Distinguishing epidemiological dependent from treatment (resistance) dependent HIV mutations: Problem Statement

Leander Schietgat[1], Kristof Theys[2], Jan Ramon[1], Hendrik Blockeel[1], and Annemie Vandamme[2]

[1]Department of Computer Science, Katholieke Universiteit Leuven
Celestijnenlaan 200A, 3001 Leuven, Belgium
{leander.schietgat,jan.ramon,hendrik.blockeel}@cs.kuleuven.be
[2]Rega Institute for Medical Research, Katholieke Universiteit Leuven
Minderbroedersstraat 10, 3000 Leuven, Belgium
{kristof.theys,annemie.vandamme}@uz.kuleuven.be

Abstract. HIV is a rapidly evolving virus leading to AIDS, a disease responsible for an estimated 2 million deaths in 2007. For optimal guidance of a patient's therapy and for the design of more effective drugs, it is important to understand how the virus becomes resistant to the current drugs and how this understanding can be used to predict therapy response. However, it is difficult to make a distinction between mutations that are inherited through epidemiological dependency and mutations that are being newly selected by a particular treatment selective pressure. Several approaches exist to separately model the two processes responsible for these mutations, but they experience problems when both processes are simultaneously at work. An integrated approach that takes into account both processes will lead to an improved understanding of the resistance development of the virus and the long-term spread of drug resistance.

1 Introduction

The Human Immunodeficiency Virus (HIV), which causes the Acquired Immunodeficiency Syndrome (AIDS), is difficult to treat due to its ability to evolve rapidly and to accumulate mutations, leading to antiviral resistance to the administered drugs. This often involves cross-resistance to other drugs in that class. To understand the process of resistance development, it is important to distinguish between two kinds of mutations. On the one hand, particular mutations are selected in a particular patient because they make the virus resistant against the administered drugs, such that patients that receive the same treatment have a more similar virus. This process is known as drug selective pressure and we refer to such mutations as *drug-related mutations*. On the other hand, mutations can be inherited from the transmission donor, such that patients that infected each other will have a more similar virus than viruses from unrelated patients. We call the corresponding mutations *polymorphisms*.

Given sequential data, there exist several methods to model the two processes responsible for these mutations independently. Firstly, data mining methods can be used to model the dependencies between mutations and the presence of drugs, which gives an understanding of drug selective pressure [1]. Secondly, methods that build phylogenetic trees, which show the ancestral relationship between different viruses, can be used to understand the process of epidemiological dependency [2]. However, as both processes happen at the same time and the difference between drug-related mutations and polymorphisms is not distinguishable from the sequence information, both kinds of methods experience problems: (1) differences in prevalence of polymorphisms between treated and untreated populations will bias the search for drug-related mutations, and (2) drug-related mutations, which have not been inherited from the transmission donor, can mislead phylogenetic analysis to cluster similarly treated patients closer than according to their epidemiological dependence. In this text, we will present some of the challenges for data mining techniques which build models from sequential intra-patient HIV sequences.

Discriminating between the two kinds of mutations will lead to more accurate models: (1) understanding how drug-related mutations give rise to resistance will lead to better treatments, and (2) a correct phylogenetic analysis will give a better insight in how the virus spreads over the world and evolves in the host or in a group of hosts. Although an integrated technique that takes both processes into account at the same time might be useful for several types of organisms, our focus is on HIV. Because of its high evolution rate, it is a key example to illustrate the setting in which intra-patient evolution due to a particular selective pressure and inter-patient evolution due to either selective pressure or genetic drift need to be teased apart.

2 Shortcomings of Existing Methods

Now that we have described the biological background of the problem, we turn to a more technical description. To briefly summarize the preceding: we have here a problem setting where the observed data are the result of two different kinds of processes, and where algorithms exist that can identify each one of these processes from data where the other process is absent, but not from data where both processes are at work.

In this section we will first cover the available data and we will then briefly discuss the methods used to model intra-patient evolution due to a particular selective pressure (in our case drug selective pressure) and inter-patient evolution. We will show what their assumptions are and how these assumptions are violated when dealing with data that contain mutations due to both types of evolution.

2.1 Available Data

Two kinds of data are relevant for this application. By genotypic data we mean the sequence of nucleotides of one or several proteins that are targeted by the

drugs. Phenotypic data contain information about the level of resistance of a virus carrying a particular sequence against one or more drugs. Testing whether a sequence is associated with phenotypic resistance against a drug is performed by monitoring the replication rate of a virus carrying that sequence in the presence of varying amounts of that drug (in vitro phenotype), or by testing the in vivo level of replicating virus (called viral load) in a patient under treatment of that drug.

There are two kinds of genotypic data. *Longitudinal datasets* consist of sequence pairs with a baseline and follow-up sequence for a single patient during a particular treatment. *Cross-sectional datasets* on the other hand use populations of unrelated sequences where each population has a specific treatment history. Cross-sectional datasets are more popular because they are generally more abundant than longitudinal datasets, since longitudinal data require keeping track of a single patient over time, with its associated privacy issues. A drawback of cross-sectional datasets is a possible effect of unknown epidemiological dependency that must be dealt with (see Section 2.2).

Throughout the section we will illustrate the problem using an example originating from Ramon et al. [3]. Consider a cross-sectional dataset of treated and untreated sequences consisting of eight nucleotides, given in Table 1. The first four nucleotides are strongly therapy-related (the wildtype is AACC while the treated sequence is mostly GGCC), while sites 5-8 evolve randomly.

Table 1. An example of a hypothetical HIV database.

ID	sequence	date	treatment	area	subtype
S1	AACCCCGA	12-01-97	untreated	Africa	C
S2	GACCCCGT	24-01-07	untreated	Africa	C
S3	GGCTGCGT	01-02-02	treated	Europe	C
S4	AACCACGT	01-02-98	untreated	Africa	C
S5	GGTTATTC	07-12-02	treated	Europe	B
S6	AACCATTT	04-12-02	untreated	Europe	B
S7	GGCTACAT	01-02-02	treated	America	B
S8	GGTTACTT	01-02-02	treated	Europe	B

2.2 Methods that Model Intra-Patient Evolution

Different evolutionary and population genetic processes are shaping viral diversity within and between hosts. However, both during inter-host and intra-host evolution selective forces (both positive and negative) and stochastic forces (genetic drift) are involved in shaping HIV genetic diversity. In this problem statement, we focus on the selective pressure[1] that is exerted upon the HIV virus by administration of antiviral drugs. To investigate this, we have information on

[1] Selective pressure is any external cause that changes reproductive success of particular variants in a population. This selective pressure is called positive, when the

whether or not a particular drug selective pressure has been active resulting in a particular sequence during a within-host evolution.

Antiviral drugs inhibit viral replication by blocking the activity of key enzymes and proteins. Treatment aims to maximally and durably suppress viral replication, leading to a decreased number of virus copies in the blood. However, ongoing replication of virus particles that can escape this suppression allows the virus population to adapt and select for one or more mutations that reduce its susceptibility to one or more drugs in the therapy. This ultimately results in therapy failure, determined by a rebound in viral load. Resistance has been described for every approved drug, and requires from only one to several mutations.

A specified selective pressure can be modeled by data mining methods such as Bayesian network learning [1], when information of genetic variation in presence and absence of this particular selective pressure is available. Mutations that are associated with resistance against a drug may be learned by comparing a population of sequences from untreated patients with a population of sequences from patients treated with that drug. A difference in prevalence of a particular mutation between these two populations may indicate a role in resistance. In addition, a structured accumulation of mutations has been observed in many cases, revealing information on drug resistance pathways. Yet, the exact order and rate are unknown for most of the drugs.

Data mining methods typically assume individuals (i.e. sequences) to be independent and identically distributed (i.i.d.), i.e. not connected somehow through their ancestry. This means that when these populations are not drawn from the same epidemiology of HIV infection, differences may also reflect epidemiological dependencies of distinct HIV populations, for example when within a particular epidemiologically closely related network, a different therapy strategy is used than in another network while both are included in the analysis. Returning to the example given in Table 1, the following rules could be learned by an association rule miner: *"If 7G then 2A"*. This rule describes a correlation between some polymorphism and a drug-related mutation. The main problem here is that a classical rule learner will not be able to distinguish therapy-related mutations from polymorphisms without knowledge about the evolution of the population. In conclusion, the epidemiological relationship between different viruses invalidates the independence between different data.

2.3 Methods that model Inter-Patient Evolution

The shape of inter-patient evolution is primarily determined by (selectively) neutral processes. Positive selection is a less potent force among patients, compared to within a patient. By inter-patient evolution we mean the accumulation of mutations due to genetic drift or unknown selective pressure (potentially even

variant increases its fitness relative to the other variants in the population, and negative, when the variant has a decreased fitness. On the other hand, accumulation of genetic variation can also be due to genetic drift, which is the process of accumulating mutations that occurs entirely due to chance events [2].

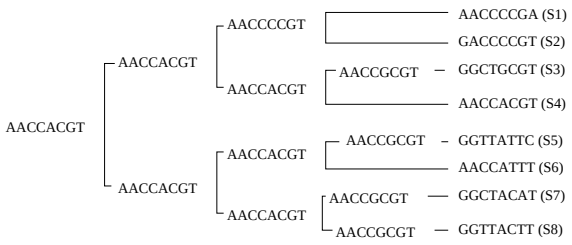

Fig. 1. A phylogenetic tree of actual evolution.

treatment selective pressure) in the viral ancestor before transmission to an infected recipient. Epidemiological relationships between patients can be modeled by a phylogenetic tree (also called an evolutionary tree). This is a tree showing the evolutionary relationships among various biological species or other entities that are believed to have a common ancestor. In a phylogenetic tree, each node with descendants represents the most recent common ancestor of the descendants, and the edge lengths in some trees correspond to genetic distance, and thus indirectly to time estimates.

A lot of research has focused on the automated construction of phylogenetic trees from data e.g., neighbor-joining (an agglomerative clustering algorithm), maximum likelihood and parsimony methods [2, 4]. The idea behind these methods is that individuals that strongly resemble each other genetically are probably closely related, i.e. they have a recent common ancestor.

Phylogenetic tree construction assumes mutations to occur independently, i.e. it is assumed that an occurrence of mutation m does not increase the probability of mutation n occurring within a certain time frame. Classical phylogenetic tree reconstruction algorithms are easily misled when this assumption is wrong. This can be illustrated in the case where drug selective pressure is present. In this case, certain specific mutations (e.g., AACC to GGTT at sites 1-4 in our example) may reoccur frequently, causing different strains of the virus to become more similar again (since they develop the same mutations). This process is known as convergent evolution. Figure 1 shows an actual tree of evolution according to our hypothetical example, while Fig. 2 shows a tree that might be learned by a standard phylogenetic tree learner. In conclusion, phylogenetic tree reconstruction algorithms tend to give converging strains a more recent common ancestor than they should.

3 Formal Problem Description

In this section, we will state the problem more formally. First, we will call an *evolutionary operator* any function P_{ev} that describes how likely it is to obtain a particular descendant sequence through evolution given a particular ancestor sequence, a particular drug treatment and an interval of time.

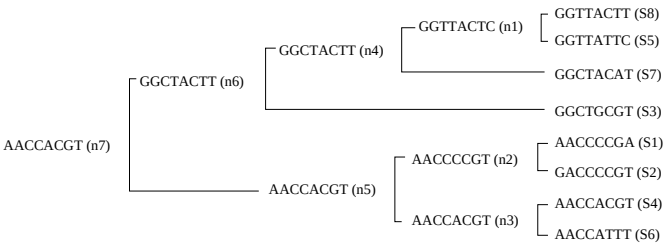

Fig. 2. A naively constructed phylogenetic tree.

The problem that we are trying to solve can now be described as follows:

Given:

- the known existence of an (unknown) evolutionary operator P_{ev}^{mut};
- the known existence of an (unknown) phylogenetic tree T_{ev}, where mutations happened both according to the evolutionary process described by P_{ev}^{mut} and according to other evolutionary operators in which we are not interested and whose existence we do not know. We do have the information along which branches P_{ev}^{mut} was present or absent;
- a dataset D consisting of the descriptions of the leaves (terminal species) of T_{ev};

Find: T_{ev} and P_{ev}^{mut}

4 Related Work

Different methods have been proposed to correct for the confounding effects of the respective processes in existing modeling methods [1]. In this section, we explain some of these methods, and we discuss advantages and disadvantages.

A first approach circumvents the problem by using longitudinal data, which includes multiple time points for a single patient. This excludes influences of different epidemiologies and therefore allows a more accurate identification of mutations associated with resistance. However, this kind of data is hard to obtain. A second approach involves the restriction of the dataset to one epidemiologically closely related cluster (such as one subtype). It reduces the confounding effect of epidemiological dependency, but it does not take into account the smaller but existing intra-subtype epidemiological dependencies. Moreover, it might be disadvantageous to leave out data and given the increasing treatment of patients infected with different subtypes, single subtype-based studies are considered too limited. A third approach involves stratification according to HIV subtype in order to achieve a similar subtype distribution in the datasets. This method also leaves out part of the data, but offers the advantage of incorporating broad HIV diversity.

The idea of using phylogenetic techniques to correct for the confounding effect of epidemiology is not new. By reconstructing the evolutionary history of sequences, one may determine whether the observed difference in prevalence of a mutation is an indication of multiple independent cases of convergent evolution, occurring at the tips of the phylogenetic tree, and thus most probably a consequence of evolution of resistance, versus an indication of inherited substitutions occurring at internal branches deeper in the phylogenetic tree. Unfortunately, as shown in Section 2.3, convergent evolution of resistance itself confounds the phylogenetic tree estimation [5], but this can initially be remedied by omitting from the analysis positions which are already known to be under drug selective pressure (hereby relying on the correctness of the selective pressure model). This introduces a chicken-and-egg problem, which stresses the need for a method that learns both the selective pressure model and a phylogenetic tree at the same time. In other work, the reconstructed evolutionary history can be used to obtain a more similar (with-in) subtype distribution across the datasets. For example, this can be done by defining a population of treatment naive sequences by sampling sequences from a larger treatment naive population which are evolutionary most closely related to the treated population [1, 6].

5 Conclusions

In this problem statement we have pointed out the need for a method that simultaneously models the two evolutionary mechanisms related to HIV. Ideally, this method can be trained from cross-sectional data.

Such a method would help to improve the untangling and discovery of resistance mutational pathways in a correct manner, which will eventually lead to better treatments. Moreover, a correct phylogenetic analysis would provide insight in how the virus spreads over the world and in a host or group of hosts. This can for example be useful when investigating the transmission history in a group of hosts.

Acknowledgements

L.S. and K.T. are supported by a PhD grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). J.R. and H.B. are post-doctoral fellows of the Fund for Scientific Research of Flanders (FWO-Vlaanderen). This work was supported by Virolab (EU IST STREP Project 027446), by the IUAP grant P6/41 and by the Katholieke Universiteit Leuven through Grant OT/04/43.

References

1. Deforche, K.: Modeling HIV resistance evolution under drug selective pressure. PhD thesis, Katholieke Universiteit Leuven (2008)

2. Salemi, M., Vandamme, A.: The Phylogenetic Handbook: A Practical Approach to DNA and Protein. Cambridge University Press (2003)
3. Ramon, J., Dubrovskaya, S., Blockeel, H.: Learning resistance mutation pathways of HIV. In: Proceedings of The Sixteenth Annual Machine Learning Conference of Belgium and the Netherlands, Amsterdam, The Netherlands. (2007) URL: http://www.cs.kuleuven.ac.be/cgi-bin-dtai/publ_info.pl?id=42687.
4. Felsenstein, J.: Inferring Phylogenies. Sinauer Associates (2003)
5. Lemey, P., Derdelinckx, I., Rambaut, A., Van Laethem, K., Dumont, S., Vermeulen, S., Van Wijngaerden, E., Vandamme, A.: Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. Journal of Virology **79** (2005) 11981–11989
6. Theys, K., Deforche, K., Libin, P., Vandamme, A.: Using ancestral state reconstruction as an alternative to correct for different epidemiologies when comparing naive versus treated sequence datasets. Statistical and Epidemiological Issues in HIV Research Workshop (2008)