Antiviral Research 106 (2014) 5–12

Contents lists available at ScienceDirect

Antiviral Research

journal homepage: www.elsevier.com/locate/antiviral

Structure-based predictors of resistance to the HIV-1 integrase inhibitor Elvitegravir

Majid Masso *, Grace Chuang, Kate Hao, Shinar Jain, Iosif I. Vaisman

Laboratory for Structural Bioinformatics, School of Systems Biology, George Mason University, 10900 University Boulevard MS 5B3, Manassas, VA 20110, USA

ARTICLE INFO

Article history:
Received 14 January 2014
Revised 14 March 2014
Accepted 17 March 2014
Available online 25 March 2014

Keywords:
Integrase
Elvitegravir
Genotype–phenotype correlation
Resistance
Computational mutagenesis
Statistical learning models

ABSTRACT

The enzyme integrase (IN) of human immunodeficiency virus type 1 (HIV-1) mediates integration of reverse transcribed viral DNA into the human genome, an essential step in the HIV-1 replication cycle. Elvitegravir (EVG) is an HIV-1 strand transfer inhibitor that binds IN and is the second drug in its class to be approved for clinical use in combination with other anti-HIV-1 medications. However, certain IN sequence mutational patterns have an effect on inhibitor binding, thereby altering the degree of IN mutant susceptibility to EVG. Employing a dataset of 115 translated IN sequences, each having a known EVG susceptibility value and consisting of a distinct set of amino acid replacements relative to the native IN, here we develop and evaluate statistical learning models for predicting the phenotypes (i.e., quantified EVG susceptibilities) of new IN mutants based solely on their genotypes (i.e., translated IN sequences). Each IN mutant is represented as a feature vector of structure-based attributes obtained via an *in silico* mutagenesis procedure that quantifies all anticipated IN residue-specific environmental perturbations from wild type upon mutation. Cross-validated performance based on four classification models show that balanced accuracy reaches 87%, while two regression models yield a Pearson's correlation coefficient as high as $r = 0.78$. At the present time, the models may potentially be useful for diagnostic purposes, but only in conjunction with other tools and techniques, including experimental phenotype assays. However, as published experimental phenotypes for new IN variants become available, a larger and more diverse training set will likely lead to significantly more accurate models.

1. Introduction

Current treatment strategies for human immunodeficiency virus type 1 (HIV-1) infection utilize combinations of drugs designed to tightly bind and inhibit proteins essential for viral replication (Cortez and Maldarelli, 2011). Cocktails that include inhibitors of HIV-1 protease and reverse transcriptase enzymes were among the first to cause a pronounced and enduring reduction of patient viral burdens, and they continue to be prescribed as highly effective options (Vella et al., 2012). Though specific amino acid substitutions have been identified in these proteins that diminish patient susceptibility to certain medications, the arsenal of all currently approved medications in each drug class now have dissimilar resistance profiles, offering patients opportunities for combating the emergence of drug resistance mutations (Johnson et al., 2013). Also, treatments are now available to interfere with HIV-1 binding and fusion to susceptible host cells, as well as with integration of the viral DNA into the human genome (Cortez and Maldarelli, 2011). The enzyme HIV-1 integrase (IN) catalyzes the latter reaction, and IN protein mutations conferring drug resistance have been identified in recent years both *in vivo* (clinic/patient) and *in vitro* (laboratory) (Stanford University HIV Drug Resistance Database, 2013). As the second medication in its class to be approved by the US Food and Drug Administration (FDA) in August 2012, the integrase inhibitor (INI) Elvitegravir (EVG) is presently only available as part of a specific 4-drug regimen packaged into a single tablet and marketed under the brand name Stribild (Arribas and Eron, 2013; Belavic, 2013). This manuscript details our development of IN structure-based predictive models of EVG resistance.

Prior to prescribing a particular multi-drug regimen to the patient, the clinician may order tests to help with selecting effective medications. Genotypic tests identify mutations in the patient's HIV-1 viral gene sequences, corresponding to single residue replacements in their translated protein products, that are already known to cause resistance to certain drugs (Cortez and Maldarelli, 2011). Phenotypic tests, on the other hand, quantify the degree to which each drug inhibits its respective mutated HIV-1 target

* Corresponding author. Tel.: +1 703 257 5756.
E-mail address: mmasso@gmu.edu (M. Masso).

http://dx.doi.org/10.1016/j.antiviral.2014.03.006
0166-3542/

relative to the fully susceptible wild type protein, irrespective of the complexity of the mutational patterns (Cortez and Maldarelli, 2011). Since phenotyping is relatively more expensive than genotyping and has a longer turnaround time, there is substantial interest in techniques that are capable of accurately predicting an unknown phenotype (i.e., degree of susceptibility to each drug) from the available genotype (i.e., genetic sequence encoding the target protein) (Prosperi and De Luca, 2012).

Here we consider a dataset of 115 translated HIV-1 IN genotypes for which quantified phenotype values are available in terms of susceptibility to EVG, with a specific focus on gene sequences that contain mutations corresponding exclusively to amino acid replacements in the IN protein products (i.e., no indels). In each case, the mutated protein sequence was threaded onto the native 3-dimensional IN structure, and a computational mutagenesis technique was employed to quantify ensuing environmental perturbations relative to wild type at all IN residue positions (Fig. 1). The IN variants were uniquely represented as feature vectors with components consisting of these residue-specific perturbations, and the dataset was used to train predictive models of resistance to EVG by implementing four supervised classification and two regression statistical learning algorithms. Cross-validation studies were performed in order to gauge the performance of the models. By using the *in silico* mutagenesis procedure to similarly generate feature vectors for new IN variants whose experimental phenotypes become available in future published studies, additional validation studies can be undertaken by using the models to predict their phenotypes. Other structure-based methods, employing well-known molecular dynamics simulation and energy minimization techniques, have also been applied recently to studying EVG resistance (Chen et al., 2013; Johnson et al., 2012; Xue et al., 2013). This study builds upon the prior successful application of our distinct structure-based approach toward developing accurate predictive models of both HIV-1 co-receptor usage (Masso and Vaisman, 2010a), as well as HIV-1 drug resistance to eight protease and eleven reverse transcriptase inhibitors (Masso and Vaisman, 2013), illustrating a "proof-of-principle" regarding the effectiveness of our technique for modeling drug resistance with the IN target structure despite a limited number of currently available dataset samples.

2. Materials and methods

2.1. Data

A total of 141 translated IN genotypes, corresponding to laboratory (Abram et al., 2013; Huang et al., 2013; Jones et al., 2007; McColl and Chen, 2010; Mesplede et al., 2013) and clinical (Huang et al., 2013; McColl and Chen, 2010) HIV-1 isolates described in 5 published studies, were obtained from the Stanford University Drug Resistance Database (Stanford University HIV Drug Resistance Database, 2013). The IN sequences are distinct, each consisting of a mutational pattern defined by amino acid residue substitutions at select positions relative to that of the native IN protein. However, for each of 24 pairs of sequences reported in one of these studies (Huang et al., 2013), the only difference between the sequences in each pair is the existence of an additional single residue substitution (S230R) in one of the sequences. This minor accessory mutation nevertheless has an impact on IN susceptibility to EVG, and since our approach utilizes a solved structure of the IN catalytic core and EVG-binding domain that excludes the C-terminus beyond position 209 (see below), these 24 sequences containing the S230R substitution were excluded. A similar situation arose for two additional pairs of sequences appearing in other studies, leading to the exclusion of 26 sequences in total, so that the final dataset employed here consists of 115 distinct IN protein sequences. Susceptibility of each IN variant to EVG was reported as a fold change (FC) value obtained using the PhenoSense assay (Monogram Biosciences, South San Francisco, CA) (Petropoulos et al., 2000). The FC value is defined as the ratio of 50% inhibitory concentration (IC_{50}) for the mutant relative to that for a drug-sensitive, wild type control. All FC values were log-transformed prior to analysis.

2.2. Computational mutagenesis methodology

The *in silico* procedure relies on a previously developed four-body, knowledge based, statistical potential. Briefly, the PISCES protein culling server (Wang and Dunbrack, 2003) was used to identify 1417 single protein chains, from high-resolution ($\leqslant$2.2 Å) X-ray crystal structures deposited in the Protein Data Bank (PDB) (Berman et al., 2000), which share low primary sequence identity

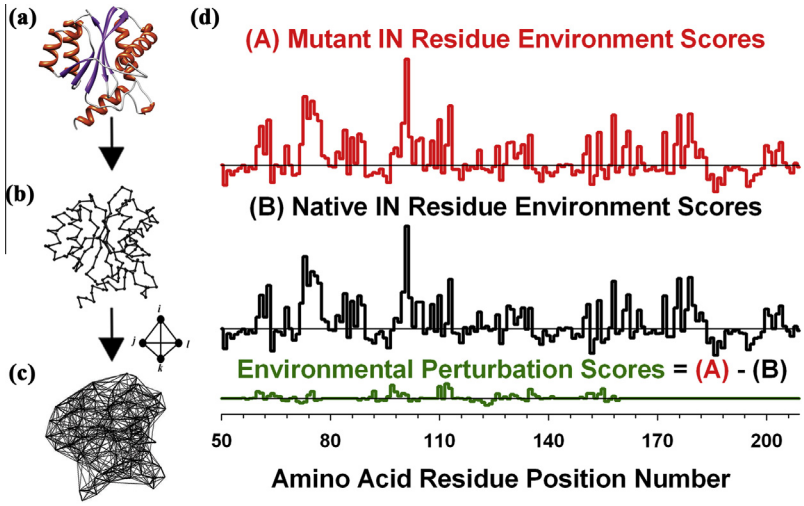

Fig. 1. HIV-1 integrase (IN) structure visualization, Delaunay tessellation, and mutant representation. (a) Ribbon diagram and (b) C-alpha trace of the HIV-1 IN catalytic domain structure (PDB accession code: 1bl3C; residues M50 – Q209). (c) C-alpha Delaunay tessellation of the IN structure superimposed over the trace. (d) Profiles of residue environment scores both for the native HIV-1 IN protein and for an IN mutant defined by the eight amino acid substitutions E92Q, T97A, L101I, T112V, S119P, T124A, I135V, and N155H. The difference (mutant – native) between both profiles defines the ensuing environmental perturbation scores upon mutation at all 160 residue positions in the IN structure.

(<30%). Each protein chain was coarse-grained at the residue level using the C-alpha coordinates, and the point-set was analyzed with the Qhull software package (Barber et al., 1996; Qhull). Qhull treats the points representing a structure as vertices and generates a convex hull of non-overlapping, space-filling, and irregular tetrahedra known in the field of computational geometry as a Delaunay tessellation (de Berg et al., 2008). Each tetrahedron objectively identifies at its four vertices a quadruplet of nearest neighbor residues in the protein (Fig. 1(c)). To ensure the exclusion of potentially false-positive quadruplet interactions, all edges longer than 12 Å were removed from each tessellation prior to analysis (Masso and Vaisman, 2010b,c). The Delaunay tessellation of an average-sized protein chain (~200 amino acids), coarse-grained at the residue level, consists of several hundred tetrahedra.

The four vertices of a tetrahedron in a protein tessellation clearly may represent repeated instances of amino acid residues (e.g., the close proximity of CCHH in many zinc finger proteins). Additionally, since we do not order the four vertices of any tetrahedron (e.g., by primary sequence number), a single representative is used to denote all possible permutations of the four letters constituting each residue quadruplet (e.g., place letters in ascending order). Given these conditions, a total of 8855 distinct four-letter subsets can be formed using a standard twenty-letter protein alphabet (Masso and Vaisman, 2008, 2010b–d). For each subset (i,j,k,l), an observed relative frequency of occurrence f_{ijkl} was calculated as the proportion of tetrahedra generated by all 1417 protein tessellations having that particular residue quadruplet represented by the four C-alpha vertices. A rate p_{ijkl} expected by chance was computed using a multinomial reference distribution, given by

$$p_{ijkl} = \frac{4!}{\prod_{n=1}^{20}(t_n!)}\prod_{n=1}^{20}a_n^{t_n}, \text{where} \sum_{n=1}^{20}a_n = 1 \text{ and } \sum_{n=1}^{20}t_n = 4.$$

In the above formula, a_n represents the proportion of amino acids of type n in the 1417 tessellated protein chains, and t_n is the number of occurrences of amino acid type n in the quadruplet. Based upon an application of the inverted Boltzmann principle (Sippl, 1993), the score $S_{ijkl} = -\log(f_{ijkl}/p_{ijkl})$ is proportional to the energy of residue quadruplet interaction, and the collection of 8855 distinct residue quadruplets with their respective scores defines the four-body statistical potential (Masso and Vaisman, 2008, 2010b–d).

A computational mutagenesis technique, employing both Delaunay tessellation and the four-body statistical potential, was applied to an X-ray structure at 2.0 Å resolution for the EVG-binding catalytic domain of HIV-1 integrase (IN) (Fig. 1(a)), consisting of residue positions M50 – Q209 (PDB accession code: 1bl3C) (Maignan et al., 1998). Upon tessellating the IN structure and filtering out edges longer than 12 Å (Fig. 1(c)), each remaining tetrahedron was assigned a score using the statistical potential based on the identity of the four residue letters at its vertices. Note that each C-alpha point in a tessellation is typically shared as a vertex by multiple adjacent tetrahedra. By adding up the scores of all tetrahedra that share a particular vertex, a residue environment score is obtained for the protein sequence position that the point represents, and such environment scores were computed for all 160 positions of the tessellated IN structure. For any IN variant defined by amino acid substitutions, this new protein sequence of identical length was threaded onto the native tessellated IN structure by re-labeling residue letters as needed at each of the 160 C-alpha vertices, and residue environment scores were subsequently recalculated for each position. The difference (mutant – wild type) between both residue environment scores at each position empirically quantifies the environmental perturbation upon mutation (Fig. 1(d)) Masso and Vaisman, 2008, 2010b–d, and these sequence position-specific perturbation scores form the attributes of a feature vector for the IN mutant.

2.3. Statistical learning and evaluation of performance

Using the Weka software package (Frank et al., 2004), we implemented four supervised classification (random forest, RF; support vector machine, SVM; decision tree, DT; and neural network, NN) and two regression (reduced-error pruned tree, REPTree; and support vector regression, SVR) statistical learning algorithms to assess how effectively the signals encoded by the attributes of the IN mutant vectors, generated by our *in silico* mutagenesis procedure, were capable of modeling the corresponding changes to EVG susceptibility. Relevant non-default parameter values employed in Weka included 100 trees (iterations) for RF; radial basis function (RBF) kernel for SVM and SVR; one hidden layer of nodes for NN; and thirty bagged (*b*ootstrap *a*ggregated) iterations for REPTree.

Cross-validation testing was used for gauging model performance. In particular, our trained models were evaluated via leave-one-out cross-validation (LOOCV) and tenfold cross-validation (10-fold CV) procedures (Witten and Frank, 2005); however, here we report only the LOOCV performance data, as classifier tuning based on 10-fold CV produced nearly identical results. With respect to our dataset, LOOCV is equivalent to 115-fold CV (i.e., each IN mutant forms a singleton), whereas 10-fold CV testing is initialized by randomly grouping the IN mutants into 10 roughly equal-sized subsets. For both methods, testing proceeds by holding-out one of the subsets, training a model using the samples from all of the other subsets combined, and then using the model to predict the phenotypes of the samples in the held-out subset. The process is iterated so that each subset serves once as the hold-out in order to have its samples predicted, and performance is then based on comparing the actual and predicted phenotypes over all 115 IN mutants in the dataset. Clearly, LOOCV is deterministic and requires the execution of only a single iteration, while 10-fold CV is best implemented by averaging the performance results over several runs (e.g., 10 iterations of 10-fold CV) to capture any variability introduced by the way in which the 115 samples are initially grouped into 10 subsets (Witten and Frank, 2005).

For the supervised classification algorithms, testing was evaluated by referring to the classes as positive (S) and negative (R), so that TP (TN) represent the total number of correct S (R) predictions, while FN (FP) correspond to the total number of respective misclassifications. Using this notation, we computed ACC = overall accuracy = (TP + TN)/(TP + FP + TN + FN), Se(S) = sensitivity = TP/(TP + FN), Sp(S) = specificity = TN/(TN + FP), and PPV(S) = positive predictive value (or precision) = TP/(TP + FP). Additionally, we calculated and reported the following quantities: the balanced accuracy rate (BAR), given by BAR = 0.5 × [Se(S) + Sp(S)]; Matthew's correlation coefficient (MCC), calculated as

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FN})(\text{TP}+\text{FP})(\text{TN}+\text{FN})(\text{TN}+\text{FP})}};$$

and the area (AUC) under the receiver operating characteristic (ROC) curve, with the latter spanning values from approximately AUC = 0.5 (random guessing) to AUC = 1.0 (perfect classifier). In the case of regression algorithms, testing was evaluated with the value of Pearson's correlation coefficient (r) between actual and predicted EVG drug susceptibility values for the IN mutants, the mean-squared error (*mse*), and the overall accuracy (ACC) of mutant classifications based on their predicted values.

3. Results

3.1. EVG susceptibility and IN attributes

PhenoSense assay results are reported based on an EVG biological cutoff of FC = 2.5, above which susceptibility to EVG is expected to progressively decrease with higher FC values (http://www.monogrambio.com/wp-content/uploads/2013/04/PS_IN_report_Watermark.pdf). Our dataset includes 115 translated IN genotypes, each with an experimentally known EVG FC phenotype obtained with the PhenoSense assay, and we incorporated three factors in deciding how to classify these IN samples as susceptible (S) or resistant (R) to EVG. First, the Stanford database utilizes the following categories: susceptible (FC < 4), intermediate (4 < FC < 20), and resistant (FC > 20) (Stanford University HIV Drug Resistance Database, 2013). Next, McColl et al. (2007) reported that patient virus obtained from virologic failure samples (i.e., those not responding to EVG treatment) showed a mean EVG FC > 151 (range: 1.21–301). Finally, the PhenoSense assay is reported to display up to a 2-fold variation in FC measurements among multiple replicates (Huang et al., 2013). Combining this information, and given that EVG susceptibility is not binary in nature but ranges across a continuum of values, we decided to use a cutoff of FC = 20 to distinguish between IN samples that are susceptible to EVG to any degree and those that are completely resistant to the drug at the highest levels. Fig. 2 (adapted from Van der Borght et al. (2013)) provides a distribution of these phenotypes and marks the location of the FC = 20 (or $\log_{10}$(FC) = 1.3) cutoff, whereby 74 (64%) of the IN mutants appear below the cutoff and are classified as susceptible (S) to EVG, while 41 (36%) are above the cutoff and considered to be completely resistant (R) to the drug. In particular, only five of the 115 IN samples in the dataset are within a 2-fold window of FC = 20, which minimizes any discordance in the classification results due to the imposition of a strict binary cutoff.

Models were developed and evaluated based on the representation of IN mutants as feature vectors in four distinct ways. First, each IN mutant was represented as a 160-dimensional vector of the environmental perturbation scores, obtained using our *in silico* mutagenesis procedure, at all residue positions appearing in the tessellated IN protein structure. Since feature vector dimensionality (i.e., number of input attributes) is greater than the number of samples, the evaluation of classifier performance can be biased. To investigate whether that is the case here, and more generally to develop models that indeed yield reliable performance measures, we employed both biological as well as algorithmic feature subset selection approaches. With respect to biological feature selection, a subset of the 160 input attributes was selected to generate an alternative 12-dimensional feature vector for each IN mutant, corresponding to perturbation scores only at documented EVG drug resistance residue positions (66, 92, 97, 121, 138, 140, 143, 147, 148, 153, 155, 157) in the IN structure (Johnson et al., 2013; Shafer and Schapiro, 2008). Alternatively, we implemented an algorithm using Weka (Frank et al., 2004) that calculated the amount of information gain associated with each input attribute, and the 160 attributes were subsequently ranked according to these values. The ordered top 20-ranked attributes correspond to residue positions (156, 145, 153, 148, 155, 143, 140, 151, 65, 142, 154, 93, 120, 118, 119, 149, 75, 67, 139, 152), which include five of the 12 EVG drug resistance positions as well as 10 others that are covalently linked through peptide bonds to such positions, and perturbation scores at these positions were extracted from the 160-dimensional vectors and used to generate 20-dimensional feature vectors for the IN mutants. Lastly, we added the perturbation scores extracted from the next top-20 ordered attributes in the ranked list, corresponding to residue positions (91, 92, 141, 76, 146, 115, 72, 150, 158, 63, 62, 77, 70, 147, 66, 94, 95, 157, 88, 100) and providing four additional EVG drug resistance positions as well as four more covalent linkages to such positions, to create a 40-dimensional feature vector for each of the IN mutants.

In addition to the input attributes (i.e., the independent variables), each IN mutant was also equipped with an output attribute (i.e., the dependent variable) in the form of either an EVG $\log_{10}$(FC) value (regression) or an R/S EVG susceptibility category (supervised classification). By learning complex nonlinear functions based on the training data, the implemented algorithms are capable of making drug susceptibility predictions for new IN mutants based on the feature vectors obtained for them.

3.2. Classification

Table 1 summarizes the LOOCV performance results both by algorithm implementation method (RF, SVM, DT, and NN) as well as by the number of residue position-specific input attributes included in the IN mutant feature vectors (160, 40, 20 and 12). The four classifiers all performed well on our training set, with each classification method reporting relatively consistent results regardless of the dimensionality (i.e., number of input attributes) of the IN mutant feature vectors. In particular, note that use of 12-dimensional input attribute vectors for representing the IN mutants, based on our biological feature selection approach, led to the weakest performance results for each method. Also, a

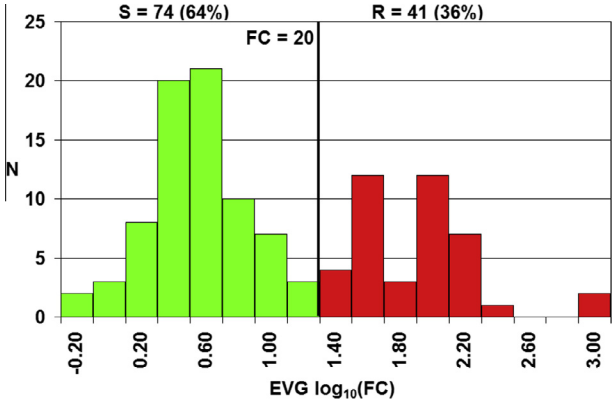

Fig. 2. Phenotype distribution of IN variants in the EVG genotype–phenotype dataset.

Table 1
Classifier performance.

Method	# Attributes	Se(S)	Sp(S)	PPV(S)	BAR	MCC	AUC
RF	160	0.92	0.76	0.87	0.84	0.69	0.87
	40	0.93	0.78	0.89	0.86	0.73	0.86
	20	0.92	0.78	0.88	0.85	0.71	0.87
	12	0.88	0.66	0.82	0.77	0.55	0.86
SVM	160	0.88	0.78	0.88	0.83	0.66	0.85
	40	0.88	0.83	0.90	0.85	0.70	0.86
	20	0.87	0.78	0.88	0.82	0.64	0.87
	12	0.87	0.73	0.85	0.80	0.60	0.85
DT	160	0.91	0.81	0.89	0.86	0.71	0.86
	40	0.89	0.76	0.87	0.82	0.66	0.85
	20	0.92	0.78	0.88	0.85	0.71	0.86
	12	0.89	0.71	0.85	0.80	0.61	0.85
NN	160	0.85	0.83	0.90	0.84	0.67	0.84
	40	0.91	0.83	0.91	0.87	0.73	0.88
	20	0.89	0.81	0.89	0.85	0.70	0.86
	12	0.84	0.68	0.83	0.76	0.52	0.76

comparison of results based on all 160 input attributes with those obtained by using subsets of these attributes resulting from algorithmic (20- and 40-dimensional vectors) or biological (12-dimensional vectors) feature selections suggests that representing each of the 115 IN mutants as a 160-dimensional vector does not lead to performance bias with this training set. Finally, given that the subset of 40 top-ranked attributes generally yields the best performance results with each of the four classifiers, and since this subset of attributes includes 9 of the 12 EVG drug resistance positions (sequence numbers 97, 121, and 138 were not included as they were ranked 73, 57 and 45, respectively) as well as 14 others that are covalently linked to such positions through peptide bonds, the results to follow in this section are based on this approach for representing the IN mutants.

Next, Fig. 3(a) offers a visual comparison of the ROC curves based on LOOCV testing that yield the AUC values presented in Table 1. Also shown are ROC curves obtained by randomly permuting the R/S class labels among the 115 IN mutants in the dataset prior to implementing LOOCV. The figure clearly depicts the degree to which the signals embedded among the 40 top-ranked attributes are capable of accurately discriminating between variants belonging to their respective R/S categories, since a control dataset based on a random R/S class label shuffling among the IN mutant feature vectors yields models that perform no better than random guessing (i.e., ROC curves along the diagonal, with AUC values in the vicinity of 0.5). The calculated BAR and MCC values for the random control dataset based on LOOCV testing further support this conclusion: RF (BAR = 0.50, MCC = −0.01), SVM (BAR = 0.56, MCC = 0.12), DT (BAR = 0.46, MCC = −0.10), and NN (BAR = 0.49, MCC = −0.07). Similar observations were made with respect to ROC curves corresponding to the AUC data presented in Table 1 for the other three input attribute vector sizes, and their subsequent comparison to results based on a control dataset generated by random shuffling the class labels among the 115 IN mutants (not shown).

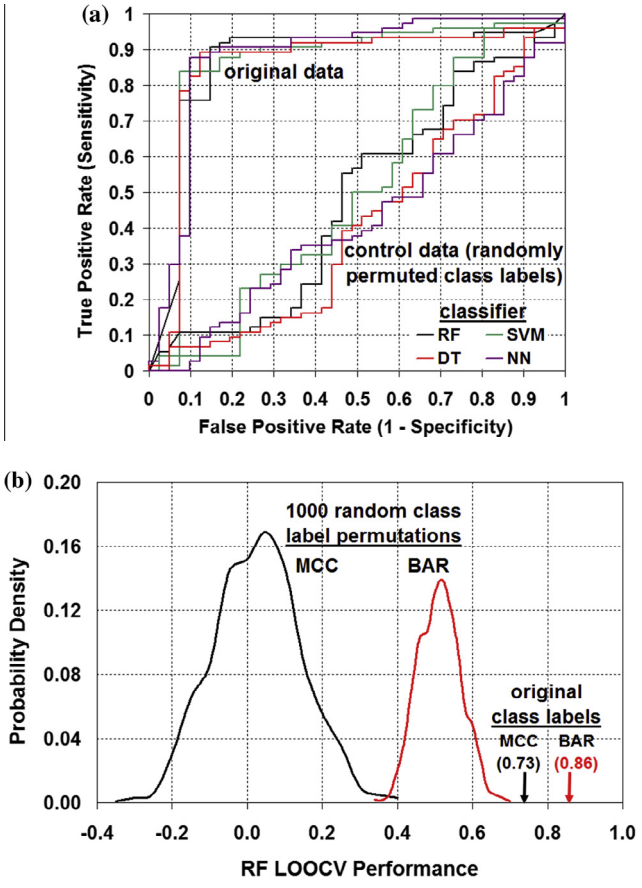

Fig. 3. (a) ROC curves based on LOOCV testing applied to both the dataset of IN mutants represented as feature vectors with 40 input attributes, as well as a control dataset obtained by randomly permuting R/S class labels among the IN mutants. (b) BAR and MCC permutation distributions based on applying the RF algorithm and LOOCV testing to 1000 control datasets (randomly permuted class labels) of IN mutants represented as feature vectors consisting of 40 input attributes.

For a more comprehensive approach that evaluates the statistical significance of results presented in Table 1, we generated 1000 control datasets via random class label shuffling among the 40-dimensional IN mutant feature vectors, assessing LOOCV performance in each case by implementing the RF algorithm. Presented in Fig. 3(b) are permutation distributions of the corresponding BAR (0.50 ± 0.06) and MCC (0.00 ± 0.12) values, both of which are centered around values indicative of random guessing and are significantly lower than comparable RF values obtained by using the actual class label arrangement in the original dataset (Table 1: BAR = 0.86 and MCC = 0.73); hence, the *p*-value for predictive power is 0.001. We obtained comparable results by implementing the other three classification algorithms, as well as by similarly applying IN mutant datasets that utilize the 160-, 20-, and 12-dimensional feature vectors upon which the performance data in Table 1 are based (not shown).

Lastly, we took a closer look at two of the models based on the DT algorithm, trained using the maximum and minimum number of attributes under consideration, whose explicit tree structures can be interpreted in a straightforward fashion. The DT models trained using either all 160 feature vector attributes (Fig. 4(a)) or the subset of 12 attributes corresponding only to EVG drug resistance positions (Fig. 4(b)) contain round tree nodes, each labeled with an IN residue number. Starting from the most information-rich root node at the top of each tree, a decision is made as to which of two directions an IN mutant follows based on the environmental perturbation score at that residue position in the feature vector for the variant. Rectangles represent the leaves of the trees, where the IN mutant is finally classified. A single number in parentheses after the class label at the leaf reports how many total variants reach the leaf, and if a secondary number is present, it provides a count of those variants that are misclassified (i.e., the resubstitution error occurring when the model trained using all 115 IN mutants is tested using the same data). The DT model based on all 160 attributes utilizes only one EVG drug resistance position (155) as a node, though all five remaining nodes correspond to positions (64, 67, 145, 146, 152) known to interact (i.e., share a tessellation edge less than 12 Å in length) with one or more of the 12 drug resistance positions in the IN structure, thus impacting their residue environment scores. On the other hand, the DT model trained using the 12-dimensional attribute vectors employs only four of the 12 key EVG drug resistance IN residue positions (92, 140, 147, 155) as nodes.

3.3. Regression

Unlike classification models, which predict IN mutants as either susceptible or resistant to EVG based upon their feature vectors of input attributes computed with the *in silico* mutagenesis procedure, regression models use the same IN mutant feature vectors to numerically predict EVG $\log_{10}$(FC) values. We trained regression models by implementing the REPTree and SVR algorithms, employing all four datasets of IN mutants (160-, 40-, 20-, and 12-dimensional feature vectors) and replacing the R/S class labels with the original log-transformed FC values.

Table 2 summarizes the results of LOOCV testing, reported using the Pearson's correlation coefficient (*r*) between the actual

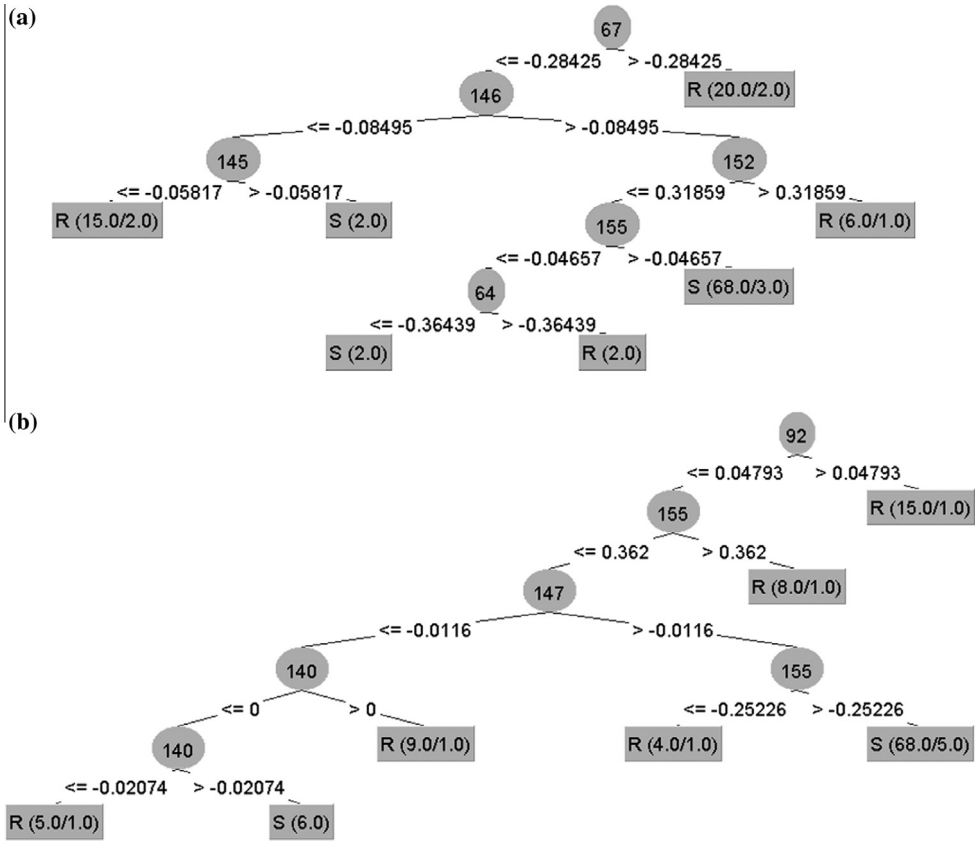

Fig. 4. DT models trained with the dataset of 115 IN mutants represented as feature vectors consisting of (a) all 160 residue-specific input attributes and (b) a subset of 12 input attributes that correspond to known EVG drug resistance residue positions.

Table 2
Regression performance.

Method	# Attributes	*r*	*mse*
REPTree	160	0.77	0.22
	40	0.77	0.22
	20	0.76	0.23
	12	0.69	0.28
SVR	160	0.78	0.20
	40	0.75	0.24
	20	0.73	0.25
	12	0.61	0.36

and predicted EVG $\log_{10}$(FC) values as well as the mean squared error (*mse*). Both methods display steady improvements in the correlation and concomitant *mse* decreases as the number of input attributes increases, with REPTree outperforming SVR at each step except for where all 160 attributes are utilized. In this specific case, the significant increase in correlation observed with SVR between 40 and 160 attributes, coupled with the corresponding REPTree results that reflect a plateau in performance, together suggest that a slight performance bias may be evident with SVR due to the fact that attribute dimensionality (160) exceeds the number of IN samples (115) as described earlier. Additionally, SVR performed significantly worse than REPTree based on the biological feature selection approach (i.e., 12 attributes) at the other extreme. These observations show REPTree to be the more stable and reliable method to be implemented on this particular dataset. Lastly, based on implementation of the REPTree algorithm with LOOCV testing on the 40-dimensional IN mutant feature vectors, Fig. 5 displays a scatter plot of the actual and predicted EVG $\log_{10}$(FC) values for the 115 IN variants.

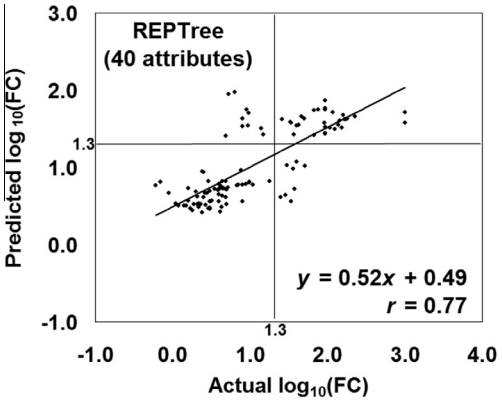

Fig. 5. Scatter plot of actual and predicted EVG $\log_{10}$(FC) values for the 115 IN variants, represented as 40-dimensional feature vectors, which was obtained by implementing the REPTree regression algorithm with LOOCV testing.

4. Discussion

The following procedure was implemented in order to directly compare the performance of the supervised classification and regression models. For regression models, predicted EVG $\log_{10}$(FC) values for the IN variants based on LOOCV testing were converted to predicted S/R classes based on their locations relative to the $\log_{10}(FC) = \log_{10}(20) = 1.3$ cutoff, and these were compared to the actual IN mutant class labels (i.e., those already used with the supervised classification models, obtained by using the experimental EVG $\log_{10}$(FC) values based on the PhenoSense assay with the 1.3 cutoff) in order to calculate overall prediction accuracy

Table 3
Overall accuracy (ACC).

Method	# Attributes				Avg.
	160	40	20	12	
RF	0.86	0.88	0.87	0.80	0.85
SVM	0.84	0.86	0.83	0.82	0.84
DT	0.86	0.84	0.86	0.83	0.85
NN	0.84	0.87	0.86	0.78	0.84
REPTree	0.84	0.84	0.87	0.79	0.84
SVR	0.84	0.82	0.86	0.78	0.83
Avg.	0.85	0.85	0.86	0.80	0.84

(ACC). Similarly, ACC performance was calculated for the four supervised classification algorithms based on LOOCV testing, simply by comparing the predicted and actual class labels corresponding to the 115 IN variants (i.e., bypassing the initial step of converting predicted values to classes as required for the regression models).

Averaging over all four supervised classification models and four input attribute sizes, the mean ACC (Table 3) is less than 1.5% (0.014375, range 0–3%) greater than the mean BAR (Table 1), supporting the assertion that despite an S/R class imbalance in the dataset, our models are optimized to avoid any potential bias for favoring predictions on the majority (S) class to the detriment of the minority (R) class. The data in Table 3 reinforces our earlier decision to focus on the IN mutant dataset that contains 40 input attributes in the analysis of supervised classification results, with 3/4 of the methods achieving the highest ACC values in this case; for DT, however, either 20 or 160 attributes yield a modest 2% increase in ACC over the use of 40 attributes (0.86 versus 0.84). Turning next to the regression methods, both REPTree and SVR achieved the highest ACC values with the use of only 20 input attributes, outperforming similar results obtained with 40 attributes by 3% and 4%, respectively (Table 3). Finally, in general all six statistical machine learning methods perform well and are relatively consistent with one another, especially when the focus is on the use of the 40- and 20-attribute subsets.

In summary, a computational mutagenesis technique was used to undertake a structure-based characterization of 115 translated HIV-1 integrase (IN) sequences containing mutations defined by amino acid replacements relative to the native IN. Each mutant IN protein was represented as a vector of 160 attributes that quantify environmental perturbations upon mutation at all constituent residue positions. These IN variants were selected for having undergone experimental phenotype testing to quantify their susceptibility to the inhibitor drug Elvitegravir (EVG) in terms of a fold change (FC), where increasing FC values correspond to decreasing susceptibility. Based on a number of considerations, a cutoff value of FC = 20 was used to categorize the IN mutants as either susceptible to any degree (S) or completely resistant (R) to EVG.

The dataset was used to train four classifiers as well as two regression models, which were evaluated based on leave-one-out cross-validation testing. Since the feature vector dimensionality (160) of each IN mutant exceeded the number of samples (115) in the dataset, both biological as well as algorithmic feature selection methods were employed, and additional models were developed and evaluated based on the use of significantly smaller subsets of the original 160 attributes. All methods performed quite well and were relatively consistent, with balanced accuracy as high as 87% among the supervised classification models and Pearson's correlation coefficient reaching 0.78 among the regression models.

These trained models can be used to generate EVG phenotype predictions for new IN variants not in the training dataset that are defined by amino acid substitutions at one or more positions. Given the speed with which both the *in silico* mutagenesis and the trained models generate results (run time is on the order of seconds), the methodology described here yields efficient EVG phenotype predictions. The current models may potentially be useful as a supplementary diagnostic tool in conjunction with existing and emerging diagnostic techniques and systems. With the future availability of a larger and more diverse training set of IN variants, significantly more accurate models can be developed for diagnostic purposes and to provide valuable insights into understanding the mechanisms of viral drug resistance.

Acknowledgements

The authors thank the Stanford University HIV Drug Resistance Database for compiling the genotype–phenotype data used in this study. G. Chuang, K. Hao, and S. Jain thank the Aspiring Scientists Summer Internship Program (ASSIP) at George Mason University for providing the opportunity to participate in this project.

References

Abram, M.E., Hluhanich, R.M., Goodman, D.D., et al., 2013. Impact of primary elvitegravir resistance-associated mutations in HIV-1 integrase on drug susceptibility and viral replication fitness. Antimicrob. Agents Chemother. 57 (6), 2654–2663.
Arribas, J.R., Eron, J., 2013. Advances in antiretroviral therapy. Curr. Opin. HIV AIDS 8 (4), 341–349.
Barber, C.B., Dobkin, D.P., Huhdanpaa, H.T., 1996. The quickhull algorithm for convex hulls. ACM Trans. Math. Software 22 (4), 469–483.
Belavic, J.M., 2013. Drug updates and approvals: 2012 in review. Nurse Pract. 38 (2), 24–42.
Berman, H.M., Westbrook, J., Feng, Z., et al., 2000. The protein data bank. Nucleic Acids Res. 28 (1), 235–242.
Chen, Q., Buolamwini, J.K., Smith, J.C., et al., 2013. Impact of resistance mutations on inhibitor binding to HIV-1 integrase. J. Chem. Inf. Model. 53 (12), 3297–3307.
Cortez, K.J., Maldarelli, F., 2011. Clinical management of HIV drug resistance. Viruses 3 (4), 347–378.
de Berg, M., Cheong, O., van Kreveld, M., et al., 2008. Computational Geometry: Algorithms and Applications. Springer-Verlag, Berlin, Germany.
Frank, E., Hall, M., Trigg, L., et al., 2004. Data mining in bioinformatics using Weka. Bioinformatics 20 (15), 2479–2481.
Huang, W., Frantzell, A., Fransen, S., et al., 2013. Multiple genetic pathways involving amino acid position 143 of HIV-1 integrase are preferentially associated with specific secondary amino acid substitutions and confer resistance to raltegravir and cross-resistance to elvitegravir. Antimicrob. Agents Chemother. 57 (9), 4105–4113.
Johnson, B.C., Metifiot, M., Pommier, Y., et al., 2012. Molecular dynamics approaches estimate the binding energy of HIV-1 integrase inhibitors and correlate with in vitro activity. Antimicrob. Agents Chemother. 56 (1), 411–419.
Johnson, V.A., Calvez, V., Gunthard, H.F., et al., 2013. Update of the drug resistance mutations in HIV-1: March 2013. Top. Antivir. Med. 21 (1), 6–14.
C. Jones, R. Ledford, F. Yu, et al., "Resistance profile of HIV-1 mutations in vitro selected by the HIV-1 integrase inhibitor, GS-9137 (JTK-303)". In: Presented at the 14th Conference on Retroviruses and Opportunistic Infections, Los Angeles, CA, February 25–28, 2007.
Maignan, S., Guilloteau, J.P., Zhou-Liu, Q., et al., 1998. Crystal structures of the catalytic domain of HIV-1 integrase free and complexed with its metal cofactor: high level of similarity of the active site with other viral integrases. J. Mol. Biol. 282 (2), 359–368.
Masso, M., Vaisman, I.I., 2008. Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. Bioinformatics 24 (18), 2002–2009.
Masso, M., Vaisman, I.I., 2010. Accurate and efficient gp120 V3 loop structure based models for the determination of HIV-1 co-receptor usage. BMC Bioinformatics 11.
Masso, M., Vaisman, I.I., 2010. Structure-based machine learning models for computational mutagenesis. In: Rangwala, H., Karypis, G. (Eds.), Introduction to Protein Structure Prediction: Methods and Algorithms. John Wiley and Sons Inc., Hoboken, New Jersey, USA, pp. 403–430.
Masso, M., Vaisman, I.I., 2010. AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements. Protein Eng. Des. Sel. 23 (8), 683–687.
Masso, M., Vaisman, I.I., 2010. Knowledge-based computational mutagenesis for predicting the disease potential of human non-synonymous single nucleotide polymorphisms. J. Theor. Biol. 266 (4), 560–568.
Masso, M., Vaisman, I.I., 2013. Sequence and structure based models of HIV-1 protease and reverse transcriptase drug resistance. BMC Genomics 14 (Suppl. 4).
McColl, D.J., Chen, X., 2010. Strand transfer inhibitors of HIV-1 integrase: bringing IN a new era of antiretroviral therapy. Antiviral Res. 85 (1), 101–118.

D. McColl, S. Fransen, S. Gupta, et al., "Resistance and cross-resistance to first generation integrase inhibitors: insights from a Phase II study of elvitegravir (GS-9137)," presented at the XVI International HIV Drug Resistance Workshop: Basic Principles & Clinical Implications, Barbados, West Indies, June 12-16, 2007.

Mesplede, T., Quashie, P.K., Osman, N., et al., 2013. Viral fitness cost prevents HIV-1 from evading dolutegravir drug pressure. Retrovirology 10.

Petropoulos, C.J., Parkin, N.T., Limoli, K.L., et al., 2000. A novel phenotypic drug susceptibility assay for human immunodeficiency virus type 1. Antimicrob. Agents Chemother. 44 (4), 920–928.

Prosperi, M.C., De Luca, A., 2012. Computational models for prediction of response to antiretroviral therapies. AIDS Rev. 14 (2), 145–153.

Qhull. Available at: <http://www.qhull.org/>.

Shafer, R.W., Schapiro, J.M., 2008. HIV-1 drug resistance mutations: an updated framework for the second decade of HAART. AIDS Rev. 10 (2), 67–84.

Sippl, M.J., 1993. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. J. Comput. Aided Mol. Des. 7 (4), 473–501.

Stanford University HIV Drug Resistance Database. Available at: <http://hivdb.stanford.edu/cgi-bin/IN_Phenotype.cgi>. Accessed October 2013.

Van der Borght, K., Verheyen, A., Feyaerts, M., et al., 2013. Quantitative prediction of integrase inhibitor resistance from genotype through consensus linear regression modeling. Virol. J. 10 (8).

Vella, S., Schwartlander, B., Sow, S.P., et al., 2012. The history of antiretroviral therapy and of its implementation in resource-limited areas of the world. AIDS 26 (10), 1231–1241.

Wang, G., Dunbrack Jr., R.L., 2003. PISCES: a protein sequence culling server. Bioinformatics 19 (12), 1589–1591.

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, Second ed. Morgan Kaufmann Publishers, San Francisco, CA.

Xue, W., Jin, X., Ning, L., et al., 2013. Exploring the molecular mechanism of cross-resistance to HIV-1 integrase strand transfer inhibitors by molecular dynamics simulation and residue interaction network analysis. J. Chem. Inf. Model. 53 (1), 210–222.