

Using a temporal Bayesian approach in order to find mutational networks in HIV

Pablo Hernandez-Leal^{a,*}, Alma Rios-Flores^a, Santiago Ávila-Rios^b,
Gustavo Reyes-Terán^b, Jesus A. González^a, Felipe Orihuela-Espina^a,
Eduardo F. Morales^a, L. Enrique Sucar^a

^a*National Institute of Astrophysics, Optics and Electronics.
Computer Science Department.*

Sta. María Tonantzintla, Puebla, Mexico

^b*Infectious Diseases Research Center at the National Institute of Respiratory Diseases
(CIENI-INER)
Mexico City, Mexico*

Abstract

Much of the HIV (Human Immunodeficiency Virus) success is due to its fast evolution which makes it drug resistant. Understanding viral evolution and its relation to pharmacology is of utmost importance in fighting diseases caused by the HIV. Although the mutations conferring drug resistance are mostly known, the dynamics of the appearance chain of those mutations remains poorly understood. Here, we apply a Temporal Nodes Bayesian Network (TNBN) to data extracted from the HIV Stanford database to explore the probabilistic relationships between mutations and antiretrovirals. We aim to unveil existing mutation networks and establish their probabilistic formation sequence. The model predictive accuracy is evaluated in terms of the relative Brier score, relative temporal error, and total number of intervals. We developed models for the protease, an important protein of the HIV. The learned models captured well known relationships, which qualitatively provide a validity measure. In our experiments, two previously unknown mutational networks were retrieved and their probabilistic temporal sequentiation uncovered. With our results, we demonstrate the usefulness of TNBN for studying drug-mutation and mutation-mutation networks. We

*Corresponding author

Email addresses: pablohl@ccc.inaoep.mx (Pablo Hernandez-Leal),
santiago.avila@cieni.org.mx (Santiago Ávila-Rios), jagonzalez@inaoep.mx (Jesus A. González), emorales@inaoep.mx (Eduardo F. Morales), esucar@inaoep.mx (L. Enrique Sucar)

expect to use our models to impact the combat against HIV infection by facilitating a better treatment planning.

Keywords: Bayesian Networks, Temporal Reasoning, Learning, HIV

1. Introduction

Viral evolution is an important aspect of the epidemiology of viral diseases such as influenza, hepatitis and human immunodeficiency virus (HIV). This evolution greatly impacts the development of successful vaccines and antiviral drugs, as mutations bestowing drug resistance or immune escape often develop early after the virus is placed under selective pressure. In HIV, this is particularly relevant as the virus ranks among the fastest evolving organisms [1]. Its remarkable viral replication capability 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. The elevated variation capability of HIV gives the virus a remarkable ability to adapt to multiple selective pressures, including the immune response and antiretroviral therapy. This intra-host genetic variation raises several questions about viral evolution, for example: How much of this diversity is shaped by the selection of the immune response and how much by the antiretroviral therapy? What is the relationship between genetic diversity and clinical outcome? And finally, is it feasible to sheer the evolution of HIV in order to reduce drug resistance? Motivated by this last question, it would be desirable to develop proactive therapies that predict the advent of mutations, ergo reducing the risk of drug resistance, rather than waiting for the virus to develop resistance to reactively change the antiretroviral regimen. If we could predict the most likely evolution of the virus in any host, then it would be plausible to select an appropriate antiretroviral regimen that prevents the appearance of mutations, effectively increasing HIV control.

In this work, a Temporal Node Bayesian Networks (TNBN) model was developed to assess the occurrence of probabilistic associations among protease mutations and protease inhibitor drugs. Results of the learning of the model are presented. Our goal was to explain mutational networks in HIV evolution in the viral protease. Our probabilistic graphical model was able to predict antiretroviral drug-associated mutational pathways in the protease gene, revealing the co-occurrence of mutations and its temporal relationships. The technical challenge was to develop a model expressive enough to capture the biological complexity, yet simple enough to facilitate

the interpretation of the results. The use of TNBNs for exposing mutational networks is, as far as the authors are aware of, an additional novelty to this work.

The rest of the paper¹ is organized as follows. Section 2 highlights some important notions regarding HIV and how it develops drug resistance. Section 3 justifies the use of TNBN over other existing graphical probabilistic approaches. Section 4 describes the TNBN model. Section 5 presents the experiments and results obtained. Finally, section 6 summarizes the findings and indicates future lines of research.

2. HIV and its defense against antiretroviral therapy

HIV is the causing agent of the disease known as Acquired Immunodeficiency Syndrome (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 (UNAIDS HIV Global Report 2010). Its structure is schematically depicted in Figure 1.

The replication cycle of HIV is characterized by a reverse-transcription step of the viral RNA genome to a double-stranded DNA 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. Many of these ARV drug resistance mutations reduce HIV susceptibility to ARV drugs by themselves, while others need to accumulate in order to cause resistance. 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 compensating mutations. Due to their polymorphic nature the frequency of compensating

¹This paper is an expanded version of the one presented in the Probabilistic Problem Solving in Biomedicine Workshop (ProBioMed-11). Now we present a more detailed description of the model, new experiments and further analysis of the results.

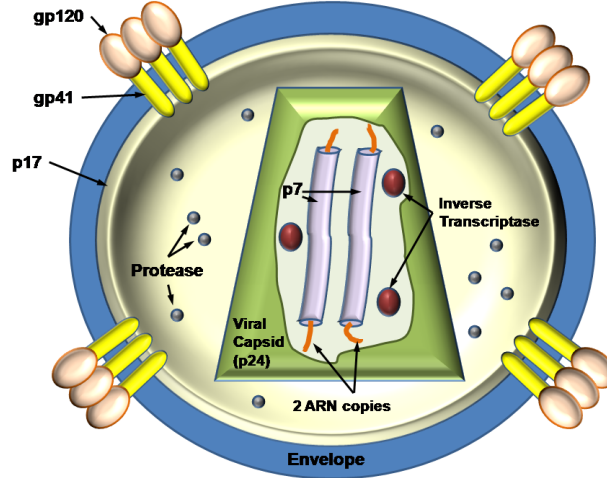


Figure 1: Schematic representation of the HIV structure. The protease and the inverse transcriptase are two main proteins of the virus. These enzymes are the target of the antiretrovirals which are studied in this paper. Some important proteins, e.g. gp120, gp41, p17 and p7 have also been illustrated.

mutations can vary between viruses circulating in different geographic areas, making it relevant to study HIV mutational networks in the context of different infected populations.

Abundant literature exists describing computational models aimed to better understand HIV evolution and immunopathogenesis. A good portion of these models is devoted to predict phenotypic HIV resistance to antiretroviral drugs using different approaches such as decision trees [3] or neural networks [4]. Other works try to identify relevant associations between clinical variables and HIV disease [5]. Surprisingly, among this wealth of literature, works aimed towards the identification of temporal relationships among mutations and drugs in HIV is almost lacking.

In [6] association rules between clinical variables and the failure of the treatment are extracted, they used 15 clinical variables from 8000 patients from data collected since 1981. The results obtained are temporal rules that have as antecedent the increasing of a subset of clinical variables and as consequent the failure of the treatment, given by side effects of the drugs or by the elevated viral count (unsuccessful therapy). None of clinical variables considered are HIV mutations.

In this paper we proposed a different approach, using Temporal Bayesian Networks in order to understand the HIV mutations process and its rela-

tionship with the antiretrovirals.

3. Bayesian Networks

Information in clinical databases is more often than not imprecise, incomplete, and with errors (noisy), and Bayesian Networks (BNs) [7] are particularly well suited to deal with uncertainty. BNs study probabilistic dependencies and relationships among domain entities. BNs models admit visual representation as a graph consisting of nodes and edges facilitating their analysis and interpretation. Nodes represent random variables and edges represent probabilistic dependencies. This graphical representation is easily understood by humans. An additional advantage is the availability of several methods to learn BN from data, e.g. [8].

BNs have proven to be successful in various domains such as medicine [9] and bioinformatics [10]. However, classical BNs are not well equipped to deal with temporal information. Dynamic Bayesian Networks (DBNs) evolved to tackle this shortcoming [11]. DBNs 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 variable changes occur infrequently, the explicit representation of DBNs becomes unnecessarily overexpressive. The alternative are Temporal Nodes Bayesian Networks [12].

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 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.

TNBN [12, 13] 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 $\mathbf{S}$. Each state is defined by an ordered pair $S = (\lambda, \tau)$, where λ is the particular value taken by v_i during its

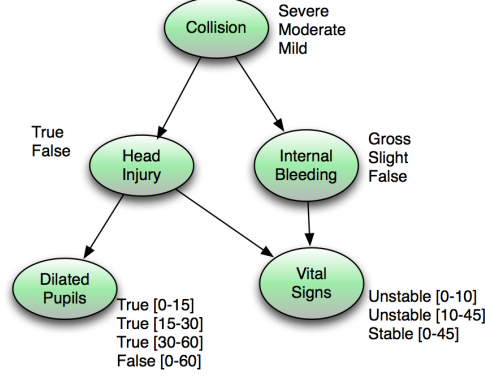


Figure 2: Each oval represents a Temporal Node. The three upper nodes (Collision, Head Injury and Internal Bleeding) are Instantaneous Nodes, so they do not have temporal intervals. The Dilated Pupils and Vital Signs are Temporal Nodes with intervals associated to them.

associated interval $\tau = [a, b]$, corresponding to the time interval in which the state change, i.e. change in value, occurs. In addition, each TN contains an extra default state $s = (\text{'no change'}, \emptyset)$ 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 $\mathbf{V}$ be a set of temporal and instantaneous nodes and $\mathbf{E}$ 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 [12], its corresponding graphical representation is shown in Figure 2.

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. To simplify the model we will 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 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 minutes 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.

There are several methods to learn BNs from data [8]. Unfortunately, the algorithms used to learn BNs cannot deal with the problem of learning temporal intervals. Then, these cannot be applied directly to learn TNBNs. The learning algorithm for TNBN used in this work has been presented in [14]. Briefly, the learning algorithm is described:

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 [15] is used to obtain an initial structure. This structure will be used in the third step, the interval learning algorithm.
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 some intervals a clustering algorithm for the temporal data is used. The approach uses a Gaussian mixture model. 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 will generate 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 Figure 5), the configurations of the parent nodes are not initially known. So, 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 the structure learning and the interval learning. However, for the experiments presented in this work, we present the results of the first iteration.

5. Experiments

5.1. Data and preprocessing

Data was obtained from the HIV Stanford Database (HIVDB) [16]. The isolates in the HIV Drug Resistance Database were obtained from longitudinal treatment profiles reporting the evolution of mutations in individual sequences.

In total data from 2373 patients with HIV subtype B was retrieved. We choose to work with this subtype because it is the most common in America [17], our geographical region of interest. For each patient data retrieved contains a history consisting of a variable number of studies. Information regarding each study consists of a treatment or cocktail of drugs administered to the patient, how long the treatment lasted in weeks, and the list of more frequent mutations in the viral population within the host at the time when the treatment was suspended (changed for a different treatment). An example of the data is presented in Table 1.

Table 1: An example of the data. It presents two patients P_1 with 3 temporal studies, and P_2 with two temporal studies.

Patient	Treatment	List of Mutations	Weeks
P_1	LPV, FPV, RTV	L63P, L10I	15
		V77I	30
		I62V	10
P_2	NFV, RTV, SQV	L10I	25
		V77I	45

For applying the learning algorithm for TNBN the data presented in Table 1 is transformed into a table similar to the one presented in Table

2. Here, each column represents a drug or mutation, each row represents a patient case, for the drugs the values are USED or NOT USED, and for the mutations the values are: APPEAR with the number of weeks that mutation appeared the first time or Default, this is when mutation did not appear in that case. The ordering provided to the K2 algorithm is: first the antiretrovirals ordered by frequency, then the mutations ordered by frequency.

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

Drug-1	Drug-2	...	Mutation-1	Mutation-2	...
USED	NOT USED		(APPEAR) 30	Default	
USED	NOT USED		(APPEAR) 40	Default	
USED	USED		Default	(APPEAR) 80	

The number of studies available varies from 1 to 10 studies per patient history. Since we are interested in temporal evolution of the mutational networks, we filtered those patients having less than 2 studies, with 973 patients outliving the conditional.

Antiretrovirals are usually classified according to the enzyme that they target. For the first experiment we focus on protease as this is the smallest of the major enzymes in terms of number of aminoacids. There exist 9 protease inhibitors (PI), namely: Amprenavir (APV), Atazanavir (ATV), Darunavir (DRV), Lopinavir (LPV), Indinavir (IDV), Nelfinavir (NFV), Ritonavir (RTV), Tripanavir (TPV) and Saquinavir (SQV). All 9 PIs will be considered during the experiments. Figure 3 presents the histogram of the administration of the different PIs in the dataset. Data from HIVDB originates from different studies and in some cases is incomplete. In this sense, Figure 3 evidences a small portion of studies only reporting the administration of a PI, but the specific antiretroviral is missing. Also there is slightly bigger portion reporting *Unknown* as the drug used. To handle this cases, if the patient case only contained *Unknown* or *None* that case was removed. However, if the case contained other drug (apart from *Unknown*), that information was used for the model.

We still have to define a target set of mutations of interest. Figure 4 presents the histogram of mutations as they appear in the dataset. 733 different mutations appeared at least once in the data. But as evident from

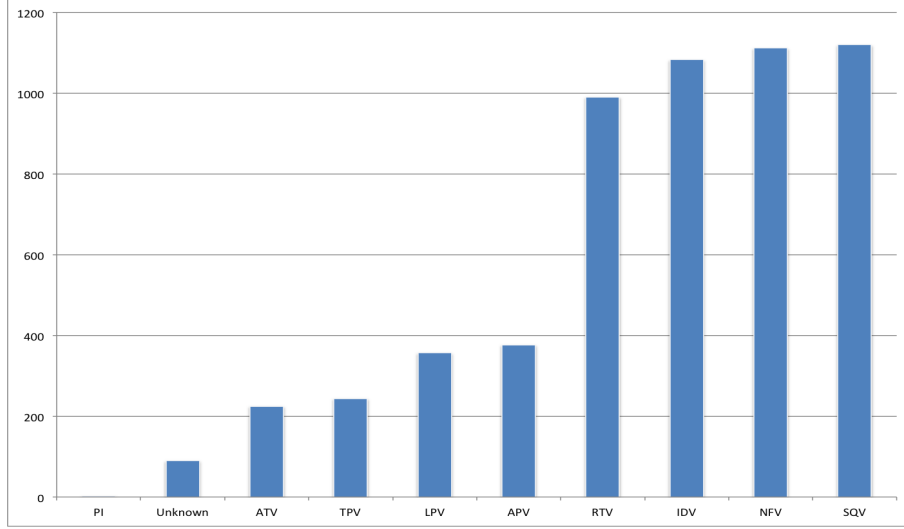


Figure 3: Histogram of the protease inhibitors administration in the full dataset including all 2373 patients.

the histogram, most mutations are rare presenting low frequency. Indeed, only few mutations exhibit high frequency. Mutation inclusion criteria is detailed below.

5.2. Model evaluation

The learning algorithm arrives to local maxima, and thus is influenced by initial parameterization. For our experiments, the number of initial intervals for each node was allowed to vary from 2 to 4 and Equal Width Discretization [18] was used to initialize those intervals. Since it does not exist a gold standard or a reference TNBN, in evaluating the model's performance three indirect measures were used: the relative Brier score (RBS), the relative time error and the total number of intervals in the model. From the 973 total patients 80% of them were used for learning and the rest for evaluation.

The Brier Score is a measure of the predictive accuracy of the network, 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:

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

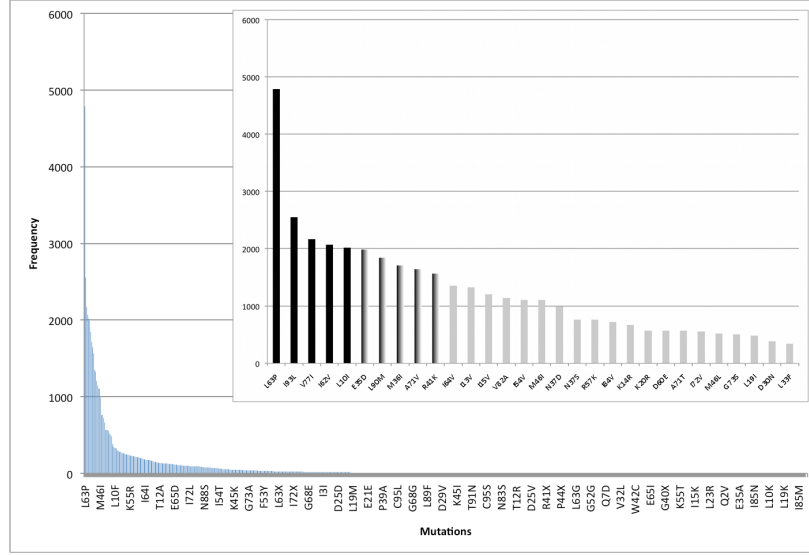


Figure 4: A group of mutations and their frequency in the full dataset including all 2373 patients. The higher frequency end of the spectrum is zoomed. Mutations used for the first experiment are shown in black, and the completing set for the second experiment are indicated in shaded black.

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) is a measure to evaluate how far the real events are from the intervals and it is defined as the difference between the real event and the middle point of the interval divided by the range of the temporal node. The range of the node is the difference between the maximum and the minimum values of the intervals in a temporal node. 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 should not be confused with the biological/physiological plausibility of the model.

5.3. Results

Two experiments have been carried out. The first experiment, with a smaller model, aims to assess the capability of TNBN for capturing known

relations and thus providing a qualitative validation of the approach. The second, with a more complete model, is aimed at uncovering the more common existing mutational networks and capturing the temporal aspect of the network formation.

In the first experiment, only mutations with more than 2000 counts were used: L63P, I93L, V77I, I62V and L10I. For tractability, in the second experiment only those mutations appearing more than 1500 times were included: L63P, I93L, V77I, I62V, L10I, E35D, L90M, M36I, A71V and R41K.

Table 3: Evaluation of the models in terms of RBS, relative time error (RTE) (in percentage) and number of intervals.

Experiment	Initial intervals	RBS	RTE	Number of total intervals
1	2	89.8	13.0	17
	3	88.3	13.6	20
	4	88.5	13.9	19
2	2	87.3	15.0	30
	3	88.5	14.7	31
	4	87.5	15.9	35

Table 3 summarizes the results for the two experiments. Figure 5 illustrates the TNBN of the first experiment exhibiting the best scores. The figure represents the network, the intervals and the prior probabilities obtained for each TN. The results show that is a small variation for different initial intervals and all the models obtained competitive predictive scores of nearly 90%.

Exploring the model reveals that RTV has arcs linking it with IDV, NFV, ATV, APV and LPV. This important relationship of RTV with other medicaments is explained due to the fact that the Ritonavir drug has been proved to boost the effect of other PIs, and therefore most of the times it is administered in combination with other drugs. The link between SQV and L10I was also already known to clinicians [19] and our model has also been successful in uncovering it. The observation of these known fact boosts our confidence that the model was meaningful; but can the model reveal new knowledge? The DRV node in the model is isolated because in the data, this drug was never given as part of a first treatment for any of the patients. This is perhaps because DRV is a relatively new drug. L63P is an extremely common mutation in the viral genotype coding the protease as revealed by the histogram in Figure 4. However, “when” this mutation appears in the

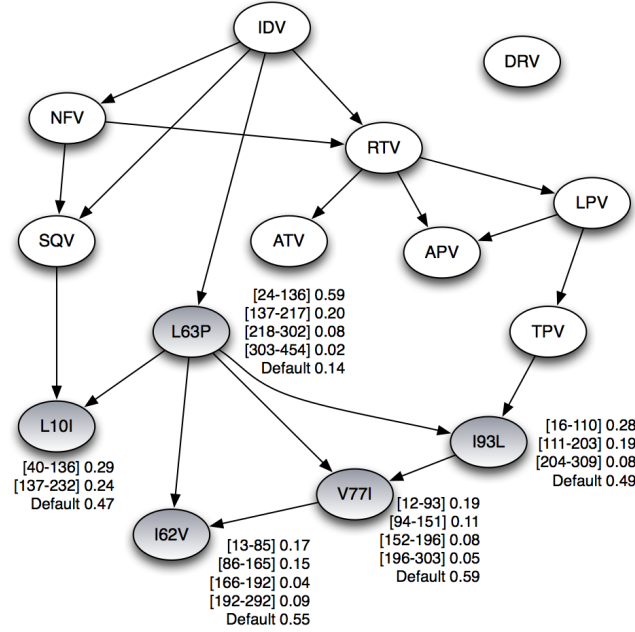


Figure 5: A learned TNBN with 9 Protease Inhibitors and 5 mutations that appear frequently. Drugs are indicated in white bubbles and mutations in gray bubbles. Learned time intervals and their probabilities for the TNs are indicated beside the bubble as is standard in the representation of TNBNs.

evolving virus remained mysterious. Our model suggests that most times this mutation tends to appear early in time, and that its probability to appear decreases over it.

Figure 6 illustrates the best TNBN model instantiation in terms of higher RBS for the second experiment. Most of the arcs from the smaller model were retained. Only the relation linking I62V and V77I and TPV with I93L are drop. Moreover, only two new arcs from SVQ and TPV to L63P appear among previously considered elements. This small variation among the two models is a good indicator of the robustness of the modeling approach. From this more complete model the prevalence of mutation L63P is evident from its relation to most drugs. It can then diverge to other mutations. There are two possible explanations for this observation; either the frequency of appearance of L63P is biasing the formation of the associations in the model -L63P almost doubles in frequency that of the following most common mutation-, or L63P is a key mutation to unleash others. Additionally, the local neighborhoods in the graph clearly reveal two mutational networks;

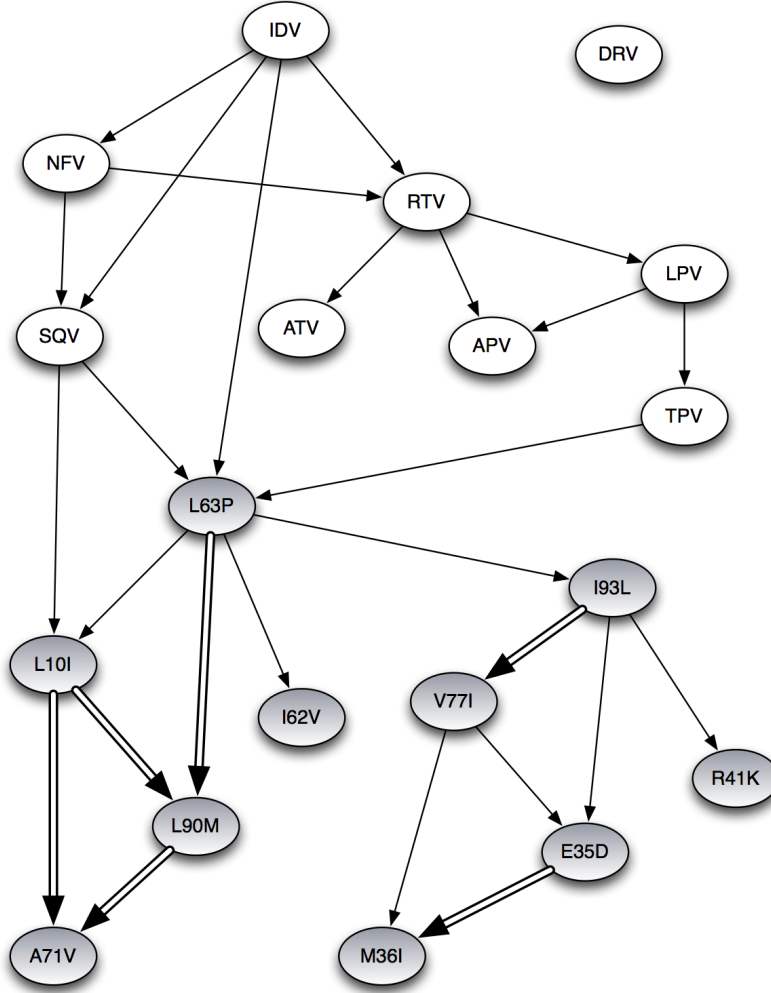


Figure 6: A learned TNBN with 9 Protease Inhibitors and 10 high frequency mutations. Drugs are indicated in white bubbles and mutations in gray bubbles. Learned time intervals for the TNs are indicated beside the bubble. Thicker arrows represent highly correlated mutations [20]. Probabilities and the Default state for the TNs are hidden for readability.

* L63P, I62V, L10I, L90M and A71V

* I93L, V77L, M36I, E35D and R41K

We are not aware of these mutational networks to have previously been reported. Moreover, the TNBN further reveals the temporal sequence of

mutation appearance.

Finally an important result is that our algorithm discovered 6 highly correlated mutations pairs presented in [20], the pairs of mutations are: (E35D, M36I), (A71V, L90M), (L10I, L90M), (L10I, A71V), (V77I, I93L), and (L63P, L90M).

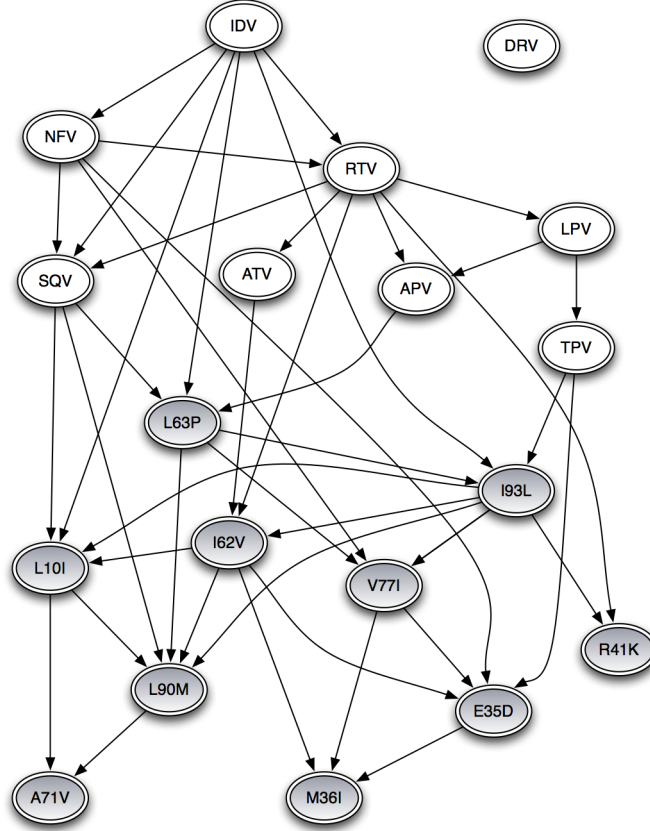


Figure 7: A learned *static* BN with 9 Protease Inhibitors and 10 mutations that appear frequently. Drugs are indicated in white bubbles, their states are Used or Not Used. Mutations are presented in gray bubbles, their states are Appeared or Not Appeared. States and probabilities are hidden for readability.

To analyze if the temporal information is giving some new information we develop a simple experiment. We remove the temporal information of the mutations. Hence, the states for both the drugs and the mutations were: Appear or Not Appeared. We used the same ordering as in the previous experiments and applied the K2 learning algorithm. The static BN learned

is presented in Figure 7. The number of arcs increased. More importantly, most of the arcs obtained with the TNBN remained. Using temporal information in this particular case yields a simpler model. Further analysis of these experiments is needed to elucidate whether this is always the case. This last exercise while interesting computationally, is far from construct validating none of the models. In this sense we are of course short from being able to determine which one is more correct, even though intuitively the simpler TNBN seems more adequate by Occam’s razor.

5.4. Finding common structures

One method for testing statistical significance of edge strengths can be evaluated using bootstrapping [10]. In non-parametric bootstrapping, the re-shuffled dataset is generated from the original dataset (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 when that edge actually appears in the set of reconstructed graphs. In the biological context, BNs should be used to understand the network of dependencies among the variables involved [10].

We performed 20-fold bootstrapping using the data that contains 10 mutations and 9 protease inhibitors. We used two thresholds for considering a relation as important. The first one is a strong relation that appears at least in 90% of the graphs, and the other is a suggestive relation, this occurs with values between 75 and 90%.

In figure 8 we present the identified relations using bootstrapping. It is important to note that our experiments obtained eight important relations between mutations, five of them: (E35D, M36I), (L10I, L90M), (L10I, A71V), (V77I, I93L), and (L63P, L90M) were previously known as highly correlated mutation pairs [20]. This is an important result that shows that our model can capture important relations using only data.

6. Conclusions

By using Temporal Nodes Bayesian Networks we have been able to unveil the two more common mutational networks present in HIV evolution as response to pharmacological selective pressure. As far as we know, we believe that these have not been previously reported. Our model has been successful in capturing relationships between mutations and protease inhibitors critically incorporating temporal information. The results obtained were compared with previously known correlated mutations and several important relations were obtained with these models. These results are encouraging,

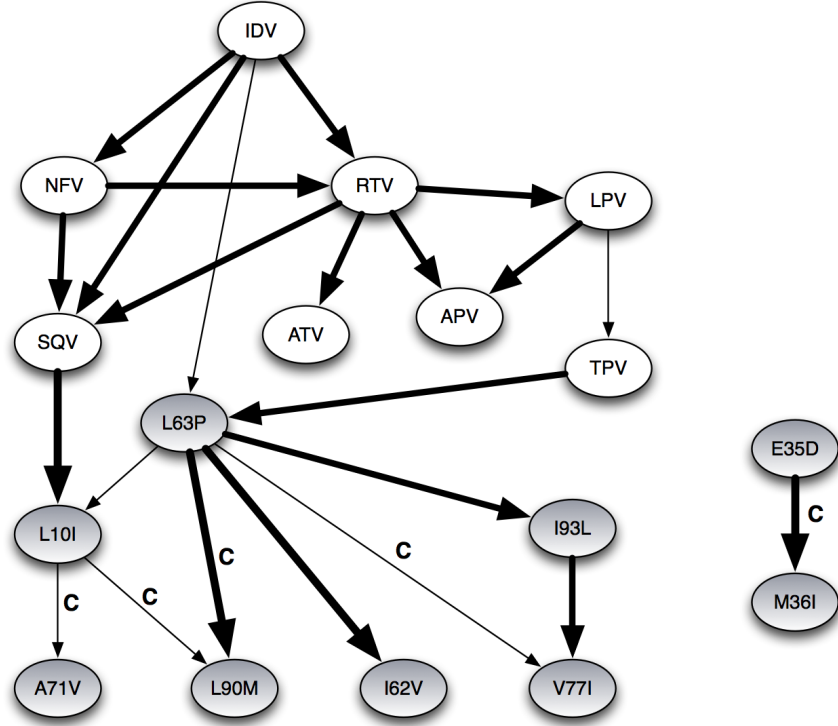


Figure 8: The patterns that appeared $\geq 90\%$ are shown in a thick arrows, those that obtained $\geq 75\%$ but $< 90\%$ are shown in thin arrows. The letter 'C' in the arrows represent highly correlated mutations [20].

presenting the model as an effective tool to explain how mutations interact with each other and providing some leverage for the clinicians in interpreting clinical tests. In this sense, the success of our model still raises more questions that demand further investigation.

The main contributions of this paper is to use a temporal probabilistic approach to understand HIV mutations. The models were developed using only data, however, they discovered some important relations of mutations.

Models such as ours are an initial step to facilitate treatment planning. If a certain mutation occurring early in a mutational network is observed during a sequentiation, one would expect the other mutations in the network to follow. Knowing the likeliness of appearance of subsequent mutations gives the therapist an idea in determining the appropriate antiretroviral regimen.

Future work plans are to learn more models according to different crite-

ria given by the experts. For the experiments presented we used the high frequency mutations and drugs. A different model could be obtained using a subset of drugs and mutations given by the experts.

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. El trabajo que ha dado lugar a este invento ha recibido apoyo del CONACyT y de la Comunidad Europea a través del FONCICYT en virtud del contrato de asignación de recursos/contrato de subvención n 95185.

References

- [1] Freeman, S., Herron, J., Payton, M.. Evolutionary analysis. Prentice Hall Upper Saddle River, NJ.; 1998. ISBN 0135680239.
- [2] Rambaut, A., Posada, D., Crandall, K.A., Holmes, E.C.. The causes and consequences of HIV evolution. *Nature Reviews Genetics* 2004;5(1):52–61.
- [3] 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):8271–8276.
- [4] Draghici, S., Potter, R.B.. Predicting HIV drug resistance with neural networks. *Bioinformatics* 2003;19(1):98–107.
- [5] Ramirez, J., Cook, D., Peterson, L., Peterson, D.. Temporal pattern discovery in course-of-disease data. *Engineering in Medicine and Biology Magazine, IEEE* 2000;19(4):63–71.
- [6] Chausa, P., Cáceres, C., Sacchi, L., León, A., García, F., Bellazzi, R., et al. Temporal Data Mining of HIV Registries: Results from a 25 Years Follow-Up. *Artificial Intelligence in Medicine* 2009;:56–60.
- [7] Pearl, J.. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann; 1988.

- [8] Neapolitan, R.. Learning Bayesian networks. Pearson Prentice Hall; 2004.
- [9] Pang, B., Zhang, D., Li, N., Wang, K.. Computerized tongue diagnosis based on Bayesian networks. Biomedical Engineering, IEEE Transactions on 2004;51(10):1803–1810.
- [10] Rodin, A., Boerwinkle, E.. Mining genetic epidemiology data with Bayesian networks: Bayesian networks and example application (plasma apoE levels). Bioinformatics 2005;21(15):3273–3278.
- [11] Dagum, P., Galper, A., Horvitz, E.. Dynamic network models for forecasting. In: Proceedings of the 8th Workshop UAI. Stanford, California, USA; 1992, p. 41–48.
- [12] Arroyo-Figueroa, G., Sucar, L.E.. A temporal Bayesian network for diagnosis and prediction. In: Proceedings of the 15th UAI Conference. Stockholm, Sweden; 1999, p. 13–22.
- [13] Galán, S., Arroyo-Figueroa, G., Díez, F., Sucar, L.. Comparison of two types of event Bayesian networks: A case study. Applied Artificial Intelligence 2007;21(3):185.
- [14] Hernandez-Leal, P., Sucar, L.E., Gonzalez, J.A.. Learning temporal nodes Bayesian networks. In: The 24th Florida Artificial Intelligence Research Society Conference (FLAIRS-24). Palm Beach, Florida, USA; 2011,.
- [15] Cooper, G., Herskovits, E.. A Bayesian method for the induction of probabilistic networks from data. Machine learning 1992;9(4):309–347.
- [16] 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.
- [17] Hemelaara, J., Gouws, E., Ghys, P.D., Osmanov, S.. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. AIDS 2006;20:W13–W23.
- [18] Liu, H., Hussain, F., Tan, C., Dash, M.. Discretization: An enabling technique. Data Mining and Knowledge Discovery 2002;6(4):393–423.
- [19] Johnson, V., Brun-Vézinet, F., Clotet, B., Günthard, H., Kuritzkes, D., Pillay, D., et al. Update of the Drug Resistance Mutations in HIV-1: December 2010. Topics in HIV medicine 2010;17:138–145.

- [20] 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.
- [21] Friedman, N., Goldszmidt, M., Wyner, A.. Data analysis with Bayesian networks: A bootstrap approach. In: Proceedings of the 15th UAI Conference. Stockholm, Sweden; 1999, p. 206–215.