Characterization of Novel HIV Drug Resistance Mutations Using Clustering, Multidimensional Scaling and SVM-Based Feature Ranking

Tobias Sing[1], Valentina Svicher[2], Niko Beerenwinkel[3], Francesca Ceccherini-Silberstein[2], Martin Däumer[4], Rolf Kaiser[4], Hauke Walter[5], Klaus Korn[5], Daniel Hoffmann[6], Mark Oette[7], Jürgen K. Rockstroh[8], Gert Fätkenheuer[4], Carlo-Federico Perno[2], and Thomas Lengauer[1]

[1] Max Planck Institute for Informatics, Saarbrücken, Germany*
[2] University of Rome "Tor Vergata", Italy
[3] University of California, Berkeley, CA, USA
[4] University of Cologne, Germany
[5] University of Erlangen-Nürnberg, Germany
[6] Center for Advanced European Studies and Research, Bonn, Germany
[7] University of Düsseldorf, Germany
[8] University of Bonn, Germany

Abstract. We present a case study on the discovery of clinically relevant domain knowledge in the field of HIV drug resistance. Novel mutations in the HIV genome associated with treatment failure were identified by mining a relational clinical database. Hierarchical cluster analysis suggests that two of these mutations form a novel mutational complex, while all others are involved in known resistance-conferring evolutionary pathways. The clustering is shown to be highly stable in a bootstrap procedure. Multidimensional scaling in mutation space indicates that certain mutations can occur within multiple pathways. Feature ranking based on support vector machines and matched genotype-phenotype pairs comprehensively reproduces current domain knowledge. Moreover, it indicates a prominent role of novel mutations in determining phenotypic resistance and in resensitization effects. These effects may be exploited deliberately to reopen lost treatment options. Together, these findings provide valuable insight into the interpretation of genotypic resistance tests.

Keywords: HIV, clustering, multidimensional scaling, support vector machines, feature ranking.

1 Introduction

1.1 Background: HIV Combination Therapy and Drug Resistance

Human immunodeficiency virus HIV-1 is the causative agent of the acquired immunodeficiciency syndrome AIDS, a disease in which persistent virus-induced

* This work was conducted in the context of the European Union Network of Excellence BioSapiens (grant no. LHSG-CT-2003-503265). T.S. would like to thank Oliver Sander for the lively and stimulating discussions on the topic of this paper.

A. Jorge et al. (Eds.): PKDD 2005, LNAI 3721, pp. 285–296, 2005.

depletion of helper T cells leads to immune failure and death due to opportunistic infections. While to date there is no cure for HIV infection, the introduction of highly active antiretroviral therapy (HAART), in which three to six antiretroviral drugs are administered in combination, has significantly improved life quality and survival time of patients. However, incomplete suppression of HIV replication by current drugs, combined with high mutation and replication rates of HIV ultimately results in the selection of viral populations carrying resistance-conferring mutations in their genomes. The fixation of these strains in the population eventually leads to therapy failure, upon which a new combination of drugs has to be chosen as next-line regimen.

1.2 Motivation: Evidence for Additional Resistance-Associated Mutations and Mutational Clusters

To date, the decision for follow-up drug combinations in patients failing therapy is routinely based on sequencing the relevant genomic region of the viral population harbored by the individual. The sequence is then analyzed to identify the presence of resistance-associated mutations for each of the 19 drugs currently available for anti-HIV therapy, by using mutation lists annually updated by the International AIDS Society (IAS) [1] or other panels of human experts.

The situation is complicated by the fact that resistance mutations do not accumulate independently from each other. Rather, they are loosely time-ordered along mutational pathways, leading to distinct mutational complexes or clusters.[1] Rational therapy planning is severely compromised by our limited understanding of these effects. Increasing evidence on additional mutations involved in the development of drug resistance [2,3], besides those listed by the IAS, provides the incentive for our present study.

1.3 Outline

We describe an approach towards the discovery and characterization of novel mutations associated with therapy failure from a large relational database, and their evolutionary and phenotypic characterization using supervised and unsupervised statistical learning methods. We focus on resistance against seven drugs from the class of nucleoside reverse transcriptase inhibitors (NRTIs), which target an HIV protein called reverse transcriptase (RT). This enzyme is responsible for translating the RNA genome of HIV back to DNA prior to its integration into the human genome. NRTIs are analogues of the natural building blocks of DNA, but lack a group essential for chain elongation. Thus, incorporation of a nucleoside analogue during DNA polymerization terminates the chain elongation process.

The knowledge discovery process described in this paper combines heterogeneous data from three different virological centers. To allow for integrated

[1] Throughout this paper, the words *complex*, *cluster*, and *pathway* are used interchangeably.

analysis, these data are stored in a relational database, whose structure is outlined in section 2. Systematic mining for mutations with differing propensities in NRTI-treated and untreated patients, respectively, as detailed in section 3, leads to the identification of 14 novel mutations associated with therapy failure. In section 4, we propose an approach towards characterizing the covariation structure of novel mutations and their association into complexes using hierarchical clustering and multidimensional scaling. Stability results are provided using a bootstrap method. Feature ranking based on support vector machines, described in section 5, allows for assessing the actual phenotypic impact of novel mutations. In section 6, we conclude by summarizing our approach, related work, and open problems.

2 The *Arevir* Database for Managing Multi-center HIV/AIDS Data

This study is based on multi-center virological data, including HIV genomic sequences from over 2500 patients, *in vitro* measurements of drug resistance [4], and clinical data such as viral load and helper T cell counts. Our relational HIV database *Arevir*, implemented in MySQL and Perl, and in use and ongoing development since 2002, provides an appropriate platform to address the challenges of data management and integration. The Arevir database schema is grouped into different modules, each consisting of a few tables, corresponding to information on patients, therapies, sequences, isolates, and specific (predicted or measured) isolate properties. Registered users can perform queries or enter data directly through a web interface. Upload of new data triggers the execution of several scripts, including programs for sequence alignment. To ensure privacy, connection between client and server is established via an SSH-tunneled Virtual Network Computing (VNC) client.[2]

3 Mining for Novel Mutations

Our approach towards identifying mutations associated with NRTI therapy is based on the assumption that these should occur with differential frequencies in treatment-naive subjects and in patients failing therapy, respectively.

Thus, mining for novel mutations was based on contrasting the frequency of the wild-type residue with that of a specific mutation in 551 isolates from drug-naive patients and 1355 isolates from patients under therapy failure, at RT positions 1–320 [5]. Chi-square tests were performed for all pairs of wild-type and mutant residues to determine mutations for which the null hypothesis that amino acid choice is independent from the patient population can be rejected. Correction for multiple testing was performed using the Benjamini-Hochberg method [6] at a false discovery rate of 0.05.

[2] Computational analyses are performed on completely anonymized data, retaining only patient identifiers instead of full names.

This procedure revealed 14 novel mutations significantly associated with NRTI treatment, in addition to those previously described in [1]: K43E/Q/N, E203D/K, H208Y, D218E[3] were virtually absent in therapy-naives ($< 0.5\%$), while K20R, V35M, T39A, K122E, and G196E were already present in the naive population with a frequency of $> 2.5\%$ but showed significant increase in treated patients. Surprisingly, mutations I50V and R83K showed significant decrease in the treated population as compared to therapy-naives.

4 Identifying Mutational Clusters

In this section we describe an unsupervised learning approach towards characterizing the covariation structure of a set of mutations and its application to the newly discovered mutations. Mutational complexes can give rise to distinct physical resistance mechanisms, but can also reflect different ways to achieve the same resistance mechanism. Indeed, the two most prominent complexes associated with NRTI resistance, the nucleoside analogue mutations (NAMs), groups 1 and 2, consisting of mutations M41L/L210W/T215Y and K70R/K219Q/D67N, respectively, both confer resistance via an identical mechanism, called primer unblocking. On the other hand, the multi-NRTI resistance complex with Q151M as the main mutation mediates a different physical mechanism in which recognition of chemically modified versions of the DNA building blocks is improved to avoid unintended integration. In essence, to appreciate the evolutionary role of novel mutations it is important to identify whether they aggregate with one of these complexes or whether they form novel clusters, possibly reflecting additional resistance mechanisms. This analysis was performed focusing on 1355 isolates from patients failing therapy.

4.1 Pairwise Covariation Patterns

Patterns of pairwise interactions among mutations associated with NRTI treatment were identified from the database using Fisher's exact test. Specifically, for each pair of mutations co-occurrence frequencies for mutated and corresponding wild-type residues were contrasted in a 2-way contingency table, from which the test statistic was computed.

A visual summary of these pairwise comparisons, part of which is shown in Fig. 1, immediately reveals the classical mutational clusters described above. It is also apparent that no significant interactions are formed between the Q151M complex and mutations from the NAM clusters, suggesting that resistance evolution along the former pathway is largely independent from the other complexes and that different pathways may act simultaneously on a sequence, at least if they mediate different physical resistance mechanisms.

In contrast, significant interactions take place across the two NAM complexes. Antagonistic interactions between the core NAM 1 mutations L210W / M41L /

[3] We use the syntax axb to denote amino acid substitutions in RT, where a is the most frequent amino acid in virus from untreated patients and b the mutated residue.

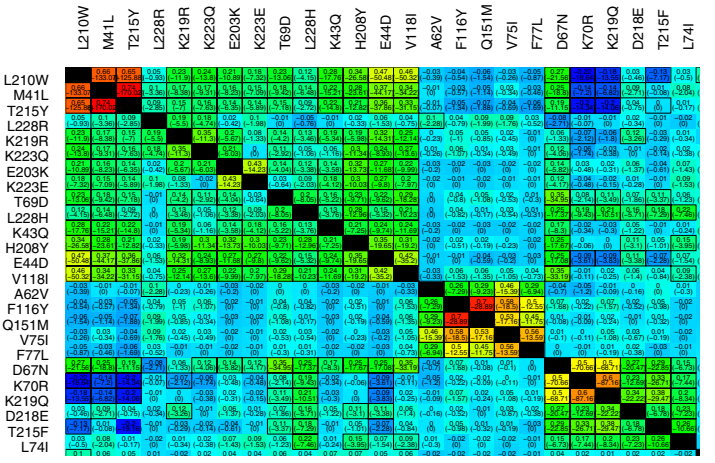

Fig. 1. Pairwise ϕ correlation coefficients between mutations (part view), with red indicating maximal observed positive covariation and blue maximal observed negative covariation. Boxes indicate pairs whose covariation behavior deviates significantly from the independence assumption, according to Fisher's exact test and correction for multiple testing using the Benjamini-Hochberg method at a false discovery rate of 0.01. The classical mutational complexes introduced in section 4 form distinct clusters, from left to right: NAM 1, Q151M multi-NRTI, NAM 2.

T215Y and NAM 2 mutations K70R and K219Q might indicate negative effects of simultaneous evolution along these two pathways, which both contribute to the primer unblocking mechanism.

4.2 Clustering Mutations

Dendrograms obtained from hierarchical clustering allow for a more detailed analysis of mutation covariation structure. The similarity between pairs of mutations was assessed using the ϕ (Matthews) correlation coefficient, as a measure of association between two binary random variables, with 1 and -1 representing maximal positive and negative association, respectively. This similarity measure was transformed into a dissimilarity δ by mapping $\phi = 1$ to $\delta = 0$ and $\phi = -1$ to $\delta = 1$, with linear interpolation in between. Since it is impossible to obtain adequate dissimilarity estimates for pairs of mutations at a single position from cross-sectional data, [4] these were treated as missing values in our approach. The resulting partial dissimilarity matrix was taken as the basis for average linkage hierarchical agglomerative clustering.[5]

The dendrogram in Fig. 2 reveals that most novel mutations group within the NAM 1 cluster (T215Y/M41L/L210W), except for D218E and F214L, which

[4] Such mutation pairs never co-occur in a sequence.

[5] In average linkage with missing values, the distance between clusters is simply the average of the *defined* distances.

aggregrate to NAM 2. Interestingly, mutations R83K and I50V, which occur more frequently in naive than in treated patients appear to form a novel outgroup.

To assess the stability of the dendrogram, 100 bootstrapped samples of RT sequences were drawn from the original 1355 sequences. Distance calculation and hierarchical clustering were performed for each of these samples as described above. Then, for each subtree of the dendrogram in Fig. 2, the fraction of bootstrap runs was counted in which the *set* of mutations defined by the subtree occurred as a subtree, without additional mutations. [6]

The four edge weights next to the root of the dendrogram show that the reported association of mutations D218E and F214L with NAM 2 is indeed highly stable across resampled data subsets, as is the grouping of other novel mutations with NAM 1, and the outgroup status of R83K and I50V. Bootstrap values for the lower dendrogram levels have been omitted for the sake of clarity; they range from 0.35 to 0.99, reflecting considerable variability of intra-cluster accumulation order. Finally, the core NAM 1 and NAM 2 mutations, respectively, are again grouped together with maximal confidence.

4.3 Multidimensional Scaling in Mutation Space

As can be seen in Fig. 1, certain mutations interact positively with mutations from both NAM pathways – an effect which might be missed in a dendrogram representation, and which can be visualized, at least to some extent, using multidimensional scaling (MDS).

The goal in MDS is, given a distance matrix D between entities, to find an embedding of these entities in $\mathbb{R}^n$ (here $n = 2$), such that the distances D' induced by the embedding match those provided in the matrix optimally, defined via minimizing a particular "stress" function. Our embedding is based on the Sammon stress function [7],

$$E(D, D') = \frac{1}{\sum_{i \neq j} D_{ij}} \sum_{i \neq j} \frac{(D_{ij} - D'_{ij})^2}{D_{ij}}, \tag{1}$$

which puts emphasis on reproducing small distances accurately. As in clustering, mutation pairs at a single position are excluded from the computation of the stress function, to avoid undue distortions.

The optimal Sammon embedding for the mutation distance matrix derived from pairwise ϕ values is shown in Fig. 3. Note that due to the non-metricity of this matrix, which violates the triangle inequality, such an embedding cannot be expected to preserve all original distances accurately, Still, the MDS plot supports the main conclusions from section 4.2, such as to the structure of the classical NAM complexes, the outgroup status of R83K and I50V, and the exclusive propensity of certain mutations, such as K43E/Q or F214L, to a unique

[6] Thus, in computing confidence values increasingly closer to the root, topology of included subtrees is deliberately ignored (otherwise, values would be monotonically decreasing from leaves to the root).

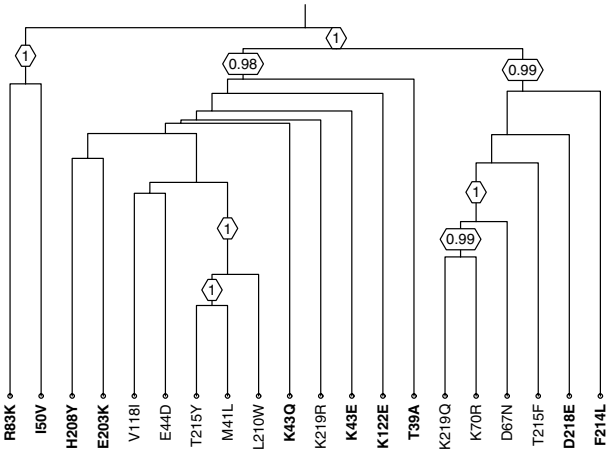

Fig. 2. Dendrogram, as obtained from average linkage hierarchical clustering, showing the clear propensity of novel mutations to cluster within one of the classical NAM complexes T215Y/M41L/L210W and K219Q/K70R/D67N, or in the case of R83K and I50V, to a distinct outgroup. Bootstrap values which are not relevant for our discussion have been removed for the sake of clarity. Distances between mutations at a single position are treated as missing values in the clustering procedure. Remarkably, such pairs of mutations can show differential clustering behavior, as is apparent in the case of K219Q/R and T215F/Y.

pathway. In addition, the plot also suggests a role in both NAM pathways for several mutations, such as H208Y, D67N, or K20R.

5 Phenotypic Characterization of Novel Mutations Using SVM-Based Feature Ranking

The analyses described above allowed us to associate novel mutations with treatment failure and to group them into distinct mutational complexes. In this section we address the question whether novel mutations contribute directly to increased resistance or merely exert compensatory functions in removing catalytic deficiencies induced by the main resistance-conferring mutations. We do so by analyzing their role in classification models for predicting phenotypic drug resistance.

Resistance of a given HIV strain against a certain drug can be measured *in vitro* by comparing the replicative capacity of the mutant strain with that of a non-resistant reference strain, at increasing drug concentrations [4]. The result of this comparison is summarized in a scalar *resistance factor*. On the basis of 650 matched genotype-phenotype pairs for each drug, we have built predictive models, using decision trees [8], and support vector machine classification and regression. These models are implemented in a publically available web server

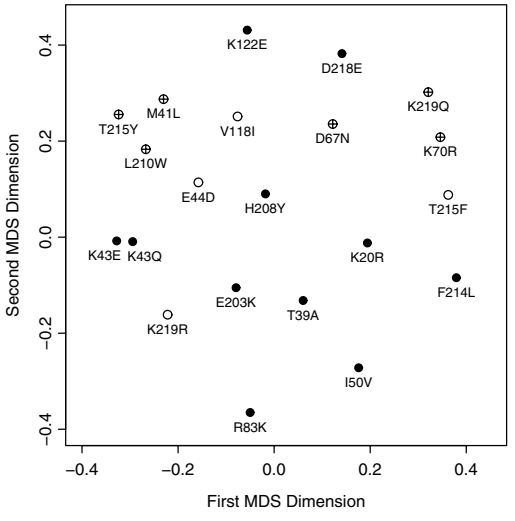

Fig. 3. Multidimensional scaling plot of novel (shown in black) and classical mutations (in white; main NAMs indicated by a cross), showing a two-dimensional embedding which optimally (according to Sammon's stress function) preserves the distances among the mutations, as derived from the ϕ correlation coefficient. Distances between mutations at a single position were treated as missing values.

called *geno2pheno* [9] (`http://www.geno2pheno.org`), which has been used over 36000 times since December 2000.

While support vector machines are widely considered as the state-of-the-art in prediction performance, there is a common attitude that these models are difficult to interpret and suffer from "the same disadvantage as neural networks, viz. that they yield black-box models" [10]. In fact, a substantial set of techniques is available for feature ranking with SVMs (e.g. [11]), by removing features or destroying their information through permutation, and even for extracting rule sets from SVMs.

In our case, using the linear kernel $k(x, y) = \langle x, y \rangle$ (standard nonlinear kernels did not significantly improve accuracy), feature ranking is particularly straightforward. Due to the bilinearity of the scalar product, the SVM decision function can be written as a linear model,

$$f(x) = \sum_i y_i \alpha_i k(x_i, x) + b = \langle \sum_i y_i \alpha_i x_i, x \rangle + b, \tag{2}$$

allowing for direct assessment of the model weights.

Figure 4 shows the result of this SVM-based feature ranking for zidovudine (ZDV), one of the seven NRTIs. All mutations associated with resistance to ZDV in the current resistance update provided by the International AIDS Society [1]

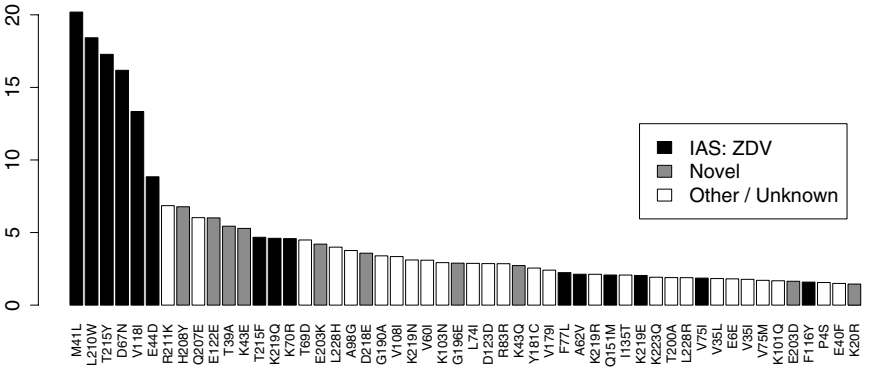

Fig. 4. Major mutations conferring resistance to zidovudine (ZDV), as obtained from SVM-based ranking of 5001 mutations. Bar heights indicate z-score-normalized feature weights (for example, mutation M41L is more than 20 standard deviations above the mean feature weight). Mutations associated with ZDV resistance by the International AIDS Society are shown in black; novel mutations identified from frequency comparisons in treated and untreated patients are shown in grey.

appear in the top 50 of 5001 features (250 positions, 20 amino acids each, plus 1 indicator for an insertion), with the first six positions exclusively occupied by classical NAM mutations (shown in black). This observation provides evidence that our models have adequately captured established domain knowledge as contributed by human experts. Remarkably, when investigating the role of novel mutations (shown in grey) in the model, we find that many of them are prominently involved in determining ZDV resistance, ranking even before several of the classical ZDV mutations.

These findings generalize to the whole NRTI drug class, as is obvious from table 1, which shows the ranks of novel mutations in the individual drug models. Table 1 also reveals some striking and unexpected differences among mutations. For example, various results suggest a close relationship of mutations H208Y and E203K, which form a tight cluster in the dendrogram, show up as neighbors in the multidimensional scaling plot, and exhibit similar rank profiles – with the notable exception of their differential impact on ddC resistance.

This surprising difference and other effects are more readily appreciated in Fig. 5, which shows the weights associated with novel mutations in the individual SVM drug models (after drug-wise z-score weight normalization for improved comparability). Indeed, increased resistance against ZDV, 3TC, and ABC upon appearance of E203K seems to coincide with *resensitization* (i.e. increased susceptibility) towards ddC. A similar, even more extreme effect can be observed in the case of T39A, for which increased resistance against ZDV and TDF again contrasts with increased ddC susceptibility. R83K shows dual behavior: increased d4T resistance and increased ZDV susceptibility. The presence of I50V is associated with increased susceptibility against all NRTIs, explaining its decreased frequency in treated patients.

Table 1. Ranks of novel mutations in SVM models for seven NRTIs, with rank 1 indicating maximal contribution to resistance, and rank 5001 maximal contribution to susceptibility. The classical mutation M184V is shown here for comparison, due to its particularly strong resensitization effect. The clinical (but not virological) relevance of results concerning ddC is limited by the limited popularity of this drug.

	ZDV	ddI	ddC	d4T	3TC	ABC	TDF
R83K	4972	3722	718	79	4973	539	154
I50V	4910	803	4702	4855	4736	4818	4899
H208Y	8	16	170	9	114	20	65
E203K	17	271	4963	103	8	19	103
K43Q	30	121	72	684	19	32	18
K43E	12	19	641	10	107	49	10
K122E	10	21	37	45	72	72	774
T39A	11	3814	4882	528	169	4017	50
D218E	20	22	103	50	25	13	659
F214L	119	898	4019	735	128	303	4844
M184V	67	2	1	4971	1	1	4994

Related effects have attracted considerable recent interest due to their possible benefits in reopening lost treatment options [12]. Arguably the most pronounced behavior can be seen in the classical mutation M184V (table 1), known to confer high-level resistance to 3TC but inducing d4T and TDF resensitization. SVM-based feature ranking reproduces this effect in a most striking manner: For ddI, ddC, 3TC, and ABC, M184V turns out to be the top resistance mutation, with contributions of 11.2,15.4,42.0, and 20.8 standard deviations above the mean. In contrast, the same mutation appears to be one of the major contributors of increased susceptibility towards d4T and TDF, 3.5 and 8.2 standard deviations *below* the mean, respectively.

6 Discussion

We have presented a case study on mining a multi-center HIV database using supervised and unsupervised methods. Previously undescribed mutations could be associated with resistance towards the drug class of nucleoside reverse transcriptase inhibitors and grouped into mutational clusters. SVM-based feature ranking on an independent data set suggests a direct contribution of novel mutations to phenotypic resistance and an involvement in resensitization effects which might be exploited in the design of antiretroviral combination therapies.

Mutation Screening. Novel mutations were found by position-wise comparisons, leaving inter-residue effects aside. It is conceivable that additional sets of mutations related to therapy failure, whose effect is too weak to discern in isolation, could be identified using other methods, such as discriminating item set miners. In fact, we have recently proposed an approach towards mining discriminating item sets, in which an overall rule weight in a mixture model of rules

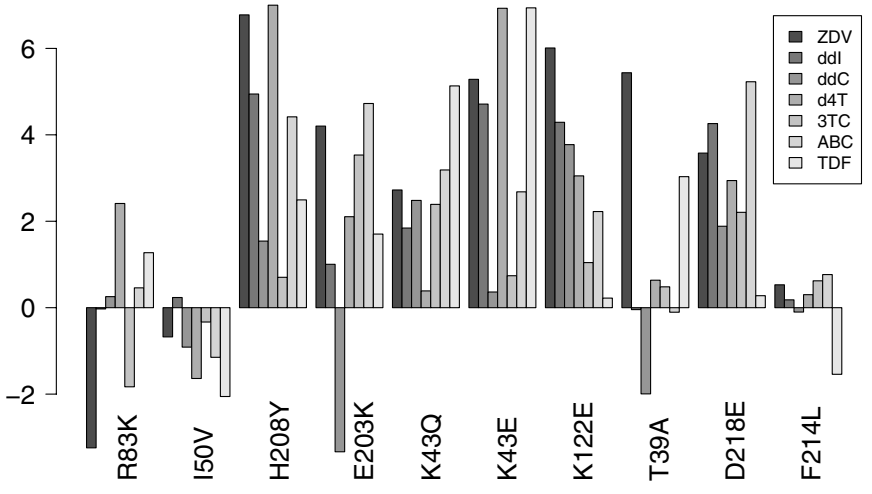

Fig. 5. Weights of novel mutations (after z-score normalization) in SVM models for seven NRTIs. For example, mutation E203K contributes significantly to ZDV resistance, while increasing susceptibility towards ddC.

is modulated by the genomic background in which a rule matches [13]. Further work will have to explore the possible benefits of using such strategies in the present context.

Covariation Versus Evolution. Dendrograms and MDS analyses describe the association of mutations into mutational complexes, but refrain from explicit statements on the accumulation order of mutations. Other approaches, most notably mutagenetic tree models [14], are explicitly tailored towards elucidating HIV evolutionary pathways from cross-sectional data as those used in our study. However, while novel mutations exhibit distinct clustering behavior, the actual order of their accumulation seems to be relatively flexible, challenging the applicability of such evolutionary models in this setting.

SVM-based Versus Correlation-Based Feature Ranking. To date, feature ranking is performed mostly using simple correlation methods, in which features are assessed in their performance to discriminate between classes *individually*, e.g. by using *mutual information*. However, as detailed in [11], feature ranking with correlation methods suffers from the implicit orthogonality assumptions that are made, in that feature weights are computed from information on a single feature in isolation, without taking into account mutual information between features. In contrast, statistical learning models such as support vector machines are inherently multivariate. Thus, their feature ranking is much less prone to be misguided by inter-feature dependencies than simple correlation methods. Further analysis of the feature rankings induced by different methods can provide valuable insights into their particular strenghts and weaknesses and suggest novel strategies for combining models from different model classes.

References

1. Johnson, V.A., Brun-Vezinet, F., Clotet, B., Conway, B., Kuritzkes, D.R., Pillay, D., Schapiro, J., Telenti, A., Richman, D.: Update of the Drug Resistance Mutations in HIV-1: 2005. Top HIV Med **13** (2005) 51–7
2. Gonzales, M.J., Wu, T.D., Taylor, J., Belitskaya, I., Kantor, R., Israelski, D., Chou, S., Zolopa, A.R., Fessel, W.J., Shafer, R.W.: Extended spectrum of HIV-1 reverse transcriptase mutations in patients receiving multiple nucleoside analog inhibitors. AIDS **17** (2003) 791–9
3. Svicher, V., Ceccherini-Silberstein, F., Erba, F., Santoro, M., Gori, C., Bellocchi, M., Giannella, S., Trotta, M., d'Arminio Monforte, A., Antinori, A., Perno, C.: Novel human immunodeficiency virus type 1 protease mutations potentially involved in resistance to protease inhibitors. Antimicrob. Agents Chemother. **49** (2005) 2015–25
4. Walter, H., Schmidt, B., Korn, K., Vandamme, A.M., Harrer, T., Uberla, K.: Rapid, phenotypic HIV-1 drug sensitivity assay for protease and reverse transcriptase inhibitors. J Clin Virol **13** (1999) 71–80
5. Svicher, V., Ceccherini-Silberstein, F., Sing, T., Santoro, M., Beerenwinkel, N., Rodriguez, F., Forbici, F., d'Arminio Monforte, A., Antinori, A., Perno, C.: Additional mutations in HIV-1 reverse transcriptase are involved in the highly ordered regulation of NRTI resistance. In: Proc. 3rd Europ. HIV Drug Resistance Workshop. (2005) Abstract 63
6. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society (Series B) **57** (1995) 289–300
7. Sammon, J.: A non-linear mapping for data structure analysis. IEEE Trans. Comput. **C-18** (1969) 401–409
8. Beerenwinkel, N., Schmidt, B., Walter, H., Kaiser, R., Lengauer, T., Hoffmann, D., Korn, K., Selbig, J.: Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. Proc Natl Acad Sci U S A **99** (2002) 8271–6
9. Beerenwinkel, N., Däumer, M., Oette, M., Korn, K., Hoffmann, D., Kaiser, R., Lengauer, T., Selbig, J., Walter, H.: Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. Nucleic Acids Res **31** (2003) 3850–5
10. Lucas, P.: Bayesian analysis, pattern analysis, and data mining in health care. Curr Opin Crit Care **10** (2004) 399–403
11. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning **46** (2002) 389–422
12. Wang, K., Samudrala, R., Mittler, J.E.: HIV-1 genotypic drug-resistance interpretation algorithms need to include hypersusceptibility-associated mutations. J Infect Dis **190** (2004) 2055–6
13. Sing, T., Beerenwinkel, N., Lengauer, T.: Learning mixtures of localized rules by maximizing the area under the ROC curve. In José Hernández-Orallo, *et al.*, ed.: 1st International Workshop on ROC Analysis in Artificial Intelligence, Valencia, Spain (2004) 89–96
14. Beerenwinkel, N., Däumer, M., Sing, T., Rahnenführer, J., Lengauer, T., Selbig, J., Hoffmann, D., Kaiser, R.: Estimating HIV Evolutionary Pathways and the Genetic Barrier to Drug Resistance. J Infect Dis **191** (2005) 1953–60