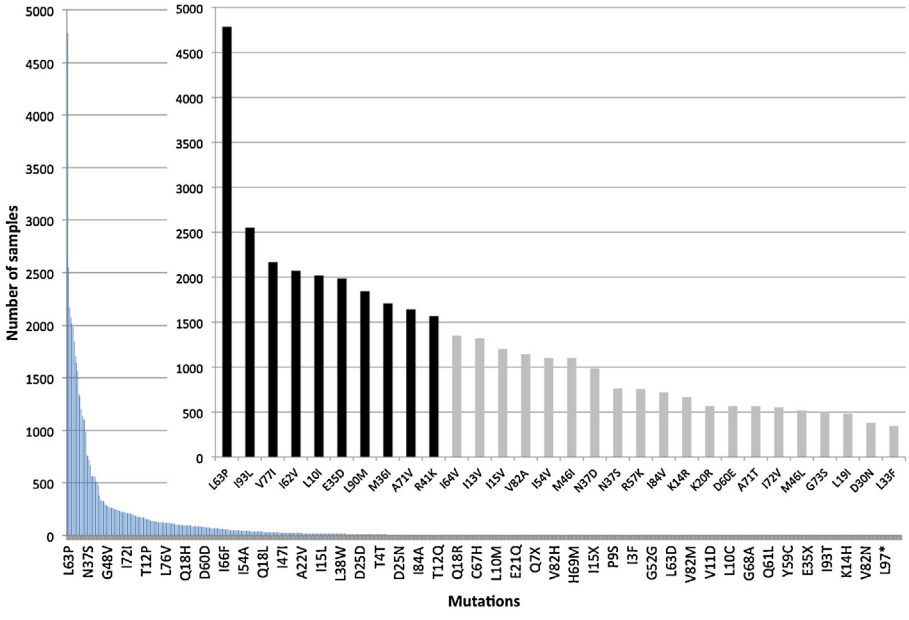

Fig. 3. A group of mutations and their frequency in the full dataset including all 2373 patients. The higher frequency end of the histogram is zoomed. Mutations used for the experiment are shown in black.

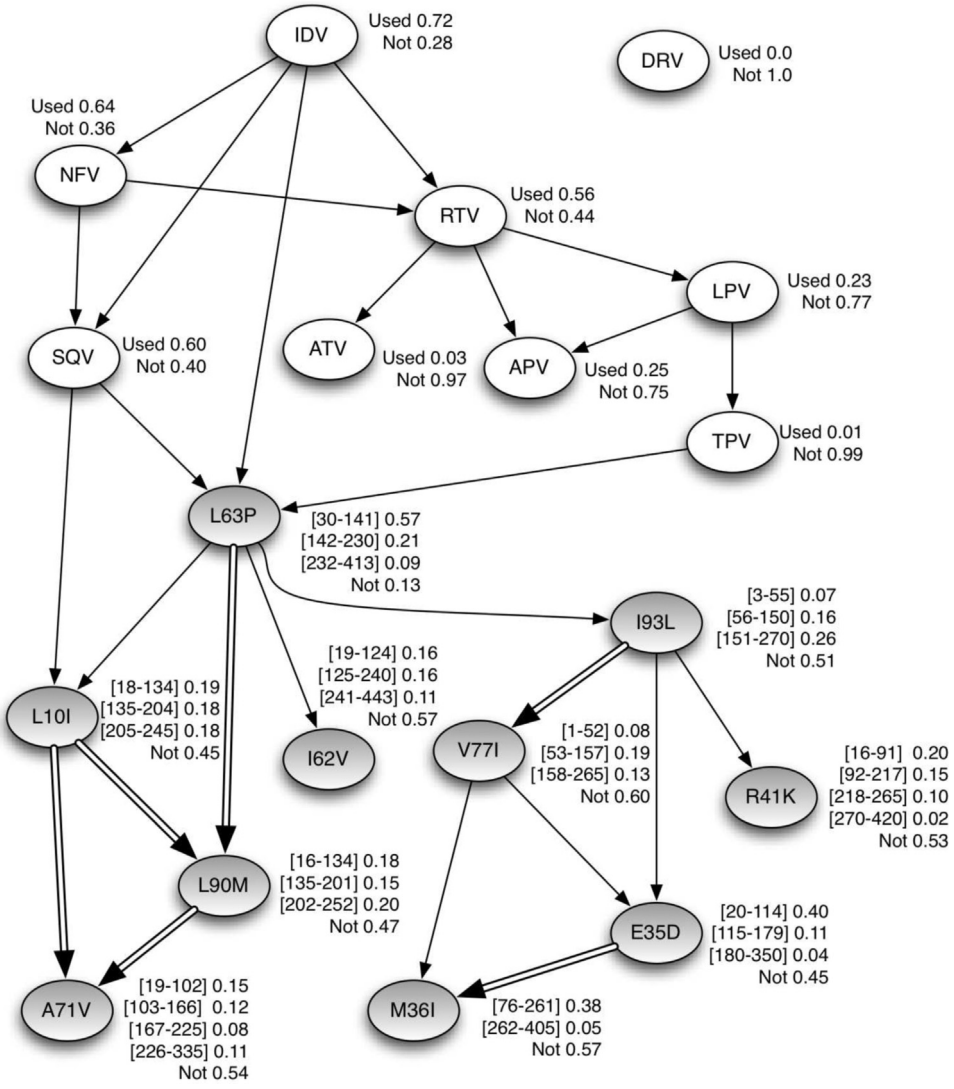

Fig. 4. A learned TNBN with 9 protease inhibitors and 10 high frequency mutations. Drugs are indicated in white ovals and mutations in gray ovals. Learned time intervals for the TNs are indicated beside the ovals. Thicker arrows represent highly correlated mutations [31].

bootstrapping and used two thresholds for classifying the significance of the observed relations. A strong relation was defined as that appearing at least in 90% of the graphs, a suggestive relation was defined as that observed between bootstrap values of 75% and 90%.

Fig. 5 presents the identified relations using bootstrapping. Remarkably, in the set of the suggestive and strong relationships between mutations, we observe five previously described highly correlated mutation pairs [31]: (E35D, M36I), (L10I, L90M), (L10I, A71V), (V77I, I93L), and (L63P, L90M). This is an important result that shows that our model is able to capture important relations from data.

5.3. Clinical analysis of the model

Some clinical interpretations can be drawn from the associations obtained by the model depicted in Fig. 4. For example, the model revealed the linking of RTV with IDV, NFV, ATV, APV and LPV. This relationship of RTV with other drugs can be explained due to the fact that RTV is widely used to boost the effect of other PIs, and therefore most of the times it is administered in combination with other drugs.

Another noteworthy observation is that the DRV node in the model is isolated because, in the data, this drug was never given as part of a first treatment regimen for any of the patients. This can be expected as DRV is a relatively new drug and its use is mostly restricted to salvage regimens.

Interestingly, the local neighborhoods in the graph clearly revealed two clusters of covarying mutations:

* L63P, I62V, L10I, L90M and A71V
* I93L, V77L, M36I, E35D and R41K

Highly significant associations between these groups of mutations have been previously observed using different models [31]. This observation suggests that our model can predict highly significant patterns of amino acid covariation.

L63P, a polymorphic mutation also selected by PIs, is a highly frequent mutation in the viruses included in the dataset (Fig. 3). Our model suggests that in most cases this mutation tends to appear early in time, and that its probability to appear decreases over time. Alternatively, L63P could represent a common polymorphism in the circulating viruses appearing in a large proportion of viruses even before being exposed to PIs. Interestingly, however,

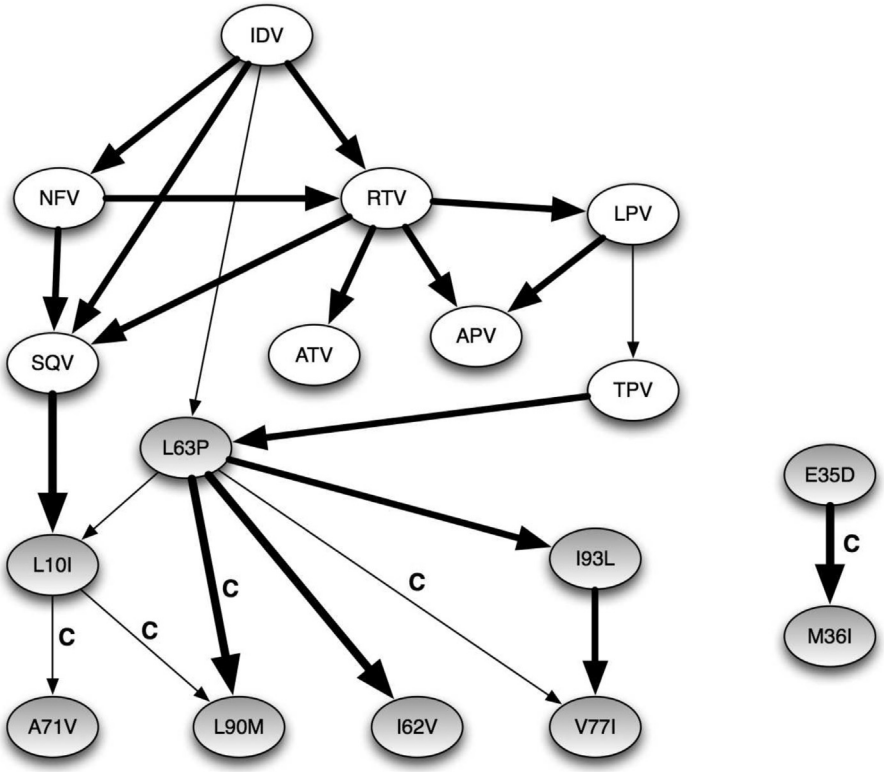

Fig. 5. A learned TNBN with bootstrapping. The patterns that appeared ≥90% are shown in a thick arrows, those that obtained ≥75% but <90% are shown in thin arrows. The letter C' in the arrows represents highly correlated mutations [31].

the model was able to predict previously described covariation of L63P with other PI-selected mutations such as L90M, L10I, A71V and I62V [31], suggesting the existence of drug-selected mutational pathways. This property of the model could make it a powerful tool to detect not only drug-associated codon covariation, but also immune-associated codon covariation. The detection of mutational pathways associated with adaptation to frequent immune responses in a given population would be relevant for vaccine research and immunogen design.

5.4. Comparison with other methods

This section, compares and contrasts the results obtained using the proposed TNBN model with three other approaches, namely: static Bayesian networks, dynamic Bayesian networks and association rules.

5.4.1. Static Bayesian network

To establish whether the temporal model is providing additional information we develop a comparison between a TNBN and a static Bayesian network. In order to apply a static learning algorithm, we remove the temporal information of the mutations. Hence, the states for both the drugs and the mutations were: `appear` or `not appear`. We used the same ordering as in the previous experiments and applied the K2 learning algorithm. The static BN learned is presented in Fig. 6.

From the model we can see that the number of arcs increased approximately in 66%. However, in this model the nodes are binary, in contrast in the TNBN model the nodes have more states that correspond to the intervals. We performed a comparison in terms of the number of values that contain the conditional probability tables for these models. For the TNBN there are 432 values whereas in the static BN there are 276 values. Despite the TNBN is not simpler in terms of parameters, the model can be intuitively understood.

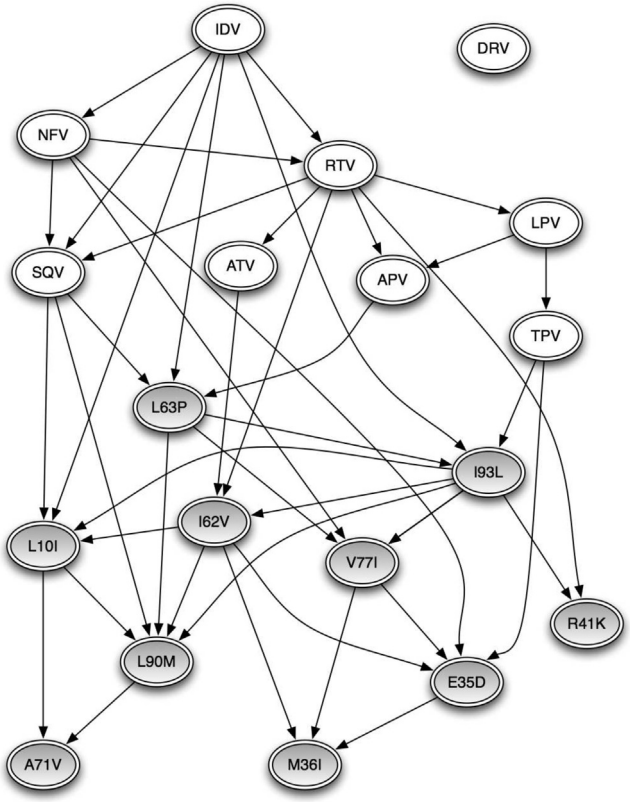

Fig. 6. A learned *static* BN with 9 protease inhibitors and 10 mutations that appear frequently. Drugs are indicated in white ovals, their states are `used` or `not used`. Mutations are shown in gray ovals, their states are `appear` or `not appear`. States and probabilities are hidden for readability.

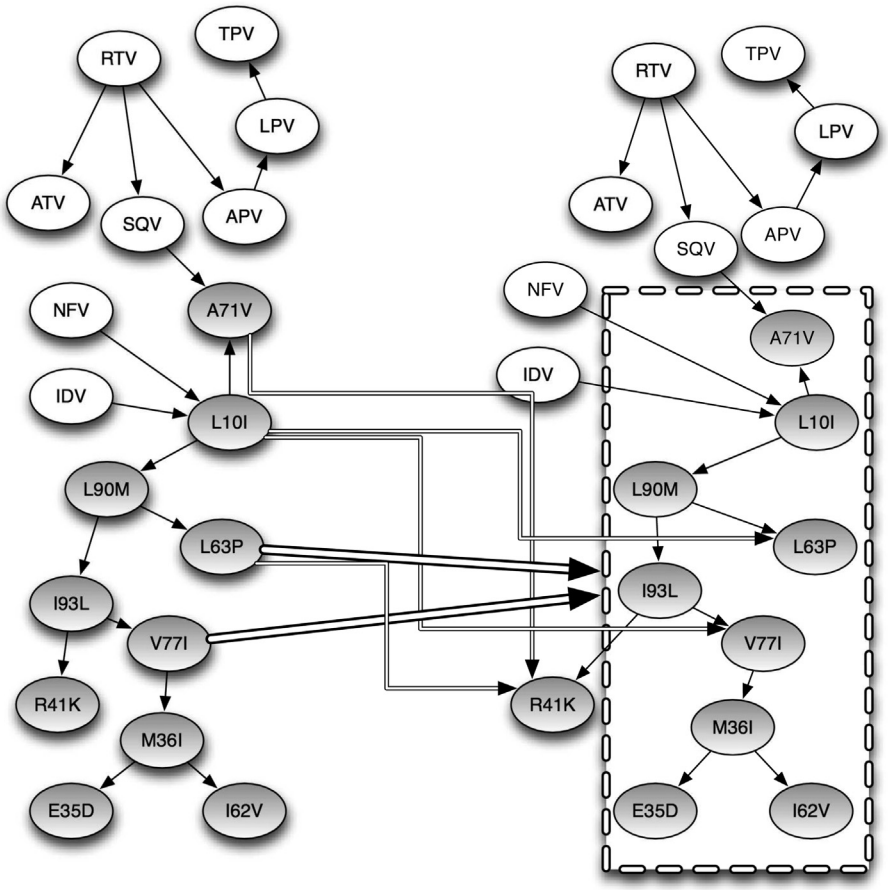

Fig. 7. A learned DBN using the same data. Persistence arcs (between the same variable in different time slices) are not shown. Thicker arrows from nodes V77I and L63P directed to the dotted rectangle represent arrows from V77I and L63P to each one of the nodes inside the rectangle.

We computed the RBS for this static BN obtaining a value of 91.3 compared to 87.3, 88.5 and 87.5 for the learned TNBNs. According to the Brier score the static BN is slightly superior (about 3 points) to the TNBN. However, we consider that this comparison is not strictly fair as the static BN only predicts the state (appear/not appear) whereas the TNBN provides additional information, not only the state but also the temporal interval when it changes.

We also performed a structural comparison. In this sense, a common measure of complexity in graph theory is density [32] defined as

$$D = \frac{|E|}{(2)|V||V-1|}$$

where $|V|$ corresponds to the number of vertices and $|E|$ to the number of edges of the graph. The maximal density is 1 for complete graphs, when the graph is fully connected, and the minimal is 0 when all nodes are isolated. This measure was computed for the TNBN in Fig. 4 and the *static* BN. The TNBN obtained $D = 0.1578$ and the BN obtained $D = 0.2456$. This measure quantitatively summarizes that the TNBN model is sparser than the respective BN, which in general implies that is easier to understand.

5.4.2. Dynamic Bayesian network

A dynamic Bayesian network extends the concept of a Bayesian network to incorporate temporal information. Just as with static BNs, a probabilistic model is created to represent a process at a single point in time. Multiple copies of this model are then generated for each time point or *slice* belonging to a temporal range of interest. Links between copies are inserted to capture temporal relations.

The learning of a DBN can be seen as a two stage process. The first stage refers to the learning of the static model and is done in an identical manner as with classic BNs. The second stage learns the transition network, that is, the temporal relations between random variables of different time slices.

For this experiment we learned a simple DBN using the Chow–Liu algorithm [33] to generate the static network, and the Rebane and Pearl algorithm [34] to learn the directions of the arcs. The transition network was learned using Kevin Murphy's Bayesian network toolbox [35] using the Bayesian information criterion to select the best parents from the previous time slice. Fig. 7 shows the learned DBN as a 2-TBN. The learned DBN shows some common relations with the TNBN; it also discovered the same well known mutational correlations that the TNBN discovered (although not always in the same direction). While the DBN successfully captured most of the relations shown in the TNBN, the DBN notoriously becomes more cluttered than the TNBN, impairing its visual interpretation. We also note that the DBN lacks the intuitive visual information to provide temporal orderings between mutations.

5.4.3. Association rules

Finally, a different approach that is not related to BNs is presented. The comparison uses association rules as the method to analyze the clinical information. In particular we used the *a priori* algorithm [36] with the same data used in the previous experiment.

Different measures have been introduced to evaluate the quality of the rules obtained by the algorithm, we used three common measures: confidence [36], lift [37], and conviction [37]. In order to define these measures, first we define the support of an itemset

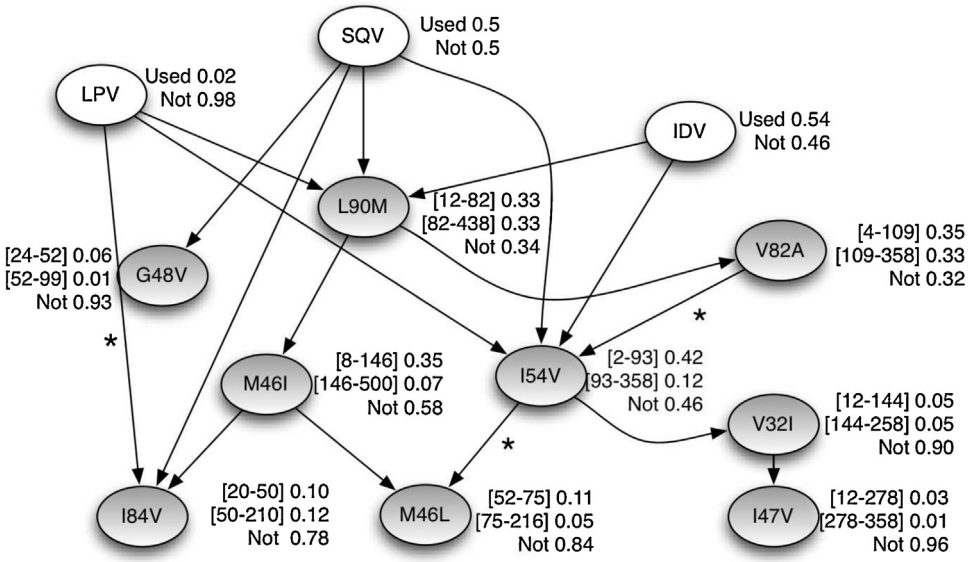

Fig. 8. A learned TNBN model. The intervals are expressed in weeks. An arc labeled with a * represents a strong relation. An arc without a * represents a suggestive relation.

supp(X) as the proportion of instances which contains that itemset, then:

- Confidence is defined as $conf(X \rightarrow Y) = supp(X \cup Y)/supp(Y)$ and can be interpreted as the probability of $P(Y|X)$.
- Lift is defined as $lift(X \rightarrow Y) = supp(X \cup Y)/supp(Y) \times supp(X)$, it can be interpreted as the ratio of the observed support of both X and Y ($X \cup Y$) divided by X and Y independently.
- Conviction is defined as $conv(X \rightarrow Y) = 1 - supp(Y)/1 - conf(X - > Y)$ and can be interpreted as the ratio of the expected frequency that X occurs without Y (the frequency of making an incorrect prediction).

In the upper part of Table 4 we present rules that obtained confidence over 0.9, in the lower part of the table, we present some rules that are consistent with the findings of our TNBN model with their respective results.

Table 4
Some association rules obtained with the *a priori* algorithm. On the upper part the rules that obtained confidence over 0.9 are presented. On the lower part, several rules that are consistent with the results of our TNBN model.

Rule	Confidence	Lift	Conviction
L10I = yes L90M = yes → L63P = yes	0.93	1.11	2.2
IDV = yes L90M = yes → L63P = yes	0.92	1.1	2
L90M = yes → L63P = yes	0.91	1.1	1.91
NFV = yes L90M = yes → L63P = yes	0.91	1.09	1.81
RTV = yes L90M = yes → L63P = yes	0.91	1.09	1.72
SQV = yes L90M = yes → L63P = yes	0.9	1.08	1.63
I62V = yes → L63P = yes	0.87	1.04	1.23
L63P = yes I93L = yes → IDV = yes	0.79	1.1	1.33
L63P = yes L10I = yes → L90M = yes	0.74	1.35	1.72
SQV = yes L63P = yes → L90M = yes	0.73	1.35	1.7
I93L = yes → IDV = yes L63P = yes	0.69	1.13	1.24
A71V = yes → L10I = yes	0.68	1.3	1.47
SQV = yes L63P = yes → L10I = yes	0.65	1.25	1.36
IDV = yes L63P = yes → L90M = yes	0.62	1.14	1.19
IDV = yes L63P = yes → L10I = yes	0.61	1.17	1.22
A71V = yes → L90M = yes	0.60	1.25	1.41
L10I = yes → A71V = yes	0.59	1.3	1.32
L90M = yes → A71V = yes	0.56	1.25	1.25
IDV = yes L63P = yes → I93L = yes	0.56	1.13	1.14
L63P = yes → L10I = yes	0.54	1.03	1.03

While association rules could obtain similar results to the TNBN model, they have several drawbacks. The first one is the difficulty of incorporating temporal information. One could think to separate each interval of the temporal nodes into a different variable and apply the same *a priori* algorithm. A limitation of this approach is that the number of variables will increases and their frequency decreases by means of the partition process.

Another issue with this approach is that it produces a large number of small rules. The number of rules could increment exponentially. Therefore analyzing the complete set of rules becomes extremely difficult. Finally, if we could select a manageable subset of rules it will not provide a global analysis of the results, they will be scattered pieces of information. In contrast, the TNBN model presents a global and easy to understand model that can provide useful temporal information.

6. Finding mutational pathways with clinical relevance

In order to asses the practical relevance of the TNBN model, a second experiment was designed using a specific set of drugs and drug resistance mutations that were selected according to the real use of PIs in the clinical setting.

6.1. Data

To test the ability of the model to predict clinically relevant data, a subset of patients from the original dataset was selected, including individuals that received ART regimes with LPV, IDV, and SQV. Relevant major drug resistance mutations associated with the selected drugs were included [36]. The mutations selected were: V32I, M46I, M46L, I47V, G48V, I54V, V82A, I84V, and L90M. Since we used a subset of drugs, the number of patients in the final dataset was reduced from the previous experiments to 300 patients.

6.2. Evaluation of the model and results

In order to evaluate the models and to measure the statistical significance of edge strengths we used non-parametric bootstrapping. Two thresholds were defined for considering a relation as important. A strong relation was defined as one that appeared at least in 90% of the graphs, and a suggestive relation was defined as one that occurred with values between 70% and 90%. In Fig. 8 a

suggestive relation is shown as an arrow, and a strong relation is presented as an arrow labeled with an asterisk (*).

Since the approach for selecting the drugs and mutations is based on experts opinion, we used a more elaborate way to obtain the order for the K2 algorithm. For this experiment we evaluated different orderings for the K2 algorithm. Specifically, we evaluated all the $\binom{9}{2}$ different combinations for the first two mutations, since they are important and the order of the rest was chosen randomly. We evaluated the models with respect to their predictive accuracy and in Fig. 8 the model with highest predictive accuracy (90.5%) is presented.

6.3. Clinical relevance

The model was able to predict clinically relevant associations between the chosen drugs and mutations. Indeed, a strong association between SQV, G48V, and I84V was readily predicted in the model, although no temporal associations were observed between the two mutations. All three drugs showed direct associations with L90M reflecting the fact that this mutation causes cross-resistance to many members of the PI family. Remarkably, the two possible mutational pathways for LPV resistance [38–40] were predicted:

- I54V → V32I → I47V
- L90M → M46IL → I84V

Whether the temporal order of mutations is clinically relevant, still needs to be further evaluated. Also, the shared mutational pathway between IDV and LPV was observed, involving mutations L90M, M46IL, I54V, V82A and I84V.

7. Conclusions

Mutational pathways provide important information for decision making in multi-drug therapy. By using temporal nodes Bayesian networks we have been able to unveil common mutational pathways present in HIV evolution as response to pharmacological selective pressure. Our model was successful in capturing relationships between mutations and protease inhibitors critically incorporating temporal information. These results are encouraging, presenting the model as an interesting tool to explain how mutations interact with each other, providing information not only in association patterns, but also in the temporal order of appearance of mutations selected by antiretroviral drugs. With a carefully selected dataset, the model could be also useful to predict timeframes between the appearance of different mutations. This could lead to recommendations in the timing of resistance testing and therapy changes in order to avoid the further loss of possible treatment options for patients with exposure to multiple antiretroviral regimens.

The main contribution of this paper is to use a temporal probabilistic approach to understand HIV mutations. The models were developed using only data, however, some important known correlated mutations were discovered, as well as other temporal relations. We also compared the TNBN approach with other models such as Bayesian networks and association rules. Despite the compared models obtained important information they were not capable of providing a global and complete model showing the temporal process of the problem.

It would be interesting to test TNBN with data involving complete antiretroviral regimens. This could be useful in retrospectively comparing the effectiveness of different regimes, predicting the patterns of mutations most frequently selected by similar regimens and the timing of appearance of these mutations in specific settings. Resistance patterns in specific target populations under different scenarios could be further assessed. If sufficient data becomes available, social factors including adherence to treatment and drug availability, as well as genetic factors modulating the pharmacodynamics could be incorporated to the model.

The relevance of TNBNs could be further extended to the vaccine field. It is well known that there is remarkable coevolution of positions in the viral genome, which depends on host genetic factors defining specific immune responses against the virus. These genetic factors, namely the human leukocyte antigen genes are highly polymorphic, with highly variable allelic frequency distributions in different populations. The possibility of discovering these coevolution patterns in the context of different populations, as well as the timing of appearance of different variants in the coevolving codons, could be a very useful tool in predicting immunogen responses and designing putative vaccines.

Future work plans are to compare two different cocktail treatments along with the temporal occurrence of drug resistant mutations, in order to predict the most effective treatment. We believe this could aid the experts in the selection of the best treatment for the patient.

Acknowledgements

This work was partly supported by the European Union and the National Council of Science and Technology of Mexico (CONACYT) under the project FONCICYT 95185. The first author is supported by a scholarship grant 234507 from CONACYT.

References

[1] Freeman S, Herron JC. Evolutionary analysis. 4th ed. Upper Saddle River, NJ: Prentice Hall; 2007.
[2] Rambaut A, Posada D, Crandall KA, Holmes EC. The causes and consequences of HIV evolution. Nature Reviews Genetics 2004;5(1):52–61.
[3] Joint United Nations Programme on HIV/AIDS (UNAIDS). Global report: UNAIDS report on the global AIDS epidemic. Geneva, Switzerland: World Health Organization; 2010.
[4] Condra J, Holder D, Schleif W, Blahy O, Danovich R, Gabryelski L, et al. Genetic correlates of in vivo viral resistance to Indinavir a human immunodeficiency virus type 1 protease inhibitor. Journal of Virology 1996;70(12):8270–6.
[5] Rhee S, Gonzales M, Kantor R, Betts B, Ravela J, Shafer R. Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Research 2003;31(1):298–303.
[6] Shafer R. Rationale and uses of a public HIV drug-resistance database. Journal of Infectious Diseases 2006;194(Suppl. 1):S51–8.
[7] Mo H, King M, King K, Molla A, Brun S, Kempf D. Selection of resistance in protease inhibitor experienced, human immunodeficiency virus type 1 infected subjects failing Lopinavir and Ritonavir based therapy: mutation patterns and baseline correlates. Journal of Virology 2005;79(6):3329–38.
[8] Beerenwinkel N, Schmidt B, Walter H, Kaiser R, Lengauer T, Hoffmann D, et al. Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. Proceedings of the National Academy of Sciences of the United States of America 2002;99(12):8271–6.
[9] Draghici S, Potter RB. Predicting HIV drug resistance with neural networks. Bioinformatics 2003;19(1):98–107.
[10] Ramirez J, Cook D, Peterson L, Peterson D. Temporal pattern discovery in course-of-disease data. Engineering in Medicine and Miology Magazine, IEEE 2000;19(4):63–71.
[11] Chausa P, Cáceres C, Sacchi L, León A, Garcí a F, Bellazzi R, et al. Temporal data mining of HIV registries: results from a 25 years follow-up. In: Combi C, Shahar Y, Abu-Hanna A, editors. Proceedings of the 12th conference on artificial intelligence in medicine. Verona, Italy: Springer; 2009. p. 56–60.
[12] Deforche K, Camacho R, Grossman Z, Silander T, Soares M, Moreau Y, et al. Bayesian network analysis of resistance pathways against HIV-1 protease inhibitors. Infection, Genetics and Evolution 2007;7(3):382–90.
[13] Deforche K, Silander T, Camacho R, Grossman Z, Soares M, Van Laethem K, et al. Analysis of HIV-1 pol sequences using Bayesian networks: implications for drug resistance. Bioinformatics 2006;22(24):2975–9.
[14] Beerenwinkel N, Rahnenführer J, Däumer M, Hoffmann D, Kaiser R, Selbig J, et al. Learning multiple evolutionary pathways from cross-sectional data. In: Bourne PE, editor. Proceedings of the eighth annual international conference on research in computational molecular biology. San Diego, CA, USA: ACM; 2004. p. 36–44.
[15] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. San Francisco, CA, USA: Morgan Kaufmann; 1988.

[16] Koller D, Friedman N. Probabilistic graphical models: principles and techniques. Cambridge, MA, USA: The MIT Press; 2009.
[17] Neapolitan R. Learning Bayesian networks. Upper Saddle River, NJ: Pearson Prentice Hall; 2004.
[18] Beinlich I, Suermondt H, Chavez R, Cooper G. The ALARM monitoring system: a case study with two probabilistic inference techniques for belief networks. In: Hunter J, Cookson J, Wyatt J, editors. Proceedings of the second European conference on artificial intelligence in medicine. 1989. p. 247–56.
[19] Lucas P, van der Gaag L, Abu-Hanna A. Bayesian networks in biomedicine and health-care. Artificial Intelligence in Medicine 2004;30(3):201–14.
[20] Neapolitan R. Probabilistic methods for bionformatics: with an introduction to Bayesian networks. Upper Saddle River, NJ: Morgan Kaufmann; 2009.
[21] Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. Journal of Computational Biology 2000;7(3-4):601–20.
[22] Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan N, Chung S, et al. A Bayesian networks approach for predicting protein–protein interactions from genomic data. Science 2003;302(5644):449–53.
[23] Rodin A, Boerwinkle E. Mining genetic epidemiology data with Bayesian networks: Bayesian networks and example application (plasma apoE levels). Bioinformatics 2005;21(15):3273–8.
[24] Dagum P, Galper A, Horvitz E. Dynamic network models for forecasting. In: Dubois D, Wellman M, editors. Proceedings of the eighth international conference on uncertainty in artificial intelligence. 1992. p. 41–8.
[25] Arroyo-Figueroa G, Sucar LE. A temporal Bayesian network for diagnosis and prediction. In: Laskey KB, Prade H, editors. Proceedings of the 15th uncertainty in artificial intelligence conference. 1999. p. 13–22.
[26] Galán S, Arroyo-Figueroa G, Dí ez F, Sucar L. Comparison of two types of event Bayesian networks: a case study. Applied Artificial Intelligence 2007;21(3):185.
[27] Hernandez-Leal P, Sucar LE, Gonzalez JA. Learning temporal nodes Bayesian networks. In: Murray RC, McCarthy PM, editors. The 24th Florida Artificial Intelligence Research Society Conference (FLAIRS-24). Palm Beach, FL, USA: AAAI Press; 2011. p. 608–13.
[28] Cooper G, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Machine Learning 1992;9(4):309–47.
[29] Hemelaara J, Gouws E, Ghys PD, Osmanov S. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. AIDS 2006;20: W13–23.
[30] Liu H, Hussain F, Tan C, Dash M. Discretization: an enabling technique. Data Mining and Knowledge Discovery 2002;6(4):393–423.
[31] Rhee S, Liu T, Holmes S, Shafer R. HIV-1 subtype B protease and reverse transcriptase amino acid covariation. PLoS Computational Biology 2007;3(5): e87.
[32] Coleman T, Moré J. Estimation of sparse Jacobian matrices and graph coloring problems. SIAM Journal on Numerical Analysis 1983;20(1):187–209.
[33] Chow C, Liu C. Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory 1968;14(3):462–7.
[34] Rebane G, Pearl J. The recovery of causal poly-trees from statistical data. In: Kanal LN, Levitt TS, Lemmer JF, editors. Third conference on uncertainty in artificial intelligence, vol. 87. 1987. p. 222–8.
[35] Murphy K. The bayes net toolbox for matlab. Computing Science and Statistics 2001;33(2):1024–34.
[36] Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Bocca JBBMJJB, Jarke M, Zaniolo C, editors. Proceedings 20th international conference on Very Large Data Bases, VLDB, vol. 1215. 1994. p. 487–99.
[37] Brin S, Motwani R, Ullman J, Tsur S. Dynamic itemset counting and implication rules for market basket data. In: Peckham J, editor. Proceedings of the 1997 ACM SIGMOD international conference on management of data, vol. 26. Tucson, Arizona: ACM; 1997. p. 255–64.
[38] Nijhuis M, Wensing A, Bierman W, De Jong D, van Rooyen W, Kagan R, et al. A novel genetic pathway involving L76V and M46I leading to Lopinavir/r resistance. In: Boucher C, Mellors J, editors. 16th international HIV drug resistance workshop, vol. 12. 2007. p. 140.
[39] Parkin N, Chappey C, Petropoulos C. Improving Lopinavir genotype algorithm through phenotype correlations: novel mutation patterns and Amprenavir cross-resistance. AIDS 2003;17(7):955.
[40] Prado J, Wrin T, Beauchaine J, Ruiz L, Petropoulos C, Frost S, et al. Amprenavir-resistant HIV-1 exhibits Lopinavir cross-resistance and reduced replication capacity. AIDS 2002;16(7):1009.