Geno2pheno: Interpreting Genotypic HIV Drug Resistance Tests

Niko Beerenwinkel, Thomas Lengauer, and Joachim Selbig, *Fraunhofer Institute for Algorithms and Scientific Computing*
Barbara Schmidt, Hauke Walter, and Klaus Korn, *German National Reference Center for Retroviruses*
Rolf Kaiser, *University of Cologne*
Daniel Hoffmann, *Center of Advanced European Studies and Research*

This intelligent system uses information encoded in the HIV genomic sequence to predict the virus's resistance or susceptibility to drugs. To make predictions, geno2pheno employs decision tree classifiers and support vector machines.

HIV-infected patients are at a high risk of developing AIDS, now the fourth-leading cause of death worldwide. In the developed world, 15 antiretroviral drugs that interfere with the virus's replication cycle are approved for use in clinical practice. These drugs belong to three distinct drug classes targeting two viral enzymes: nucleoside and nonnucleoside reverse transcriptase (RT) inhibitors (NRTI and NNRTI, respectively) and protease inhibitors (PI). (See the "Abbreviations" sidebar for other abbreviations in this article). Even with the currently recommended combination therapies consisting of three to five drugs, few patients achieve long-term suppression of plasma virus levels below detectable limits.

Drug resistance[1] is a major factor contributing to therapy failure. The genetic basis of drug resistance is HIV's high mutation rate and very high replication rate. Researchers have estimated that each single mutation in the 9-kbp (kilobase pair) viral genome appears once daily in each infected individual. Some mutations lead to a slightly altered 3D protein structure that enables the viral enzyme to fulfill its task even in an inhibitor's presence (see Figure 1). These mutants have a selective advantage under drug pressure and become dominant in the virus population. So, persistent viral replication due to subinhibitory drug levels or host immune failure leads to the evolution of drug-resistant variants and consequently to therapy failure.

To support the diagnosis of HIV resistance or susceptibility to antiretroviral agents, we have developed *geno2pheno*. This intelligent system uses two machine learning techniques—decision trees and linear support vector machines—to predict phenotypic resistance from information encoded in the viral genomic sequence. We applied both techniques to more than 400 genotype–phenotype pairs for 13 drugs. Our results show that geno2pheno performed well for all but three of the drugs.

Resistance testing

To find a new, potent drug combination after therapy failure, current treatment guidelines recommend resistance testing. Existing resistance-testing methods are based on either directly measuring viral activity in the presence and absence of a drug (*phenotyping*) or scanning the viral genome for resistance-associated mutations (*genotyping*). Clinical studies have demonstrated that both phenotypic and genotypic resistance testing can significantly improve treatment response.

Phenotyping

Phenotyping is considered the "gold standard" for resistance testing. Most assays use recombinant-

1094-7167/01/$10.00 © 2001 IEEE

Abbreviations

ABC	Abacavir
CTSHIV	Customized Treatment Strategies for HIV
ddC	Zalcitabine
ddI	Didanosine
DLV	Delavirdine
DNA	Deoxyribonucleic acid
d4T	Stavudine
EFV	Efavirenz
F	Phenylalanine
G	Glycine
HIV	Human Immunodeficiency Virus
I	Isoleucine
IDV	Indinavir
L	Leucine
M	Methionine
NFV	Nelfinavir
NNRTI	Nonnucleoside reverse transcriptase inhibitor
NRTI	Nucleoside reverse transcriptase inhibitor
NVP	Nevirapine
PI	Protease inhibitor
R	Arginine
RF	Resistance factor
RT	Reverse transcriptase
RTV	Ritonavir
SQV	Saquinavir
T	Threonine
V	Valine
ZDV	Zidovudine
3TC	Lamivudine

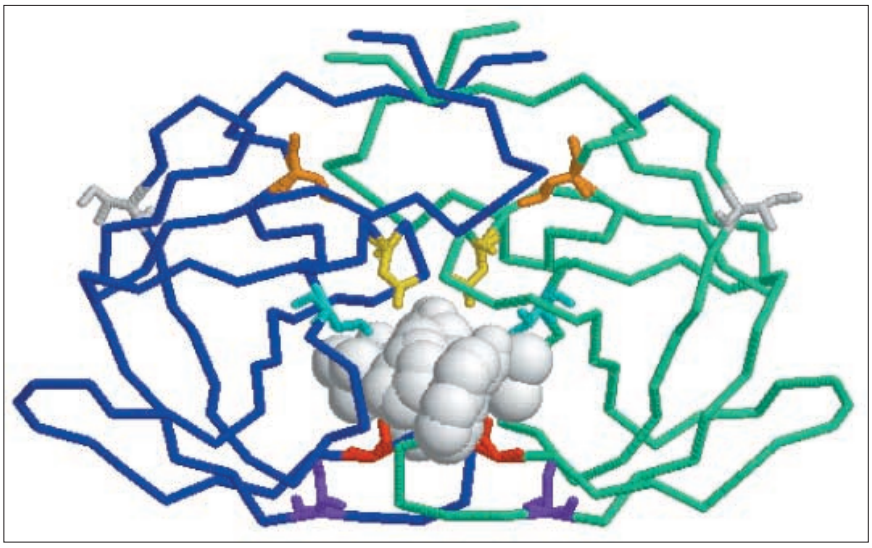

Figure 1. Dimeric protease in complex with two molecules of the drug saquinavir (the light-gray spheres). The backbones of the two polypeptide chains are blue and green, respectively. The yellow sticks represent active site residue, aspartic acid, at position 25. The positions of the residues occurring in the saquinavir classification model (see Figure 3) appear here as colored sticks: 48 glycine (G) in red, 54 isoleucine (I) in purple, 72 isoleucine in light gray, 84 isoleucine in cyan, and 90 leucine (L) in orange.

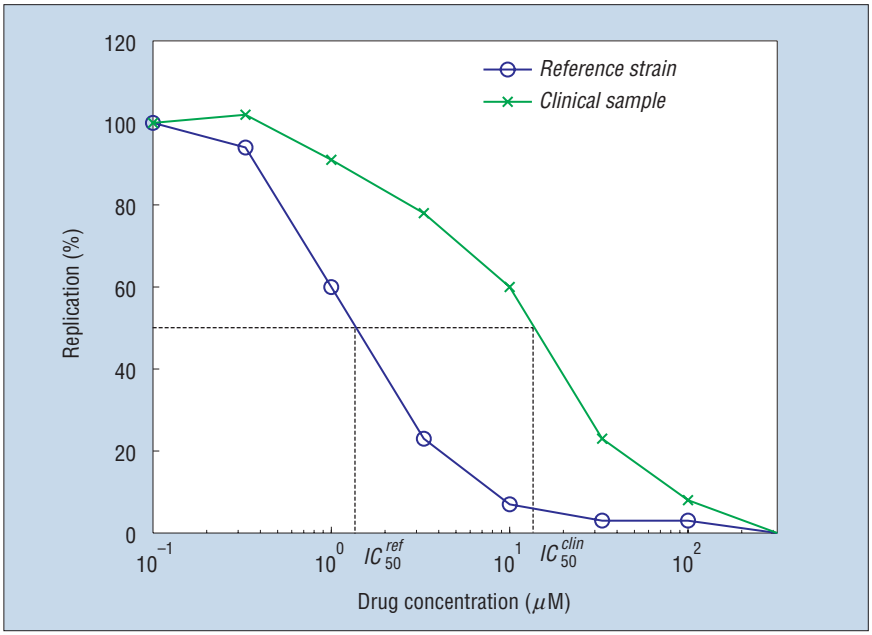

Figure 2. Drug response curves for a reference virus (blue) and a viral population extracted from a clinical sample (green). The dashed lines mark the two IC_{50} values (16.289 micromoles [μM] for the clinical sample and 1.625 μM for the reference sample), which denote the drug concentration needed to inhibit viral replication by 50 percent. In this example, the resistance factor is $RF = \frac{16.289}{1.625} = 10.03$.

virus techniques directly measuring viral replication in the presence of increasing drug concentration.[2] This results in two dose-response curves, one for the clinical sample under investigation and one for a susceptible reference virus. Resistance is usually expressed in terms of the *resistance factor*:

$$RF = \frac{IC_{50}^{clin}}{IC_{50}^{ref}},$$

where IC_{50} denotes the drug concentration needed to inhibit viral replication by 50 percent (see Figure 2). Phenotypic assays yield an easily interpreted quantitative measure of the degree of resistance to each drug. However, they are time consuming (four to eight weeks) and expensive ($750 to $1,000 for the full range of approved drugs).

Genotyping

In contrast, genotypic assays can provide results within a few days, are less expensive (between $250 and $500), and are available as standardized commercial kits for routine diagnostics. Genotypic resistance testing includes sequencing the relevant parts of the gene coding for the viral drug targets (protease and RT) and recording the differences in amino acids compared to a reference strain. Such differences might represent naturally occurring variants or might be associated with drug resistance. Genotyping also detects

Related Work

Many mutations have already been associated with drug resistance. Such findings can result from observing the emergence of a certain mutation either in samples derived from patients under a well-defined therapy (usually a monotherapy–treatment with just one drug) or in cell culture under continuous drug pressure. Another way of associating mutations with drug resistance consists of inserting mutations into a susceptible viral clone (site-directed mutagenesis) and then comparing the phenotypes of the wild-type (the standard form) and mutant.

Lookup tables

This type of knowledge is available from the scientific literature and has been compiled in lookup tables[1] and databases. The Los Alamos resistance database (http://hiv-web.lanl.gov), for example, contains 159 entries for the protease and 270 for the reverse transcriptase, each associating an amino acid substitution with resistance to a certain drug.

However, it is unclear how helpful these associations are in real-life applications. Clinical samples contain complex mutational patterns, rendering interpretation difficult. Drug resistance can sometimes be due to a single amino acid substitution; in other cases, the accumulation of a number of mutations seems necessary. Moreover, the effect of some mutations depends on the presence or absence of other mutations. Indeed, some mutations reverse resistance caused by others. So, a mutation cannot be considered independently of its background sequence. Furthermore, some mutations or combinations of mutations might also cause resistance to drugs the patient has not yet been exposed to, leading to considerable cross-resistance, especially among nonnucleoside reverse transcriptase inhibitors and protease inhibitors.

Rule-based systems

Several research groups have set up scoring systems for relating sequence variations to phenotypic drug resistance or likelihood of therapy failure. Two freely available systems rely on the systematic incorporation of published data.

Richard Lathrop and his colleagues have proposed CTSHIV (Customized Treatment Strategies for HIV), a rule-based system designed to overcome one of the difficulties with lookup tables—namely, how to use this knowledge.[2] Basically, CTSHIV encodes associations between amino acid changes and drug resistance into rules and applies these rules to the genotype under consideration and nearby mutants. It uses a branch-and-bound algorithm to identify drug combinations that could avoid additional resistance mutations.

Robert Shafer, Duane Jung, and Bradley Betts use *mutation scoring tables* to calculate from each sequence a score that is translated into one of five classes ranging from susceptible to high-level resistant.[3] Experts derive the scoring tables from published data on correlations between genotype and phenotype, treatment history, and clinical outcome.

Both methods provide a rational approach to incorporating the knowledge from the scientific literature. However, they depend largely on the published data's quality and applicability and the chosen rules or scores.

References

1. R.F. Schinazi, B. Larder, and J.W. Mellors, "Mutations in Retroviral Genes Associated with Drug Resistance: 2000–2001 Update," *Int'l Antiviral News*, vol. 8, no. 5, May 2000, pp. 65–92.
2. R.H. Lathrop et al., "Knowledge-Based Avoidance of Drug-Resistant HIV Mutants," *Proc. 15th Nat'l Conf. Artificial Intelligence/10th Conf. Innovative Applications of Artificial Intelligence*, AAAI Press, Menlo Park, Calif., 1998, pp. 1071–1078.
3. R.W. Shafer, D.R. Jung, and B.J. Betts, "Human Immunodeficiency Virus Type 1 Reverse Transcriptase and Protease Mutation Search Engine for Queries," *Nature Medicine*, vol. 6, no. 11, Nov. 2000, pp. 1290–1292.

nucleotide mixtures (*wobbles*), which reflect HIV's quasi-species nature. Genotypic assays usually identify variants representing at least 30 percent of the virus population. The challenge with using genotypic assays is the interpretation of sequence information.

Geno2pheno

Our intelligent system provides genotype interpretation in terms of the in vitro phenotype. That is, it predicts phenotypic drug resistance from sequence information by analyzing a large set of the observed genotype–phenotype pairs. Unlike existing rule-based systems, geno2pheno does not rely on data from lookup tables or other published data, which might be of limited use (see the "Related Work" sidebar).

We model drug resistance with the classes susceptible or resistant. This formulation as a binary classification problem requires mapping RFs obtained from phenotypic testing onto susceptible and resistant by choosing cutoff values for those factors. Defining appropriate cutoffs is a difficult task that depends on the assay, on the drug, and on what the resulting classes are supposed to mean (resistant could, for example, mean "predictive of therapy failure" or "significantly above expectation among drug-naïves [patients who haven't ever taken a certain drug]"). Interpreting a phenotypic test involves essentially the same task.

Here we regard the cutoff value c as a parameter. So, for each drug d we must learn the function

$$F_{d,c} : S \rightarrow \{\text{susceptible}, \text{resistant}\}$$

defined on the set S of all sequences coding for an HIV protease or RT. The genotype–phenotype pairs $(s, F_{d,c}(s))$, $s \in S$, that make up the training data can be determined only by experimental methods, which are always error-prone. Sequence data are well reproducible, but for RFs, researchers have observed coefficients of variation between 10 and 60 percent (depending on both the drug and the resistance level).[2] We applied two machine learning techniques—decision tree classification and linear support vector machine classification—that follow different strategies to handle this type of high-dimensional, noisy data. (Dechau Wang, Stuart Bloor, and Brendan Larder follow a similar approach; they use an artificial neural network to predict phenotypic drug resistance from genotypes.[3])

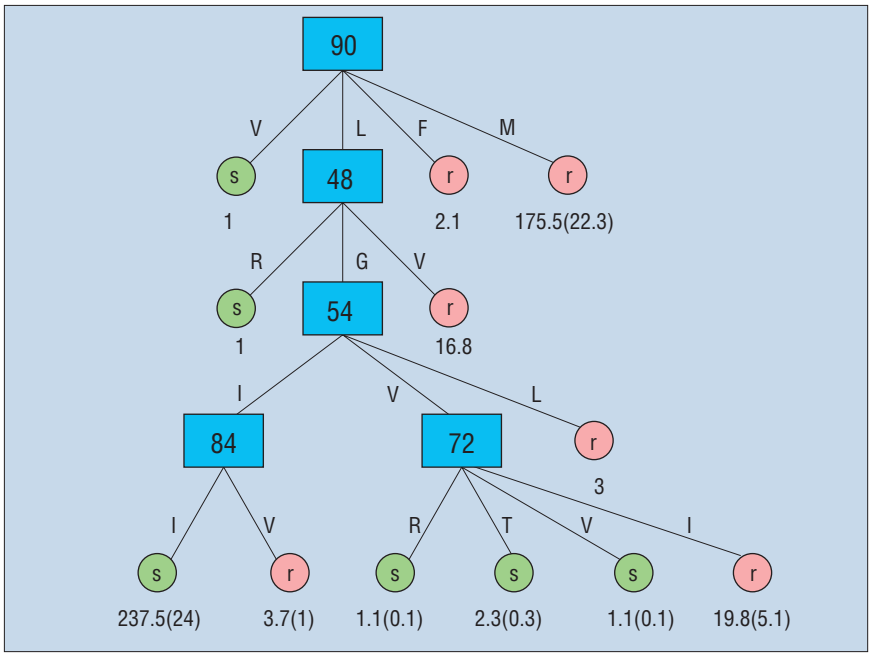

Figure 3. A decision tree classifier for resistance to the drug saquinavir. Interior vertices represent attributes; the numbers in the blue rectangles indicate the protease sequence positions. Leaves represent classes, either resistant (red circles), defined by a resistance factor *RF* ≥ 3.5, or susceptible (green circles). The letter next to an edge indicates its attribute value—that is, the amino acid (for an explanation of the letters, see the "Abbreviations" sidebar). The numbers *N* or *N*(*E*) under each leaf denote the number *N* of samples classified by this path and the estimated number of errors *E* due to this classification (for more details, see Ross Quinlan's book on C4.5[4]).

Testing the system

To test both machine learning techniques, we investigated 471 clinical samples derived from patients after therapy failure. Genotypic data were available for the complete protease and the first 220 to 250 amino acids of the RT. We obtained phenotypic results for 13 antiretroviral agents:

- NRTIs—zidovudine (ZDV), zalcitabine (ddC), didanosine (ddI), stavudine (d4T), lamivudine (3TC), and abacavir (ABC);
- NNRTIs—nevirapine (NVP), delavirdine (DLV), and efavirenz (EFV); and
- PIs—saquinavir (SQV), indinavir (IDV), ritonavir (RTV), and nelfinavir (NFV).

This resulted in 443 to 469 genotype–phenotype pairs.

We processed sequence data by aligning each DNA sequence against a reference strain and translating it into amino acids. We found one sample with a deletion of a single amino acid and eight samples with a member of a known family of insertions between RT positions 67 and 70. (All position numbers refer to residue positions relative to HXB2; for more details, see http://hiv-web.lanl.gov/HTML/reviews/HXB2.html.) The researchers conducting the experiments chose cutoffs on the basis of published data on test variability and treatment response.

Decision trees. We represent sequences by one attribute X^j for each amino acid sequence position ($j = 1, \ldots, 99$ for all 99 positions of the protease, and $j = 1, \ldots, 250$ for the first 250 positions of the RT). Each attribute's value is one of the 20 naturally occurring amino acids. We assume that unknown or ambiguous positions due to wobbles in the DNA sequence are distributed probabilistically according to the known attribute values. For the RT, an additional binary attribute X^0 indicates an insertion of the above type. We denote by $Y \in$ {resistant, susceptible} the phenotypic class we want to predict from the attributes X^j.

We use C4.5[4] to generate decision trees by recursively splitting the training set following a heuristic divide-and-conquer strategy. The test criterion for a split is the attribute for which the *gain ratio*,

$$\frac{I(X^j, Y)}{H(X^j)},$$

is maximal. The gain ratio is the quotient of information $I(X^j, Y)$ that attribute X^j provides about the class Y and the entropy $H(X^j)$ of X^j. To avoid overfitting, our method prunes trees by removing subtrees that it estimates will increase the error rate minimally (*reduced-error pruning*). This estimate also allows the calculation of confidence factors for each prediction based on the tree. Geometrically, a decision tree corresponds to a partitioning of the input space—that is, the space spanned by the discrete attributes X^j—into rectangular (in some non-Euclidean sense) regions, each labeled with one of the two classes.

For our data set, we obtained 13 decision trees with four to seven interior vertices (Figure 3 shows one example). Thus, we identified genotypic patterns characteristic of drug resistance and susceptibility that incorporate far fewer sequence positions than are associated with resistance in lookup tables. Several decision trees for drugs in the same drug class resemble each other, reflecting cross-resistance.

Support vector machines. Because the input space X for a support vector machine[5] is a real vector space, we map protein sequences onto vectors by introducing 20 indicator variables (one for each amino acid) for each sequence position and an additional variable for insertions in the RT. So, each sequence is represented by a vector $x = (x^1, \ldots, x^n)$ of dimension $n = 99 \cdot 20 = 1{,}980$ for the protease and $n = 250 \cdot 20 + 1 = 5{,}001$ for the RT. An indicator variable x^j for a certain sequence position and a certain amino acid is set to one if the amino acid is found at that position, and to zero otherwise. For ambiguous and unknown positions, we denote W as the set of possible amino acids (either those derived from translating wobbles in the nucleotide sequence or all amino acids if the position is completely unknown). To indicate the presence of an amino acid that appears in W, we assign the value $1/|W|$ to each variable.

For support vector machines, the output space is usually denoted $Y = \{-1, +1\}$. So, we assign resistant to the positive class and susceptible to the negative class. Thus, this technique trains on the set

$$\left\{(x_i, y_i) \middle| x_i = (x_i^1, \ldots, x_i^n), y_i = \pm 1\right\}_{i=1, \ldots, m}$$

of m samples.

We try to solve the classification problem in X by learning a linear function

$$f(x) = w \cdot x + b,\ w \in \mathbf{R}^n,\ b \in \mathbf{R},$$

such that

$$\text{sign}\, f(x_i) = y_i,\ i = 1, \ldots, m.$$

Geometrically, this amounts to separating X into two parts representing the two classes with a hyperplane with normal w and distance b from the origin (see Figure 4).

Suppose such separating hyperplanes exist. We define the margin of a separating hyperplane as the minimum distance between all points x_i and the hyperplane. The learning strategy of support vector machines is to choose as the classifier a hyperplane with the maximal margin. Although this might seem a reasonable heuristic approach, it is actually well founded in statistical learning theory. Basically, bounds on the generalization error are given in terms of the classification function's complexity, which we can minimize for linear functions by maximizing the margin. In particular, these bounds do not depend on the dimension of the input space X, which makes this strategy appealing for high-dimensional data.

To modify this approach for linearly inseparable data, we introduce slack variables and simultaneously maximize the margin and minimize the classification error on the training set. This includes introducing a regularization parameter C that controls the trade-off between these two objectives. The resulting optimization problem is a quadratic program, usually solved in the form of its Lagrangian dual:

$$\min_{\alpha} \quad \sum_{i=1}^{m}\sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j x_i \cdot x_j - \sum_{i=1}^{m} \alpha_i$$

$$\text{s.t.} \quad \sum_{i=1}^{m} y_i \alpha_i = 0$$

$$0 \le a_i \le C, \quad i = 1, \cdots, m.$$

To solve this problem, we use Joachim's SVM^{light}.[6] For each drug, we obtain a linear decision function

$$f(x) = \sum_{i=1}^{m} y_i \alpha_i x_i \cdot x + b$$

(see Figure 4). This knowledge representation is not as easy to interpret as a decision tree. However, the coefficients α_i provide some information about how much influence each training sample has on the decision

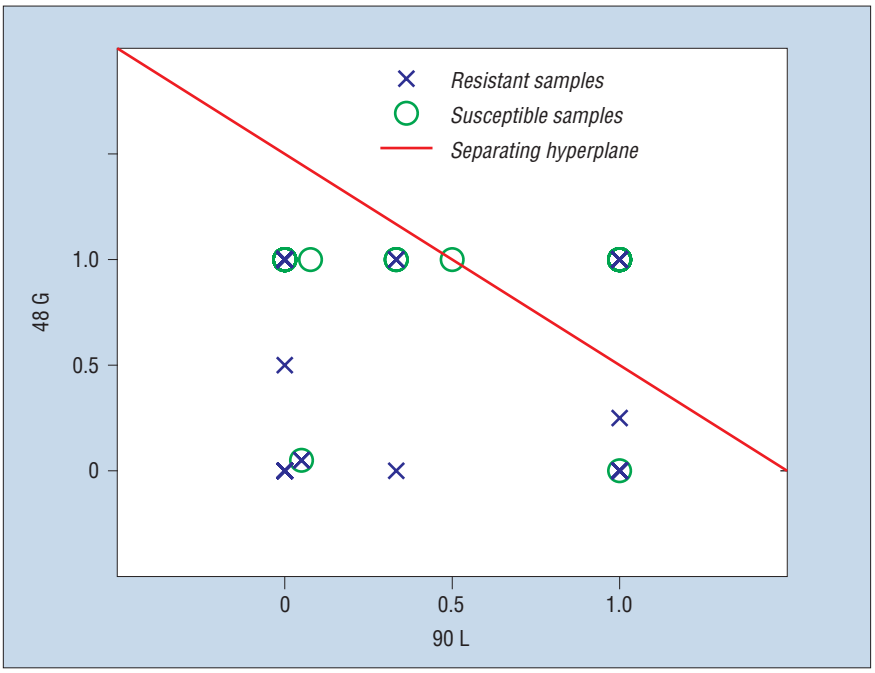

Figure 4. A simplified example of linear separation. Sequences have been mapped into a 2D vector space spanned by the indicator variables for leucine (L) at protease position 90 and glycine (G) at position 48. Both variables indicate *wild-type* (the standard form). Blue crosses indicate saquinavir-resistant samples (defined by a resistance factor *RF* ≥ 3.5); green circles indicate susceptible samples. The quadratic program's solution is the separating hyperplane $\{x \in \mathbf{R}^2 \mid (-2, -2)^t \cdot x - 3 = 0\}$, shown in red. This simplified classifier is estimated to predict resistance for unseen sequences with a 14.4 percent error rate.

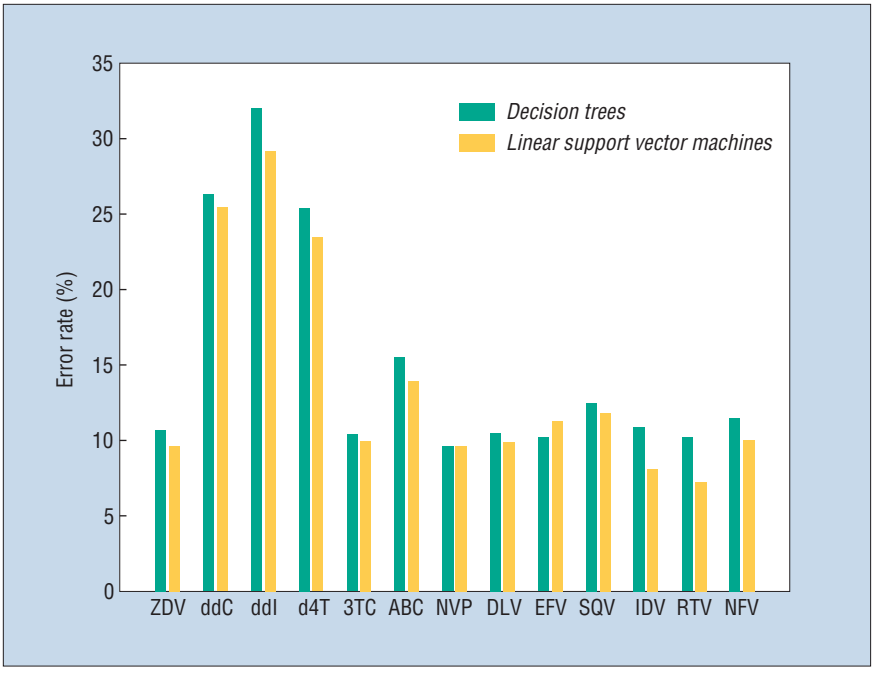

Figure 5. Leave-one-out error rates for the decision tree classifiers (green) and linear support vector machines (yellow). The class resistant was defined by a resistance factor *RF* ≥ 8.5 for the drugs ZDV, 3TC, NVP, DLV, and EFV; *RF* ≥ 2.5 for ddC, ddI, d4T, and ABC; and *RF* ≥ 3.5 for SQV, IDV, RTV, and NFV.

function. Points x_i with $\alpha_i > 0$ are *support vectors*. In the linearly separable case, they are exactly those points whose distance to the hyperplane is equal to the margin.

Validation. We estimated the generalization error of the two different families of models in leave-one-out experiments (see Figure 5). For the decision tree classifiers, error rates ranged between 9.6 and 15.5 percent for all drugs except for the nucleoside analogs ddC, ddI, and d4T, which had error rates of 25.4 to 32.0 percent. For nine out of 13 drugs, error rates were below 12.5 percent. Thus, with the exception of ddC, ddI, and d4T, decision trees were able to extract genotypic patterns that can reliably predict phenotypic resistance.

Linear separation with support vector machines performed slightly better for all but one drug, but the difference was not significant (t-test, 95 percent confidence level). For most drugs, leave-one-out testing estimated generalization errors between 7.3 and 12.4 percent, but ddC, ddI, and d4T again showed exceptionally high error rates of 23.5 to 29.2 percent. Remarkably, these latter drugs also showed exceptionally high coefficients of variation in the phenotypic assay.[2]

Web implementation

We have implemented geno2pheno as a Web-based system (http://cartan.gmd.de/geno2pheno.html). When users submit a DNA sequence coding for an HIV-1 protease or RT, the system both returns an alignment to a reference strain and reports insertions, deletions, and substitutions. Users can define the classes to be predicted by modifying cutoff values and can choose one of the two prediction methods we describe in this article. While the support vector machines probably produce more reliable predictions, the decision tree predictions come with a confidence factor. So, both predictions are valuable.

Decision trees appear to be appropriate for phenotype prediction because they can easily handle discrete data and unknown attribute values. Researchers have previously applied them to protein classification tasks such as protein secondary-structure prediction.[7] Physicians and biomedical researchers can easily interpret decision tree models. The extracted knowledge encoded in trees can easily be transformed into rules, which makes it applicable in different contexts such as the CTSHIV system (see the "Related Work" sidebar) or the development of new, potent drugs.

Support vector machines do not allow for knowledge extraction so naturally, but they improve the prediction quality. A good learning strategy and an efficient optimization algorithm make this approach practical for the high-dimensional sequence data. We have not yet exploited support vector machines' full flexibility. Learning a broad class of nonlinear decision functions (including, for example, polynomial and radial basis functions) would be straightforward; it involves replacing the quadratic program's inner product with a nonlinear kernel function.[5] We hope to further improve the system's predictive power by using an appropriate kernel function.

The ultimate goal of interpreting genotypic data is to optimize therapies for the individual patient. To do this will require not just predicting phenotypic drug resistance, as we've described here, but also determining cutoff values that predict therapy failure. However, these prediction methods promise to be useful when researchers directly investigate correlations between sequence variations and therapy response by incorporating additional data such as viral load measurements and drug treatment histories. ■

Acknowledgments

Financial support for this research came from the Bayerisches Staatsministerium für Kultus, Erziehung und Wissenschaft, the Deutsche Forschungsgemeinschaft, and the Robert Koch Institute, Berlin.

References

1. A.-M. Vandamme, K. Van Laethem, and E. De Clercq, "Managing Resistance to Anti-HIV Drugs: An Important Consideration for Effective Disease Management," *Drugs*, vol. 57, no. 3, Mar. 1999, pp. 337–361.
2. H. Walter et al., "Rapid, Phenotypic HIV-1 Drug Sensitivity Assay for Protease and Reverse Transcriptase Inhibitors," *J. Clinical Virology*, vol. 13, nos. 1–2, June 1999, pp. 71–80.
3. D. Wang, S. Bloor, and B.A. Larder, "The Application of Neural Networks in Predicting Phenotypic Resistance from Genotypes for

The Authors

Niko Beerenwinkel works in the computational biology group at the Max Planck Institute for Informatics. He concentrates on the development and application of machine learning techniques to identify and characterize genotype–phenotype relations in HIV, to optimize therapy. He received his diploma in mathematics at the University of Bonn and joined the Institute for Algorithms and Scientific Computing at the German National Research Center for Information Technology as a bioinformatics PhD student. Contact him at the Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, D-66123, Saarbrücken, Germany; niko.beerenwinkel@mpi-sb.mpg.de.

Barbara Schmidt heads a research group and works as the coordinator of the German National Reference Center for Retroviruses. Her research focuses mainly on genotypic and phenotypic resistance, the development and comparison of systems for the prediction of phenotype from genotypic data, and the evaluation of resistance profiles of new antiretroviral drugs. She received her MD with a thesis in vascular surgery from the University of Erlangen-Nürnberg. Contact her at the Univ. of Erlangen-Nürnberg, Schlossgarten 4, D-91054 Erlangen, Germany; baschmidt@viro.med.unierlangen.de.

Hauke Walter is a scientific assistant in the Department for Diagnostic Services at the Institute of Clinical and Molecular Virology of the University of Erlangen-Nürnberg. His work focuses on the comparison of phenotypic and genotypic HIV resistance, the comparison of interpretation systems for genotypic HIV resistance data, and changes in fitness of the virus due to drug resistance. He received his MD from the University of Erlangen-Nürnberg with a thesis in HIV drug resistance. Contact him at the Univ. of Erlangen-Nürnberg, Schlossgarten 4, D-91054 Erlangen, Germany; hewalter@viro.med.uni-erlangen.de.

Rolf Kaiser heads the Virus Resistance Group in the University of Cologne's Institute of Virology, where he investigates HIV's resistance to antiretroviral drugs. He previously investigated the evolution of HIV from a unique source of infection and the safety of blood products. He received his PhD in human genetics from the University of Bonn. Contact him at the Inst. of Virology, Univ. of Cologne, Fürst Pückler Str. 56, D-50935 Köln, Germany; rolf.kaiser@medizin.uni-koeln.de.

Thomas Lengauer is a director at the Max Planck Institute for Informatics. His major research interests include protein structure and function prediction and computational drug screening and design. Previously, he was a full professor at the University of Paderborn and a director of the Institute for Algorithms and Scientific Computing at the German National Research Center for Computer Science. He received his PhD in mathematics from the University of Berlin and his PhD in computer science from Stanford University. Contact him at the Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, D-66123 Saarbrücken, Germany; lengauer@mpi-sb.mpg.de.

Daniel Hoffmann heads the protein-folding group at the Center of Advanced European Studies and Research. He works on developing methods for structural biology and proteomics (the study of protein expression patterns) using combinations of experimental and theoretical methods. He also teaches courses on molecular modeling and bioinformatics at the University of Bonn. He previously worked in Thomas Lengauer's bioinformatics group at the German National Research Center for Information Technology. He received his Diploma in physics from the University of Heidelberg and earned his PhD in chemistry at the Free University of Berlin's Institute of Crystallography and Department of Chemistry. Contact him at the Center of Advanced European Studies and Research, Friedensplatz 16, D-53111 Bonn, Germany; daniel.hoffmann@caesar.de.

Klaus Korn is the head of diagnostic services at the Institute of Clinical and Molecular Virology at the German National Reference Center for Retroviruses. His major research interests are the development of new diagnostic tests, the molecular epidemiology of viral infections, and the investigation of antiviral drug resistance with particular emphasis on HIV and the hepatitis B virus. He previously was a scientific assistant at the Institute of Clinical and Molecular Virology at the University of Erlangen-Nürnberg. He earned his MD from the Johann Wolfgang Goethe University in Frankfurt. Contact him at the Inst. of Clinical and Molecular Virology, German Nat'l Reference Center for Retroviruses, Univ. of Erlangen-Nürnberg, Schlossgarten 4, D-91054 Erlangen, Germany; kskorn@viro.med.uni-erlangen.de.

Joachim Selbig is a research scientist and a project leader in Thomas Lengauer's bioinformatics group at the German National Research Center for Information Technology. He also teaches courses on selected topics of molecular modeling and bioinformatics at the University of Bonn and at bioinformatics summer schools. His main research interest is the application of machine learning and applied graph theory to biochemical and biophysical problems. He has contributed to several bioinformatics tools for protein structure analysis and prediction. He received his Diploma in physics from the University of Leipzig and his PhD in computer science from the Academy of Sciences in Berlin. Contact him at the Fraunhofer Inst. for Algorithms and Scientific Computing, Schloss Birlinghoven, D-53754 Sankt Augustin, Germany, joachim.selbig@scai.fraunhofer.de.

HIV-1 Protease Inhibitors," *Antiviral Therapy*, vol. 5, supplement 3, 2000, pp. 51–52.

4. J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Francisco, 1993.

5. C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, June 1998, pp. 121–167.

6. T. Joachims, "Making Large-Scale Support Vector Machine Learning Practical," *Advances in Kernel Methods: Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, eds., MIT Press, Cambridge, Mass., 1999, pp. 169–184.

7. J. Selbig, T. Mevissen, and T. Lengauer, "Decision Tree-Based Formation of Consensus Protein Secondary Structure Prediction," *Bioinformatics*, vol. 15, no. 12, Dec. 1999, pp. 1039–1046.