Finding Relational Associations in HIV Resistance Mutation Data

Lothar Richter, Regina Augustin, and Stefan Kramer

Technische Universität München, Institut für Informatik
Boltzmannstr. 3, 85748 Garching bei München, Germany
{richter, kramer}@in.tum.de

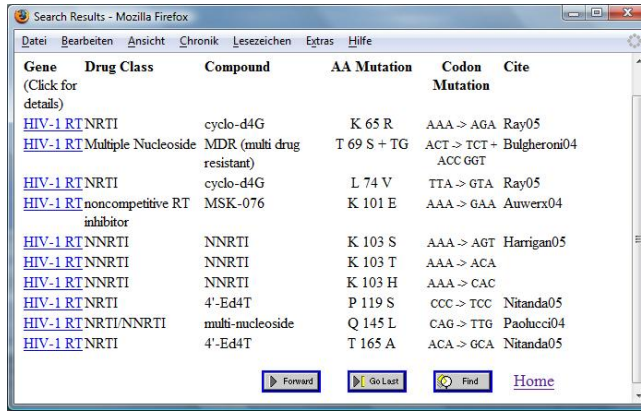
Abstract. HIV therapy optimization is a hard task due to rapidly evolving mutations leading to drug resistance. Over the past five years, several machine learning approaches have been developed for decision support, mostly to predict therapy failure from the genotypic sequence of viral proteins and additional factors. In this paper, we define a relational representation for an important part of the data, namely the sequences of a viral protein (reverse transcriptase), their mutations, and the drug resistance(s) associated with those mutations. The data were retrieved from the Los Alamos National Laboratories' (LANL) HIV databases. In contrast to existing work in this area, we do not aim directly for predictive modeling, but take one step back and apply descriptive mining methods to develop a better understanding of the correlations and associations between mutations and resistances. In our particular application, we use the Warmr algorithm to detect non-trivial patterns connecting mutations and resistances. Our findings suggest that well-known facts can be rediscovered, but also hint at the potential of discovering yet unknown associations.

1 Introduction

The optimization of HIV therapy is a crucial task, as the virus rapidly develops mutations to evade drug pressure [4]. Several machine learning approaches have been developed for decision support in this area. The task is typically predictive modeling [5, 4, 2], for instance, to predict the resistance against one or several drugs from the genotypic sequence of viral proteins and other data. While this is the ultimate goal in this application, we take one step back here and aim to find all possible correlations and associations between mutations and resistance, which has been done so far only in a limited form [5]. To do so, we define a relational representation for an important part of the data, namely the sequences of a viral protein (reverse transcriptase), their mutations, and the drug resistance(s) associated with those mutations. In the following, we present some background of this work, the raw data, the relational data representation, and some highlight results from applying standard relational association rules to the problem.

Presented at the 19th International Conference on Inductive Logic Programming (ILP'09).

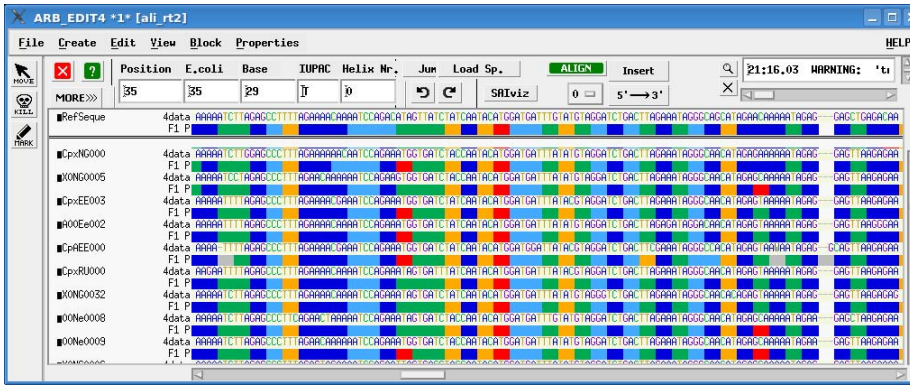
Fig. 1. Recent screenshot of LANL Resistance Database web interface showing the first ten results



2 Background

The infection with HI virus sooner or later causes the disease AIDS, which is still beyond remedy. HIV belongs to the group of retroviruses which establishes itself permanently in cells of the host's immune system. This causes a chronic infection which leads to a depletion of essential immune system cells, and the AIDS characteristics become observable. Up to now, no treatment for complete remedy or vaccination against HIV is available. Current medication strategies try to defer the point where AIDS symptoms begin to show and extend the survival time of patients. To achieve this, several drugs suppressing different steps of virus replication and infection were developed. These drugs are predominantly active against one of a few target proteins – e.g., reverse transcriptase, HIV protease or the envelope protein – which are all essential for a successful virus propagation. In this work, we present the analysis of gene sequences for reverse transcriptase genes. This is the enzyme which converts the viral RNA into DNA, which is necessary for the integration into the host cell's genome, and hence essential for replication. The integrated copy of the virus DNA acts later on as a template for the production of new virus RNA offspring. Due to a high mutation rate during replication, alterations of the gene sequences occur frequently and some of them confer resistance against a certain drug. If such a mutation occurs during drug treatment, the resistant virus strain becomes selectively propagated, the drug treatment is not effective any longer, and a new drug has to be administered. A strategy to overcome this problem is the concurrent application of several drugs, each acting against a different protein and target site. The rationale for this is that a single point mutation cannot confer resistance against drugs targeting different sites in one step. However, even this strategy does not work indefinitely, but only impedes the selection of a resistant virus strain for an extended period. We studied the occurrence of mutations conferring resistance against a certain number of drugs, and looked for rules correlating mutations and resistances against drugs.

Fig. 2. Screenshot of reverse transcriptase DNA alignment with corresponding amino acid color blocks



3 Dataset and Preprocessing

In the following, we describe the data we used and how we defined a suitable representation for mining first-order association rules with Warmr [3].

The gene sequence and mutation data we used in this study were retrieved from the Los Alamos National Laboratories (LANL) on November 28th 2006. The main information was derived from the LANL HIV resistance database (for a recent screen shot see Figure 1). Each entry of the database describes an observed resistance mutation and has eight attributes: affected gene, drug class, string identifier, wild type amino acid, position, altered amino acid, underlying nucleotide sequence alteration and literature reference. Wild type in this context means that the sequence was retrieved from the reference virus strain, which is one of the first isolates and which can therefore be regarded as original or wild. Currently, the database contains information about 370 known resistance mutations.

In addition, DNA sequences corresponding to the reverse transcriptase gene region were selected via the web interface of LANL, and subsequently re-aligned manually. This part of the data comprises 2,339 DNA sequences of a length from 1,320 to 1,443 nucleotides, which corresponds to protein sequences from 440 to 481 amino acid residues. The DNA sequences were aligned using the program package ARB [6], taking into account the genes' reading frame to avoid unnecessary frame shifts. The alignments are necessary to make sequence positions comparable and to identify mutations, i.e., deviations from the wild type at a given position. A screenshot displaying the DNA alignment and the corresponding amino acid sequence as colored blocks is shown in Figure 2. This alignment was exported to an amino acid alignment consisting of 481 positions and further processed.

Because positions for mutations are given with respect to the wild type sequence (HXBII), a head line was inserted which holds the corresponding wild type residue number if there is an amino acid in the wild type sequence, or

Table 1. Tabular representation of amino acid sequence alignment with only a few selected columns and rows

0,	WildType,	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	-	219
1,	RefSeque,	E	L	R	Q	H	L	L	R	W	G	L	T	T	P	D	-	K
796 ,	CpxNG000,	E	L	R	E	H	L	L	K	W	G	F	A	T	P	D	-	K
1992,	XONG0000,	E	L	R	E	H	L	L	K	W	G	F	T	T	P	D	-	K
1426,	CpxEE003,	E	L	R	E	H	L	L	K	W	G	F	T	T	P	D	-	K
1427,	A00Ee002,	E	L	R	E	H	L	L	K	W	G	F	T	T	P	D	-	K
1975,	CpAEE000,	Q	L	R	E	H	L	L	E	W	G	I	P	X	P	R	X,	K
2175,	CpxRU000,	E	L	R	E	H	L	L	K	W	G	F	T	T	P	D	-	K
2023,	XONG0032,	E	L	R	E	H	L	L	K	W	G	F	T	T	P	D	-	K
837,	00Ne0008,	E	L	R	E	H	L	L	K	W	G	F	T	T	P	D	-	K
838,	00Ne0009,	E	L	R	E	H	L	L	K	W	G	F	T	T	P	D	-	K

Table 2. Schemata of Warmr input relations

`key(Sequence)`	index number to address the sequences
`has_mutation(Sequence, Mutation)`	connects a sequence identifier with a mutation identifier
`resistance_against(Mutation, Drug)`	links a mutation with the resistance against a drug
`mutation_property(Mutation, AA1, Position, AA2)`	defines the properties of a mutation, where an amino acid at a certain position was substituted by another

a '–' -sign, if the alignment shows a gap for the wild type sequence at a certain position. The resulting table has the following layout (for an example, see Table 1): Line 1 contains the reference sequence indices in the alignment, followed by the values of the aligned 2,339 gene sequences. For identification, an index starting at 0 is inserted in column 1, and a string identifier in column 2. The remaining columns contain the values of the aligned amino acids of the above mentioned sequences. This table is used to detect mutations (deviations from the wild type), which are stored in the relations `has_mutation(Sequence, Mutation)` and `mutation_property(Mutation, AA1, Position, AA2)` (for details, see below).

The resulting relational representation consists of three relations: The base relation `has_mutation` connects sequences and mutations, the second relation `mutation_properties` describes the properties of mutations, and the third relation `resistance_against` connects mutations and their resistance against a certain drug. The relations are summarized in Table 2. Typical instances from these relations could be the following: Sequence number 8 carries mutation number 8 (`has_mut(8,8)`), where mutation number 8 confers resistance against a substance from the NRTI group (`res_against(8, 'Nucleoside RT Inhibitor(NRTI)')`). Mutation number 8 is further described as an amino acid change from 'I' to 'R' at position 51 (`mut_prop(8,'I',51,'R')`).

4 Results and Discussion

For our analysis, we used Warmr in the implementation of ACE 1.2.11. With a minimum support of 0.1, we obtained 4,931 frequent queries after a running

Table 3. Sample association rules found

Number	Natural language description	*Supp.*	*Conf.*	*Lift*
(1)	If position 177 is changed to E and position 286 changed to A, then a change of position 335 to D conferring resistance against NRTI is found.	0.29	0.83	1.90
(2a)	If 177 E, 292 I and 35 T then a mutation 335 D with NRTI resistance is found.	0.28	0.87	2.00
(2b)	If 177 E, 292 I and 291 D then a mutation 335 D with NRTI resistance is found.	0.29	0.87	2.00
(2c)	If 177 E, 292 I and 293 V, then a mutation 335 D with NRTI resistance is found.	0.30	0.86	1.97
(3)	If 6 D, 11 T, 35 T and 39 K, then 43 E responsible for resistance against NRTI.	0.06	0.96	8.7
(4)	If 41 L, then 215 Y.	0.06	0.82	6.8
(5a)	If 41 L and 215 Y, then 277 K.	0.05	0.79	1.5
(5b)	If 41 L and 215 Y, then 293 V.	0.04	0.75	1.1

time of two days on a standard Linux machine.[1] The output of Warmr consists of frequent queries and association rules. Because association rules are more interesting than frequent queries in our application, we will focus on the rules here. To estimate the "interestingness" of a rule, we also calculated the lift measure for each of the resulting 5,096 rules. In the rules, each mutation is given with the position relative to the wild type sequence and the altered amino acid.

Remarkably, some of the rules concern resistance mutations that were not known at the time of data retrieval. This is the case for the first rule presented here (see rule (1) in Table 3; resistance mutation 335 D was not part of the analyzed data), which is very similar to three variants of another rule (see rules (2a) to (2c)). These results show the potential of the approach, as patients that are already positive for mutations 177 E and 292 I or 286 A might better not be treated with NRTI, because the development of a resistance caused by mutation 335 D is rather likely. Another interesting rule with a newly discovered resistance is rule (3), which reflects a very tight coupling with mutation 43 E, as the frequency of the body alone is already 0.063.

Additionally to those findings which elucidated correlations between mutations and newly discovered resistances, we also found well-known correlations in the data. The mutations 41 L and 215 Y (see rule (4), both linked with resistance against NRTI) have also been described as highly correlated before [7]. In addition to this rule, there is an interesting extension (see rules (5a) and (5b)). Mutation 277 K is an alteration with respect to the wild type sequence and is not described as conferring resistance yet. Nevertheless, 277 K has shown a high correlation to known resistance mutations and may turn out to be a resistance mutation in the future, or may give strong hints for an evolving resistance based on mutations 41 L or 215 Y in the further course of a patient's treatment. For

[1] For subsequent experiments, the minimum support was lowered.

a more detailed analysis, we refer to the diploma thesis of the second author of this paper [1].

5 Conclusion

We presented a relational representation of data derived from the LANL HIV databases, and an analysis of the data using descriptive mining methods, to discover new correlations and associations between mutations and resistances against HIV drugs. Given the relevance of the application and the complex structure of the data, we believe it is a rewarding new field for ILP and relational learning methods. In particular, these methods lend themselves to the discovery of co-occurring mutations, potentially also giving hints for viral evolution paths.

The work presented in this paper could be extended in several ways: First, it would be interesting to consider a richer representation of proteins, for instance, taking into account amino acid properties. However, it has to be noted that only sequence information is known.[2] Second, the chemical structure of the inhibitor could be included. Third, it is straightforward to extend this type of analysis to other viral genes, e.g., protease. Fourth, the EuResist database (see `http://www.euresist.org/`) contains more detailed information than the LANL HIV databases, including therapy and patient data. To make the proposed approach scalable to such large-scale data, one would have to preprocess them appropriately and use suitable abstractions.

References

1. Regina Augustin. *Data Mining in HIV-Resistenzmutationen.* Diploma Thesis, Technische Universität München, 2008.
2. Steffen Bickel, Jasmina Bogojeska, Thomas Lengauer, and Tobias Scheffer. Multi-task learning for HIV therapy screening. In *Proceedings of the Twenty-Fifth International Conference on Machine Learning (ICML 2008)*, pages 56–63, 2008.
3. Luc Dehaspe and Hannu Toivonen. Discovery of frequent Datalog patterns. *Data Min. Knowl. Discov.*, 3(1):7–36, 1999.
4. Michal Rosen-Zvi *et al.* Selecting anti-HIV therapies based on a variety of genomic and clinical factors. In *Proceedings of the 16th International Conference on Intelligent Systems for Molecular Biology (ISMB 2008)*, pages 399–406, 2008.
5. Tobias Sing *et al.* Characterization of novel HIV drug resistance mutations using clustering, multidimensional scaling and SVM-based feature ranking. In *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2005)*, pages 285–296, 2005.
6. Wolfgang Ludwig *et al.* ARB: a software environment for sequence data. *Nucleic Acids Research*, 32(4):1363–1371, 2004.
7. Thomas Lengauer and Tobias Sing. Bioinformatics-assisted anti-HIV therapy. *Nature Reviews Microbiology*, 4:790–797, October 2006.

[2] Protein structure is, in general, only known for the wild type.