Mining complex genotypic features for predicting HIV-1 drug resistance

Hirotu Saigo1, Takeaki Uno2 and Koji Tsuda1,*

1Max Planck Institute for Biological Cybernetics, Spemannstraße 38, 72076 Tübingen, Germany and 2National Institute of Informatics, 2-1-2, Hittotsubashi, Chiyoda-ku, Tokyo, Japan

ABSTRACT

Motivation: Human immunodeficiency virus type 1 (HIV-1) evolves in human body, and its exposure to a drug often causes mutations that enhance the resistance against the drug. To design an effective pharmacotherapy for an individual patient, it is important to accurately predict the drug resistance based on genotype data. Notably, the resistance is not just the simple sum of the effects of all mutations. Structural biological studies suggest that the association of mutations is crucial: even if mutations A or B alone do not affect the resistance, a significant change might happen when the two mutations occur together. Linear regression methods cannot take the associations into account, while decision tree methods can reveal only limited associations. Kernel methods and neural networks implicitly use all possible associations for prediction, but cannot select salient associations explicitly.

Results: Our method, itemset boosting, performs linear regression in the complete space of power sets of mutations. It implements a forward feature selection procedure where, in each iteration, one mutation combination is found by an efficient branch-and-bound search. This method uses all possible combinations, and salient associations are explicitly shown. In experiments, our method worked particularly well for predicting the resistance of nucleotide reverse transcriptase inhibitors (NRTIs). Furthermore, it successfully recovered many mutation associations known in biological literature.

Availability: http://www.kyb.mpg.de/bs/people/hirotu/iboost/

Contact: koji.tsuda@tuebingen.mpg.de

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Due to the development of antiretroviral drugs, the progress of diseases among HIV-infected patients can be slowed down nowadays. Currently 20 antiretroviral drugs are approved: 8 protease inhibitors (PIs), 7 nucleoside and 1 nucleotide reverse transcriptase inhibitors (NRTIs), 3 non-nucleoside reverse transcriptase inhibitors (NNRTIs) and 1 fusion inhibitor. However, the long-term use of the same drug leads to a new problem: the mutations that occurred under drug pressure confers resistance to the drug. Moreover, the mutations that arose during exposure to a drug do not only confer the resistance to the specific drug, but also often to an entire drug class, i.e. cross-resistance.

Since the cost of identifying the genotypes of HIVs in a patient is relatively low, the computational prediction of drug resistance from genotype data is considered useful for designing an effective pharmacotherapy. So far, several machine-learning methods have been applied to solve the problem, e.g. linear regression (LARS) (Rabinowitz et al., 2006; Rhee et al., 2006), decision trees (Beerenwinkel et al., 2002), neural networks (NN) (Wang and Larder, 2003) and support vector regression (SVR) (Beerenwinkel et al., 2003; Sing and Beerenwinkel, 2007; Sing et al., 2005), Bayesian networks (Deforche et al., 2006). Statistical approaches are also attempted (Dirienzo et al., 2003; Foulkes and De Gruttola, 2002).

In computational biology, it is very important that the obtained prediction rule is interpretable, and understood as biologically meaningful. However, none of the existing methods can achieve high accuracy and good interpretability at the same time (see Section 2 for details).

We propose a new method called itemset boosting for predicting a target variable from a genotype (i.e. a set of mutations). It is basically a linear regression method in the space of power sets of mutations. In addition to high accuracy, our method can discover salient mutation associations, which helps biologists evaluate the biological relevance of the prediction rule.

In this article, we particularly focus on the prediction problems of HIV-1 drug susceptibilities, though our method can be applied to other problems, e.g. multiple SNP analysis for human disease association (Brinza et al., 2006). A main reason is that the underlying biological mechanism of drug resistance is relatively well understood, thus we can claim the biological implications of our prediction rules in a verifiable manner. Also, particularly in the reverse transcriptase (RT), it is known that single mutation cannot affect the susceptibility much. Only the combinations of multiple mutations can achieve the drug resistance. In such a dataset, it is necessary to use a non-linear regression method that can take into account mutation associations.

In experiments using the data from the Stanford HIV Drug Resistance Database (Rhee et al., 2003), the accuracy of our method was competitive among existing state-of-the-art methods. In addition, it will be shown that the obtained
mutation associations are consistent with previously reported ones in biological literature.

This article is organized as follows. In Section 2, we review the existing methods used to predict the HIV-1 drug resistance. In Section 3, a new method combining boosting and itemset mining is introduced. Experimental results and in-depth discussion about discovered mutation associations are shown in Section 4. We conclude with discussion in Section 5.

2 EXISTING APPROACHES

In this section, we review existing approaches and point out their drawbacks to motivate the use of our method. We start with defining some notations. Each mutation is conventionally represented using two letters and one digit like M41L, meaning that the amino acid M at position 41 in the wild type has been changed to L. Taking all different mutations in all sequences at hand, we can represent a sequence by a feature vector of binary mutation indicators,

\[ x = (x_1, \ldots, x_d), \quad x_i \in \{0, 1\}, \]

where \( d \) is the number of all mutations. Recently, Rhee et al. (Rhee et al., 2006) applied linear regression methods to predict the drug resistance. Their prediction functions are of the following form:

\[ f(x) = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d = w^T x, \]

where \([w_i]_{i=1}^d\) are regression coefficients. They reported the coefficients in the article, but some important mutations did not show up with high coefficients. For example, the weight of mutations T215Y is not high, though it is known to confer high-level resistance in conjunction with 69i (insertion between residues 69 and 70) for NRTIs (Iversen et al., 1996). It is due to the fact that linear regression does not consider associations among mutations. Later in Section 4, we will show that those mutations can be found by our approach.

A decision tree (Beerenwinkel et al., 2002) is a regression method that can take the associations into account. In the examplar decision tree shown in Figure 1, the following rules are defined:

\[ x_1 \land x_6 \Rightarrow y_1, \quad x_1 \land \neg x_6 \Rightarrow y_2, \quad \neg x_1 \land \neg x_7 \Rightarrow y_3 \]
\[ \neg x_1 \land x_7 \land x_2 \Rightarrow y_4, \quad \neg x_1 \land x_7 \land \neg x_2 \Rightarrow y_4. \]

The symbol \( \Rightarrow \) means that if the condition on the left-hand side holds, then the value on the right-hand side is obtained as the predicted resistance. These rules employ Boolean clauses with logical AND and NOT operators (\( \land, \neg \)), which are called complex genotypic features in this article. However, decision trees stand on a naive assumption that only one complex feature can determine the resistance completely. Furthermore, complex features in the decision tree have a limited form, e.g. every feature in (2) has \( x_1 \), the root node. This limitation is effective for avoiding combinatorial searches, but often harms the prediction accuracy of decision trees. To be more flexible, boosting is often applied to decision trees for better accuracy (Dietterich, 2000), but interpretability is lost in this case.

![Fig. 1. Examplar decision tree. The solid line is followed, if the corresponding Boolean variable is 1, and the broken line, otherwise.](image)

Kernel methods (Scölkopf and Smola, 2002) such as support vector machines have a prediction function as follows:

\[ f(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i). \]

where \( x \) is a training example. \( K \) is a non-linear kernel function that quantifies the similarity between \( x \) and \( x_i \) such as polynomial kernels and Gaussian kernels. Neural networks implicitly depend on numerous complex features. In fact, the prediction accuracy of kernel methods is relatively high as shown in our experiments. However, biologists are interested in discovering mutation associations in most cases. A crucial drawback is that kernel methods cannot select salient combinations, since the (implicit) coefficients are distributed over numerous complex features.

3 METHODS

Our method, itemset boosting, constructs the feature space progressively by adding a complex feature in each iteration. Basically, we follow the strategy of linear programming boosting (LPBoost) (Demiriz et al., 2002), where a next feature is selected to minimize the gap between current predictions and given target values. Unlike standard AdaBoost (Freund and Schapire, 1996), LPBoost updates the parameters by mathematical programming. Thereby the regression function is learned quickly by only a few iterations. Unless the number of mutations is trivially small, we need an efficient procedure to find a complex feature that optimizes a given score function at each iteration. For this purpose, we adopt LCM (Linear time Closed itemset Miner) (Uno et al., 2005), which is one of the best itemset mining algorithms. Typically, itemset mining is used to find interesting combinations of items out of numerous shopping records. In the considered HIV case, a sequence and mutations correspond to a shopping record and bought items, respectively.

3.1 Set representation

Our training data is represented as \((x_i, y_i)\), where \( x_i \in \{0, 1\}^d \) and \( y_i \in \mathbb{R} \) and denote the mutations of the \( i \)-th sequence and its drug resistance (i.e. the target value). The binary vector \( x_i \) is equivalently represented as a set \( X_i \), containing all the indices \( j \) such that \( x_{ij} = 1 \). If we do not use the NOT operator, i.e. \( \neg \), a complex feature can also be represented as a set \( t \), e.g. \( x_1 \land x_4 \land x_6 \rightarrow \{1, 4, 6\} \). The complex feature corresponding to the set \( t \) is written as

\[ R(t \subseteq X), \text{where } R(\cdot) \text{ becomes } 1 \text{ if the condition inside holds and } 0 \text{ otherwise.} \]
3.2 Itemset boosting

Boosting methods construct a linear combination of weak hypotheses to come up with a better prediction. In our case, a weak hypothesis corresponding to each complex feature is

\[ z_i(X) = 2(t \leq X) - 1. \]  

As a result, each hypothesis \( z_i(X) \) has either -1 or 1. One can use a simpler choice, \( z_i(X) = K(t \leq X) \), but we obtained better results with the -1/1 model.

Let \( T \) be the set of all complex features appearing in at least one of \( X_1, \ldots, X_n \). For a complex feature \( t \), define as \( L(t) \) the set of its locations (i.e., the set of sequences including \( t \)). It is often the case that two complex features \( t \) and \( t' \) have the same locations, producing redundancy in representation (1). To reduce redundancy, \( T \) is restricted to the set of closed complex features \( T \), where a closed feature is the largest one among the complex features sharing the same locations. We would like to construct a sparse regression function depending on only small number of complex features. For interpretability, we adopt the following linear function

\[ f(X) = \sum_{i \in T} a_i z_i(X) + b, \]  

where \( a, b \) are weight parameters to be learned. The learning problem is formulated as follows:

\[ \arg\min_{\mathbf{a}, b} \sum_{i \in T} |a_i| + \frac{C}{2} \sum_{i=1}^{n} \left( \sum_{t \in \mathcal{L}(i)} a_i z_i + b - y_i \right)^2, \tag{4} \]

where \( z_i := z_i(X) \). Using the L1-norm regularizer (the first term), sparsity is enforced to the parameters. This is exactly the same learning problem as that of LASSO (Tibshirani, 1996). However, we need to devise a new way of solving it due to an extremely high dimensionality of our feature space. The problem is rewritten as the following quadratic program.

\[ \min_{\mathbf{a}, b} \sum_{i \in T} (a_i^+ + a_i^-) + \frac{C}{2} \sum_{i=1}^{n} \xi_i \tag{5} \]

subject to

\[ \sum_{i \in \mathcal{L}(t)} a_i z_i + b - y_i \leq \xi_i, \quad i = 1, \ldots, n \tag{6} \]

\[ y_i - \sum_{i \in \mathcal{L}(t)} a_i z_i - b \leq \xi_i, \quad i = 1, \ldots, n \tag{7} \]

\[ a^+, a^- \geq 0, \quad \xi_i \geq 0, \tag{8} \]

where \( \xi_i \) is a slack variable, \( a_i = a_i^+ - a_i^- \). The above quadratic program has \( |T| \) variables and \( n \) constraints. Since \( |T| \) can be extremely large, directly solving this primal problem is hard. Thus, we consider the dual problem:

\[ \min_{\mathbf{u}, \mathbf{v}} \frac{1}{2n} \sum_{i=1}^{n} (u_i^+ + u_i^-)^2 - \sum_{i=1}^{n} y_i (u_i^+ - u_i^-) \tag{9} \]

subject to

\[ -1 \leq \sum_{i=1}^{n} (u_i^+ - u_i^-) z_{2i} \leq 1, \quad \forall t \in T \tag{10} \]

\[ \sum_{i=1}^{n} u_i^+ - u_i^- = 0, \quad u^+, u^- \geq 0 \tag{11} \]

where \( u_i^+ \) and \( u_i^- \) are Lagrange multipliers for the constraints (6) and (7), respectively. See the Supplementary Material for the derivation of the dual. Once the dual problem is solved, the primal solution \( a \) and \( b \) are recovered from the Lagrange multipliers of the dual problem.

Though the dual problem has too many constraints, it can be efficiently solved by an iterative procedure called the column generation algorithm (Demiriz et al., 2002). First of all, an initial solution of \( \mathbf{u} \) is obtained from the problem with no constraints (10). In each iteration, one finds the most violated constraint based on the current value of \( \mathbf{u} \), and add the found constraint to the quadratic program. In our case, a constraint corresponds to a complex feature, so we need to solve the following search problem,

\[ \arg\max_{t \in T} g(t), \quad \text{where } g(t) \text{ is the gain function} \]

\[ g(t) = \sum_{i=1}^{n} a_i u_i z_{2i}. \tag{12} \]

Then, a dual quadratic program with a limited number of constraints is solved and the obtained solution will be used in the next search.

The iteration will be continued until the dual parameter \( \mathbf{u} \) converges. Since our formulation is quadratic programming problem, global optimum is obtained. Our algorithm is summarized in Algorithm 1.

### 3.3 Searching for complex features

We use an itemset mining method (Uno et al., 2005) for searching the optimal complex feature in each iteration. Naively, one can enumerate all complex features and compute the gain (12) for each of them, and take the maximum. Enumeration of complex features, or itemsets, is a central issue of data mining and numerous methods have been proposed (Uno et al., 2005). Basically, the itemsets are organized in a search tree, and it is traversed in a depth-first manner (Fig. 2).

To be efficient, the size of the search tree has to be limited by tree pruning: suppose that one has traversed the tree up to the node with complex feature \( t \) and the maximum gain found so far is \( g_{\text{cur}} \). If there are no complex features among the supersets \( t' \) of the feature \( t \) which exceed the gain \( g_{\text{cur}} \), one can quit the traversal here and prune the downstream. To judge if the tree can be pruned at \( t \), we use the following theorem (Kudo et al., 2005; Morishita, 2001).

**Theorem 1.** Define \( y_i = \text{sgn}(u_i) \) and \( d_i = |u_i| \). For any complex feature \( t' \) such that \( t \subseteq t' \), \( g(t') < g_{\text{cur}} \) if

\[ \max |g^+(t), g^-(t)| < g_{\text{cur}}. \tag{13} \]

**Algorithm 1 Itemset Boosting**

1: procedure **ItemsetBoosting** \((X, y)\)
2: \( F_0 = \emptyset, w_i^{(0)} = y_i / \sum_{i=1}^{n} |y_i| \)
3: loop
4: \( i^* \leftarrow \text{LCM}(u, X) \) \( \triangleright \) Call Itemset Miner
5: if \( \sum_{i=1}^{n} a_i u_i z_{2i} < 1 \) then
6: break \( \triangleright \) No more hypotheses
7: end if
8: \( H \leftarrow H \cup \{i^*\} \) \( \triangleright \) Update the hypothesis set
9: \( (u, a, b) \leftarrow \text{Solve the dual problem} \) (9), (10), (11)
10: end loop
11: return \((u, a, b)\)
12: end procedure
where

\[ g^+(t) = 2 \sum_{i=1}^{n} d_i - \sum_{i=1}^{n} u_i \]

\[ g^-(t) = 2 \sum_{i=1}^{n} d_i - \sum_{i=1}^{n} u_i \]

Other conditions such as the maximum number of mutations in a complex feature and the minimum number of occurrence can be used in combination with the pruning condition (13). However, we did not use any further conditions to restrict possible associations a priori. Theoretically, the computational cost of itemset mining is NP-hard. However, the scalability depends on the properties of the dataset. In sparse data where the overlap between the sets is small, itemset mining can deal with millions of shopping records. Our HIV dataset is relatively small: hundreds of possible mutations and hundreds of sequences, so it is well within the reach of itemset mining methods. We will benchmark our method in Section 4.6.

4 RESULTS

The HIV-1 sequence data is available from Stanford HIV Drug Resistance Database (Rhee et al., 2003), where both sequences and in vitro susceptibility results are available, and mutations are defined as amino acid differences from the subtype B consensus wild-type sequence (http://hivdb.stanford.edu/pages/asi/releaseNotes/). The data used in this article is the same one as prepared and used by Rhee et al. in (Rhee et al., 2006), where drug resistance represents the fold change defined as the ratio of the IC\(_{50}\) of an isolate and the standard wild-type control isolate. They provide three datasets with different number of mutation points. In the following experiments, we chose the mutation set called ‘Complete’, because it contains the largest number of mutations and is more challenging for mutation association discovery. In our experiments, the resistance to each of 16 drugs is independently predicted based on the mutations of a sequence.

4.1 Prediction accuracy: linear versus non-linear

First, we compare the cross-validation regression accuracies of a variety of methods. Especially, we are interested in how the accuracy improves by taking complex features into account. As linear methods that do not use complex features, we adopt linear SVR, ridge regression (Ridge) and least angle regression (LARS). For the first two methods, we use our implementation and compute the accuracy in 5-fold cross-validation. For LARS, we quote the accuracies from the Supplementary Material of Rhee et al. (Rhee et al., 2006), which is also computed by 5-fold cross-validation. As non-linear methods that use complex features either implicitly or explicitly, we adopt non-linear SVR with two different kernels, polynomial and Gaussian, and our itemset boosting (iBoost).

The performance of the each prediction model is reported by \( r^2 \) value, which is defined as:

\[ r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{\sum_{i=1}^{n}(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i)^2} \]

Table 1 shows the results classified into three classes of drugs: NRTIs, NNRTIs and PIs. The number of coefficients of iBoost is different among drugs, because the number of selected features depends on the used dataset.

Interestingly, we obtained completely different results for each drug class. In NRTIs, itemset boosting performed best (0.769) and the SVR-RBF was next to best (0.747). We also performed \( t \)-test to close look at the statistical difference between iBoost and the best-performing linear method (SVR-linear), or the best-performing non-linear method (SVR-RBF). Table 2 shows the statistical superiority of iBoost at 5% significance level for 4 out of 6 drugs in the NRTI class against SVR-linear and 2 out of 6 drugs against SVR-RBF. The overall good performance of non-linear methods in NRTIs indicates that mutation associations are crucial in predicting resistance. Furthermore, the best performance of iBoost suggests that this dataset matches the method’s assumption that only a small number of complex features are salient. We will take a close look at the discovered complex features in Section 4.2.

In contrast, for NNRTIs, linear methods worked the best, and we observed only a few mutation combinations in the coefficients of iBoost (See Fig. 2 in the Supplementary Material). Therefore, for this type of drugs, mutation associations are not important factors. Our coefficients are concentrated on single mutations, avoiding the generation of unnecessary complex features. It shows that our method does not complicate the prediction rule if simple features are good enough.

For PIs, the result is somewhere in between. The non-linear methods performed better, but the difference is not as large as NRTI drugs. SVR performed better than iBoost, suggesting the difficulty of identifying salient mutation associations. As pointed out in the previous study (Rhee et al., 2006), relatively poor performance of all the methods in ATV and TDF would be due to the shortage of the data.

4.1.1 Setting parameters

In SVR, the epsilon parameter is fixed at 0.1, and the regularization parameter \( C \) was chosen from \{0.1, 1, 10, 100\}. For ridge regression, the ridge parameter was chosen from \{0.1, 1, 10, 100\} independently for each drug. For itemset boosting, regularization parameter \( C \) was chosen from \{10, 50, 150, 1000\}. In all methods, the parameter setting that showed the best cross-validation accuracy was chosen.

4.2 Results for NRTIs

The top 20 coefficients of iBoost for NRTIs are presented in Figure 3. In order to obtain stable coefficients, we ran leave-one-out perturbation and took the mean of the coefficients. The coefficients are sorted according to their absolute values and shown together with SD and the percentage of the appearance during the leave-one-out perturbation. In our method, the bias term \( b \) in (3) is adjusted automatically. As a result, we observed only a few small negative coefficients (none in top 20) in all experiments. This result implies that the drug resistance can be mostly explained with the accumulation of positive effects. Without the bias term \( b \), negative coefficients always appear as in Rhee et al. (2006). The bias term is used.
because it contributes to minimize the number of non-zero coefficients. We discovered many mutation associations, including large ones consisting of more than five mutations. Before discussing about discovered mutations, we briefly review biological mechanisms of NRTI drug resistance (Vivet-Boudou et al., 2006).

The reverse transcriptase encoded by HIV-1 virus is a major target for the antiretroviral drugs. Figure 4 shows the polymerase active site of the reverse transcriptase as well as major drug resistance mutations. Phosphorylated NRTIs bind to the active site, and compete with natural deoxynucleotide triphosphates (dNTPs). Once NRTIs are incorporated into the growing DNA chain, they act as chain terminators, resulting in blocking further extension of the DNA. HIV-1 RT has two mechanisms to neutralize NRTIs; (i) the phosphorolytic cleavage of an analogue-blocked primer, (ii) the discrimination of dNTP analogue (phosphorylated NRTIs) against the natural dNTP.

4.2.1 Thymidine-associated mutations (TAMs) The mutation associations related to TAM are marked in red in Figure 3. In our results, combination of TAMs themselves or combination of 69i and members of TAMs are observed in all drugs except for TDF.

The former mechanism is caused by the thymidine-associated mutations (TAMs), which consists of mutations M41L, D67N, K70R, L210W, T215Y/F and K219Q (Vivet-Boudou et al., 2006). They are shown as red chains in Figure 4. Single mutation is insufficient for the excision of the incorporated NRTIs, so the occurrence of multiple mutations is required for drug resistance to take place. The latter mechanism is mainly due to the Q151M complex {Q151M, A62V, V75I, F77L, F116Y} (Iversen et al., 1996), shown in blue in Figure 4.
combination clearly shown in our results. Indeed 69i associated with TAMs is reported to generate resistance to all NRTIs (Vivet-Boudou et al., 2006), thus considered as a major cause of treatment failure using NRTIs.

4.2.2 *The Q151M complex* The positions 75, 77, 116 and 151 form a functional unit called the Q151M complex (Iversen et al., 1996). In our results for AZT, all the Q151M complex members are recovered together, and smaller combinations affect the ability of the enzyme to discriminate between deoxynucleotide triphosphates (dNTP) and nucleoside analog RT inhibitors (Iversen et al., 1996). In Figure 3, the Q151M complex are marked in blue. It is observed that the Q151M complex received high coefficients in ABC, AZT, DDI and D4T. Iversen et al. (Iversen et al., 1996) showed in their biological experiments that the mutation at position 151 alone can confer high-level resistance, but the combination with 75, 77 and 116 increases the resistance significantly. This effect is also observed in our analysis; Q151M, \{F77L, Q151M\} and \{V75I, F77L, F116Y, Q151M\} are individually assigned high coefficients, which represents additive effects of accumulated mutations. However, the mutations V75I and F77L are missing in the linear regression results of Rhee et al. (Rhee et al., 2006).

4.2.3 Other remarks Vivet-Boudou et al. (Vivet-Boudou et al., 2006) summarized known mutation associations that cause cross-resistance. The complex features that are related to those known associations are highlighted in green boxes in Figure 3. They are either (i) subsets of the Q151M complex, or (ii) 69i and members of TAMs (mainly T215Y).

In Figure 3 of Beerenwinkel et al. (2002), the results of decision tree analysis are presented. Their decision tree for AZT [denoted as ZDV in (Beerenwinkel et al., 2002)] contains positions {215, 77, 75} in the same tree, though 215 and {75, 77} belong to distinctly different mechanisms. From this decision tree alone, one cannot see the underlying two mechanisms because of the inherent restrictions of mutation associations discussed in Section 2.
In HIV therapies, several drugs are often used together in order to suppress mutations that are likely to occur under a single drug (monotherapy) pressure. Especially, the combination of AZT and 3TC is well studied, because of their different mutation profiles. Also in our analysis, these drugs show clearly different profiles: the mutations at position 184 are crucial for 3TC, whereas the Q151M complex and the mutation at position 215 are important for AZT.

### 4.3 Results for PIs

Figure 1 in the Supplementary Material shows top complex features of iBoost for PIs. As expected from high prediction accuracies of linear methods, we found fewer associations than NRTIs.

Our dataset contains polymorphisms, i.e. the mutations that can occur without drug pressure. We found that the combinations including polymorphisms are prevalent in many PIs. Especially, positions 36, 63, 77 and 93 are known as the resistance-associated protease polymorphism positions (Kozal, 2004). These mutations are observed before the treatment, but help increasing the resistance to drugs. Rhee et al. (Rhee et al., 2006) mainly used the dataset without polymorphic positions to limit the dimensionality of feature space. However, our analysis suggests that polymorphic positions are strongly related to the resistance for several drugs.

Mutations at {10, 46, 54, 82, 84, 90} are known as strongly associated with cross-resistance (Kozal, 2004). Our coefficients recovered those mutations in all PIs.

### 4.4 Results for NNRTIs

The results for NNRTIs are shown in Figure 2 in the Supplementary Material. Discovered features are mostly single mutations. It would be worthwhile to note that we recovered so-called ‘primary mutations’ {K103N, Y181C/L, Y188C, G190A} (Sardana et al., 1992). NNRTIs bind to the hydrophobic pocket adjacent to the active site and induce structural modifications that decrease the nucleotide incorporation rate by displacement of the catalytic aspartate residues. Primary mutations are considered located around the pocket and prevent NNRTIs from binding to the allosteric pocket. The primary mutations are considered to confer cross-resistance, and consists of the resistance factors for NNRTIs. TAMs appearing in EFV would be an artifact caused by the combined use with NRTIs.

### 4.5 Possible false positives

Like all statistical methods, our complex features can include false positives. However, we did not find any features clearly identified as false positives, partly because the underlying biological mechanism is not completely known. For example, the mutation 103N appearing in AZT and TDF is a candidate for a false positive. The association between 103N and 210W in AZT cannot be found in literature. However, since 103N is known as a primary resistance mutation which occurs under exposure to NNRTI drugs, it might be a true association occurred under a specific combinatorial therapy using AZT and one of NNRTIs.

### 4.6 Computation time

Figure 5 shows the computation time needed for learning from the 3TC dataset (633 isolates with 371 mutations) averaged over the 5-fold cross-validations. We measured the consumed time with different maximum itemset size constraints, though we used no constraints in the reported experiments, i.e. ‘No limit’ in the plot. Since most of large itemsets are not considered due to the tree pruning, the effect of the maximum size constraint is rather limited. In total, it took around 4 min which is much slower than, e.g. support vector machines (3 s). However, this computational time is still in a practical range. Contrary to our expectation, the quadratic program took the most of computation time. Here, we used a general solver SeDuMi (Sturm, 1999), so the development of customized algorithms may improve the efficiency dramatically.

### 5 DISCUSSION

We have presented itemset boosting, a non-linear regression method that can select salient complex features. Decision trees and related methods avoid combinatorial search, thus the representation ability is limited. Our idea is not to avoid the search, but confront it using recently developed itemset mining methods. As a result, our computational cost is higher indeed, but more precise profiles can be obtained as shown in our NRTI results. Today, common statistical approaches for combinatorial therapy consist of two stages. At first, each drug resistance is estimated, then determine a
combination of drugs in a regimen (Lengauer and Sing, 2006). Our method fits in the first stage, and would be useful for building a transparent decision support system in clinical practice.

As mentioned, our method is also applicable to similar problems, such as p53 mutation analysis (Danziger et al., 2006), RNAi efficacy prediction (Vert et al., 2006), discovery of salient motif combinations (Zhu et al., 2005), multiple-SNP analysis (Brinza et al., 2006), and multiple loci analysis (Nakaya et al., 2000).

Technically, our method has a clear modular structure. It is divided into a mathematical program and an itemset mining algorithm. Therefore, a variety of machine-learning problems can be devised by just slightly modifying the mathematical program, e.g. robust regression, ordinary regression, classification and multi-task learning. Moreover, it is possible to include inequality constraints on drug resistance like knowledge-based support vector machines (Le et al., 2006).

ACKNOWLEDGEMENTS

The authors would like to thank Michael Habeck for figure preparation. Special thanks to Sebastian Nowozin for coding and proof reading.

Conflict of Interest: none declared.

REFERENCES


