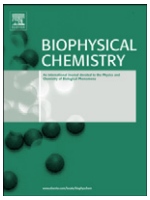
Biophysical Chemistry 153 (2011) 168–172

Contents lists available at ScienceDirect

Biophysical Chemistry

journal homepage: http://www.elsevier.com/locate/biophyschem



A combined sequence–structure approach for predicting resistance to the non-nucleoside HIV-1 reverse transcriptase inhibitor Nevirapine

Vadim L. Ravich, Majid Masso, Iosif I. Vaisman *

Laboratory for Structural Bioinformatics, Department of Bioinformatics and Computational Biology, George Mason University, 10900 University Blvd., MSN 5B3, Manassas, VA 20110, United States

ARTICLE INFO

Article history:
Received 22 September 2010
Received in revised form 5 November 2010
Accepted 12 November 2010
Available online 23 November 2010

Keywords:
Delaunay tessellation
Knowledge-based statistical potential
Computational mutagenesis
Machine learning
HIV-1 drug resistance
Prediction

ABSTRACT

The development of drug resistance to antiretroviral medications used to treat infection with HIV-1 is a major concern. Given the cost and time constraints associated with phenotypic resistance testing, computational approaches leading to accurate predictive models of resistance based on a patient's mutational patterns in the target protein would provide a welcome alternative. A combined sequence–structure computational mutagenesis procedure is used to generate attribute vectors for each of 222 mutational patterns of HIV-1 reverse transcriptase that were isolated and sequenced from patients. Phenotypic fold-levels of resistance to the non-nucleoside inhibitor Nevirapine are known for over 25% of these mutants, whose values are used to assign each assayed mutant to a drug susceptibility class, either sensitive or resistant. Support vector machine and random forest supervised learning algorithms applied to this subset respectively classify mutants based on drug susceptibility with 85% and 92% cross-validation accuracy. The trained models are used to predict susceptibility to Nevirapine for all remaining mutant isolates, and predictions are in agreement for 90% of the test cases.

1. Introduction

The HIV-1 reverse transcriptase (RT) is an important target enzyme for nearly all combination antiretroviral therapies that are currently available to treat patients [1]. In the earliest stages following infection of a host cell, RT is responsible for converting the RNA viral genome of HIV-1 into DNA for subsequent integration into the host genome. In addition to RT, the *pol* gene of HIV-1 encodes the protease and integrase enzymes, which are also crucial for viral replication and targets for pharmaceutical inhibitor drugs [2]. The functional RT enzyme is a heterodimer consisting of a p66 subunit that is enzymatically active and a p51 subunit that provides structural stability (Fig. 1A [3]). The larger chain contains both an N-terminal polymerase domain comprising 440 amino acid residues as well as a C-terminal RNase H domain that spans 120 residues [4]. The palm of the p66 subunit includes the polymerase active site, characterized by the catalytic aspartic triad formed by Asp110, Asp185, and Asp186 [5], where the latter two residues participate in a highly conserved YXDD motif across retroviral RTs [6,7].

Commercially available non-nucleoside reverse transcriptase inhibitor (NNRTI) drugs bind to a hydrophobic region located in the palm subdomain of the p66 subunit, approximately 10 Å away from the polymerase active site [8]. In particular, the drug Nevirapine (NVP) makes a total of 38 atomic contacts with residues in the palm and thumb subdomains. A beta-sheet within the palm is shifted as a result of NNRTI binding, which alters the geometry of the active site and deactivates polymerase activity [9]. The majority of mutations in RT associated with NNRTI resistance occur at residue positions making direct contact with the particular drug, including Leu100, Lys103, Val106, Val108, Tyr181, Tyr188, Gly190, Pro225, Met230, and Pro236 [10,11]. Amino acid replacements at these positions interfere with NNRTI binding by eliminating atomic contacts as well as by altering the size and shape of the hydrophobic region. Analysis of crystallographic structures has revealed that drug resistance mutations do not substantially change protein conformation but introduce local geometric variations around mutation sites, inducing a change in local van der Waals forces and hydrogen bonding patterns [12].

Given the clinical imperative for prescribing to HIV-1 infected patients an effective cocktail of antiretroviral medications to which they are susceptible, genotypic and phenotypic assays are now available to assess the degree to which RT enzymes harboring single or multiple amino acid substitutions are susceptible to inhibitor drugs [13]. Genotyping consists of sequencing patient RT isolates in order to determine if there are mutations present that are already known to be associated with resistance, while phenotyping involves directly measuring and comparing the susceptibility of an RT mutant to an inhibitor relative to a drug-sensitive RT control. Since phenotypic assays are expensive and can take up to two weeks to complete, reports detailing computational techniques for rapidly predicting phenotype from genotype have started to appear in the literature [14–22]. Additionally,

* Corresponding author. Tel.: +1 703 993 8431; fax: +1 703 993 8401.
E-mail address: ivaisman@gmu.edu (I.I. Vaisman).

0301-4622/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2010.11.004

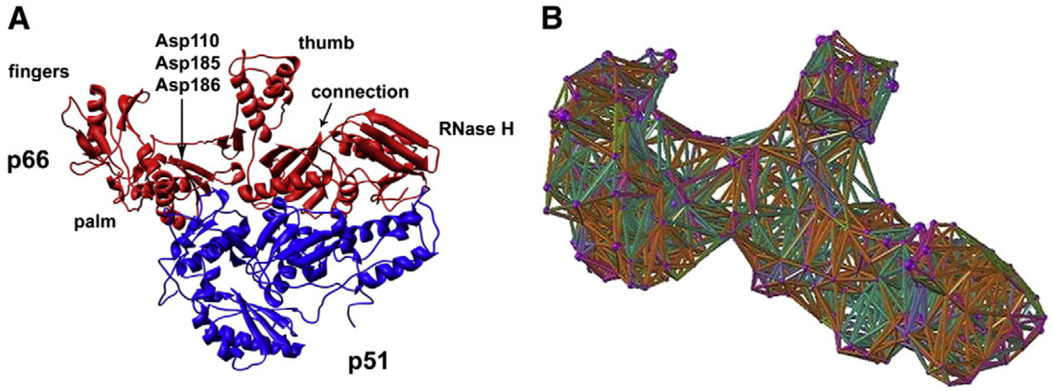


Fig. 1. (A) Ribbon diagram of the HIV-1 RT heterodimer (PDB accession code 1rtj) identifying pertinent structural and functional elements, and (B) Delaunay tessellation of the p66 subunit of RT subject to an edge-length cutoff of 12 Å.

structure-based approaches to predicting resistance have also been described [12,17,22–28].

Our work focuses on a combined sequence–structure approach to predicting NVP drug resistance, based on a novel computational mutagenesis methodology that utilizes a Delaunay tessellation-derived four-body statistical potential. For each mutational pattern in the p66 subunit of RT, the approach yields an attribute vector whose components quantify environmental perturbations at all RT positions due to the residue substitutions. Attribute vectors are generated for 222 distinct mutational patterns in the p66 subunit of RT isolated and sequenced from patients, where each mutant consists of up to eight amino acid substitutions. For 59 of these mutants, the fold-levels of resistance to NVP are already known from experimental phenotypic assays and are used to classify each mutant as either sensitive (S) or resistant (R) to the inhibitor. This subset of mutants, represented by their attribute vectors, is used to train predictive models based on the support vector machine (SVM) [29] and random forest (RF) [30] supervised classification algorithms. We undertake a detailed evaluation of model performance, which includes investigating the relative contributions of sequence and structure to the overall model accuracy.

2. Materials and methods

2.1. Delaunay tessellation and the four-body statistical potential

Given the 3-dimensional (3D) C_α atomic coordinates of all constituent amino acids in a protein structure, the Delaunay tessellation of this point-set is a computational geometry construct that yields a space-filling tiling of non-overlapping irregular tetrahedral simplices (Fig. 1B). The points are utilized as vertices for the tetrahedra, and since the points represent their respective amino acids, each tetrahedron objectively identifies a quadruplet of nearest-neighbor residues. Adjacent tetrahedral simplices may share one point, one edge (i.e., two points), or one triangular facet (i.e., three points); hence, each point is generally shared as a vertex simultaneously by numerous tetrahedral simplices in a tessellation. The four-body statistical potential, based on the tessellations of over 1200 structurally diverse protein structures with low sequence similarity selected from the Protein Data Bank (PDB) [31], empirically quantifies the energy of interaction (a log-likelihood score) for each of the 8855 permutation-free amino acid quadruplet subsets (obtained by sampling with replacement from the standard 20-letter protein alphabet) via an application of the Boltzmann principle from statistical mechanics [32]. The Quickhull algorithm [33] is used to tessellate the protein structures, and we prepared an ad-hoc suite of Java and Perl programs for the subsequent data analyses.

2.2. Computational mutagenesis and mutant residual profile vectors

In the Delaunay tessellation of the native p66 subunit of HIV-1 RT (Fig. 1, PDB accession code 1rtj, chain A [34]), each tetrahedral simplex is assigned with a score by identifying the four amino acids at the vertices and referring to the four-body statistical potential for the log-likelihood score of that residue quadruplet. For each position in the protein structure, a *residue environment score* is calculated by adding up the scores of all the tetrahedral simplices that share its C_α coordinate as a vertex, and a vector consisting of the collective environment scores for all p66 subunit positions is referred to as a *potential profile* [35]. A potential profile for any p66 mutant due to residue substitutions is obtained by recalculating all residue environment scores after replacing amino acid identity labels at the appropriate C_α coordinates in the tessellation to reflect the mutation (i.e., after threading a new RT sequence onto the C_α coordinates of the native structure). We define the *residual profile* of any p66 mutant as the difference between the mutant and native potential profiles, and we refer to the residual profile component values as *environmental change (EC) scores* for the corresponding positions in the mutant protein. These residual profiles are precisely the attribute vectors with which we represent the collection of RT mutants. The components with nonzero EC scores in a mutant residual profile identify the mutated residue positions as well as their topological nearest neighbors based on the Delaunay tessellation (i.e., those with which they participate as vertices to form tetrahedra), and the values of these nonzero EC scores are a consequence of the types of amino acid replacements at the mutated positions and their compatibilities with neighboring residues. Hence, sequence and structure information is encoded in mutant residual profiles.

2.3. Experimental phenotypic data

The Stanford HIV Drug Resistance Database (http://hivdb.stanford.edu/) provides a tabulation of 222 distinct mutational patterns for the p66 subunit of HIV-1 RT that were isolated and sequenced multiple times from over 4000 patients enrolled in large-scale clinical trials [36]. Also reported in the table for a subset of 59 of these mutants are the results of phenotypic tests, using the Monogram Biosciences PhenoSense assay, that provide measurements of susceptibility to the inhibitor NVP [37]. Many of these mutational patterns had been isolated from multiple patients, and these mutants were typically assayed numerous times. Results are reported in the table as a median fold change in susceptibility to NVP, defined as the ratio of the 50% inhibitory concentration (IC_{50}) for the mutant relative to that for a drug-sensitive reference HIV-1 RT control. The repeated fold change measurements for each of the mutants display a small mean absolute deviation from the median fold change [36]. Consistent with the manufacturer of the assay, we label mutants with mean fold change less than a threshold value of 4.5 as sensitive (S), and the remaining mutants are labeled as resistant (R) [37,38]. For the subset of assayed mutants that are NVP-resistant, the most commonly observed substitutions, at positions mentioned in the Introduction known to interact with the inhibitor, include the following: L100I, K103N, V106A, V108I, Y181C, Y188L, G190A, P225H, M230L, and P236L. More generally, among the remaining non-assayed mutants with currently unknown NVP susceptibility, we observe a similar result with respect to the most frequently occurring substitutions, with the exceptions of V106I and P236S at these respective positions.

2.4. Supervised classification and performance measures

The WEKA suite of machine learning tools [39] is used for implementing the support vector machine (SVM) [29] and random forest (RF) [30] supervised classification algorithms. These two state-of-the-art classifiers have been applied successfully and have displayed superior performance in various studies [40–43], and here we select the following parameters with these methods. The SVM classifier for this work uses a polynomial kernel to nonlinearly map the RT mutant residual profiles into a higher dimensional feature space, where a hyperplane is constructed that yields a maximal margin of separation between the R/S mutant classes and corresponds to a nonlinear decision boundary in the original space. The RF algorithm parameters include a forest of ten classification trees with ten attributes randomly selected for splitting at each node, and predictions are based on a majority vote. Classifier performance is evaluated via k-fold cross-validation, a well-documented statistical testing procedure whereby the dataset mutants are first randomly grouped into k equally sized subsets. The subsets are stratified for k greater than 1, meaning that the ratio of the class sizes in the original dataset is maintained in each of the subsets. Next, one subset is held-out for prediction with a model trained by a dataset consisting of the combined elements of the other k-1 subsets, the process is repeated k times so that each subset is held-out exactly once for prediction, and the procedure concludes with each of the original dataset elements having associated with it a single class prediction that can be compared with its actual class. Due to the bias-variance tradeoff, classification performance is evaluated based on both stratified tenfold cross-validation (10-fold CV, higher bias in the expected prediction error but lower variance) as well as leave-one-out CV (LOOCV, approximately unbiased but higher variance) approaches.

Given that TP (TN) represents the total number of correctly predicted R (S) mutants, and FN (FP) refers to the total number of respectively misclassified mutants, the overall accuracy (Q), the balanced error rate (BER), and the Matthew's correlation coefficient (MCC) are calculated as follows:

$$Q = \frac{TP + TN}{TP + FN + FP + TN}$$

$$BER = \frac{1}{2} \times \left(\frac{FN}{FN + TP} + \frac{FP}{FP + TN}\right)$$

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

The latter two values are used in place of reporting measures such as sensitivity, specificity, and precision for each of the classes individually, and they are especially useful for unbalanced class distributions. For this reason, we also report the area (AUC) under the receiver operating characteristic (ROC) curve, a plot of the true positive rate (i.e., sensitivity) versus the false positive rate (i.e., 1 − specificity) in the unit square. An AUC value near 0.5 is suggestive of a model that does not perform better than random guessing, while a value of 1.0 indicates a perfect classifier. Finally, a chi-square test can be applied to assess the statistical significance of computed MCC values, where the test statistic is given by $\chi^2 = N \times MCC^2$ (N = dataset size) with one degree of freedom [44].

3. Results and discussion

3.1. Evaluation of SVM and RF model performances

Given the highly skewed distribution of classes in the HIV-1 RT training set (13 S and 46 R mutants, Supplementary material Table S1), cost sensitive learning is applied to both the SVM and RF algorithms in order to optimize the performance for both classes. Specifically, the cost of misclassifying an S mutant is weighted three times as heavily as that of an R mutant. Additionally, we apply a bagging (bootstrap aggregating) procedure with both methods to reduce variance by generating ten bootstrap datasets of mutants from the original training set and aggregating results via majority vote. Each bootstrap dataset is the same size as the original training dataset and is generated by randomly selecting with replacement elements from the original set.

Table 1 summarizes the classification performance with the incorporation of the techniques described earlier to both learning schemes. Overall prediction accuracy (Q) with the 10-fold CV and LOOCV testing methods is identically 0.85 based on SVM and varies from 0.90 to 0.92, respectively, with RF. The RF model outperforms that of SVM in both cases primarily due to an increase in the sensitivity of R class mutant predictions. Sensitivity of S class mutant predictions remains at 0.85 over both learning methods and both testing approaches, while sensitivity of R class mutant predictions, which is also 0.85 using SVM and both testing approaches, increases to 0.91 and 0.94 using RF with 10-fold CV and LOOCV, respectively. These results are reflected in Table 1 through BER values that are lower with RF classification relative to SVM. All MCC values are statistically different from zero ($p < 0.0001$), which indicates that RF and SVM predictions are significantly more correlated with the data compared to random guessing. For the purpose of comparison to an established yet orthogonal method, Beerenwinkel et al. [20] trained a decision tree classifier for predicting NVP resistance using a sequence-based approach and a significantly larger dataset (N = 457), and they reported LOOCV performance measures of Q = 0.90 and S/R class sensitivity values of 0.97/0.82. Lastly, Fig. 2 depicts the ROC curves (labeled "Original Dataset") associated with SVM and RF 10-fold CV, respectively, with AUC values of 0.90 and 0.93.

3.2. Control datasets

The first control set consists of performing a random shuffling of the 13 S and 46 R class labels among the 59 HIV-1 RT mutants in the original training set. Evaluation of SVM and RF performances on this control by using the 10-fold CV approach described in the previous section leads to the ROC curves shown in Fig. 2 (labeled "Shuffled Classes") with AUC values of 0.48 and 0.54, respectively. These AUC values suggest that SVM and RF models trained with this control are not expected to perform any better than random guessing, highlighting the significance of results

Table 1
SVM and RF performance measures.

Method	Q	BER	MCC	AUC
10-fold CV				
SVM	0.85	0.15	0.62	0.90
RF	0.90	0.12	0.72	0.93
LOOCV				
SVM	0.85	0.15	0.62	0.90
RF	0.92	0.11	0.76	0.93

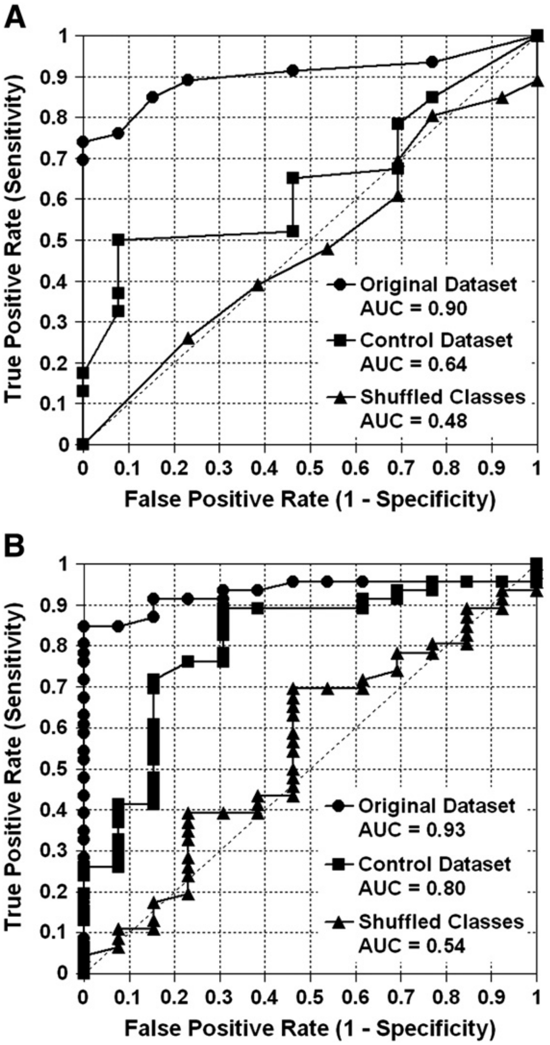


Fig. 2. ROC curves associated with (A) SVM and (B) RF supervised classifications.

obtained with the original training set based on the strength of the signals encoded in the mutant residual profiles and the degree of signal disparity between the S and R classes.

Next, we consider a control for assessing the relative contributions of sequence and structure information encoded in the HIV-1 RT mutant residual profile vectors to overall performance. In particular, the residual profile components with nonzero EC scores identify the locations of both the mutated residue positions as well as their nearest neighbors based on the Delaunay tessellation of the p66 subunit structure of RT. On the other hand, the actual values of these nonzero EC scores are a consequence of the specific types of amino acid replacements at the mutated positions in the primary sequence of the protein; alternative substitutions at the identical set of positions lead to a different set nonzero EC scores, but their component locations within the residual profile vector are unaltered. Hence, we generate a control set that measures the structural contribution to performance by randomly altering the values of all nonzero EC scores in the 59 residual profiles that constitute the original training set. Again, application of the 10-fold CV approach to this control leads to the SVM and RF ROC curves shown in Fig. 2 ("Control Dataset"), with AUC values of 0.64 and 0.80, respectively. These results suggest that structure-based signals contribute 38% $(100\times(\mathrm{AUC}_{\mathrm{control}}-\mathrm{AUC}_{\mathrm{shuffled}})/(\mathrm{AUC}_{\mathrm{original}}-\mathrm{AUC}_{\mathrm{shuffled}})=100\times(0.64-0.48)/(0.90-0.48))$ to SVM performance and 67% $(100\times(0.80-0.54)/(0.93-0.54))$ to RF performance. In the case of RF, our result here is relatively similar to that obtained for a similar analysis of training sets for HIV-1 protease mutants with experimentally known fold-levels of resistance to seven inhibitors (74% overall mean structural contribution) [43].

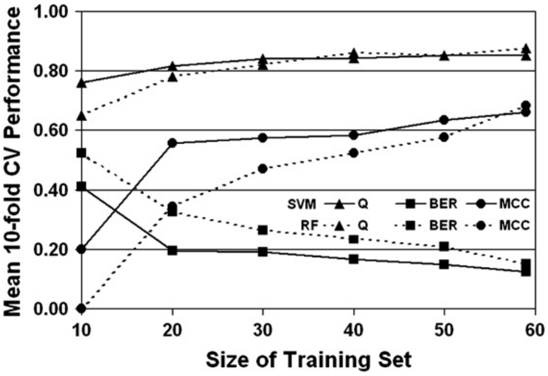


Fig. 3. Learning curves associated with SVM and RF supervised classifications.

3.3. Learning curves

Starting with a stratified random sampling of 10 mutants chosen from the original training set, 10-fold CV is applied ten times, from which we calculate mean performance values and respective standard deviations. At each subsequent iteration, we return to the original training set of 59 mutants and increment the size of the sampled training set by 10 mutants, and mean performance and standard deviation are again computed based on ten runs of 10-fold CV. Fig. 3 shows the resulting SVM and RF learning curves, plots of mean performance (Q, BER, and MCC) versus training set size. The plots reflect a substantial improvement in mean performance values as the size of the training set increases, accompanied by a decrease in variability (not shown) due to larger proportions of mutants being selected from the original training set. Based upon the plots presented in Fig. 3, the future inclusion of additional mutants to the current training set may further benefit model performance.

3.4. SVM and RF model predictions

Lastly, we utilize trained SVM and RF models for predicting the S/R class memberships of the 163 uncharacterized HIV-1 RT patient mutational patterns, by supplying each of the models with an independent test set consisting of residual profiles for these mutants. With 90% of these mutants (147 out of 163), both the SVM and RF models yield the same class prediction (Supplementary material Table S2). One straightforward reason for not obtaining a prediction match for 16 of the mutants may be due to the fact that the RF algorithm demonstrated a slightly better performance than SVM. A more subtle explanation could be that for these mutants, future phenotypic testing may reveal that their fold-levels of resistance are near the threshold value of 4.5 that is used for separating S from R mutants in the training set. Indeed, this collection of RT mutants will provide a reliable and independent real-world validation dataset once phenotyping results are performed and made available to compare with these predictions.

4. Conclusions

A computational mutagenesis methodology incorporating both sequence and structure yields residual profile attribute vectors for representing HIV-1 RT mutants isolated and sequenced from patients. This proof-of-principle report illustrates the relevance of signals encoded in these mutant residual profiles by evaluating the performance of SVM and RF classifiers trained using a set of RT mutants with known sensitivity (S) or resistance (R) to the inhibitor NVP. Cross-validation accuracies (85–92%), balanced error rates (0.11–0.15), Matthew's correlation coefficients (0.62–0.76), and areas under the ROC curve

(0.90–0.93) are all suggestive of reliable predictive models. Two control datasets, separately generated from the original training set by random permutations of the S and R class labels as well as random changes to nonzero residual profile components, respectively reflect classification significance of the trained models and gauge relative contribution of structure information to overall model performance. Learning curves reveal that training set size significantly influences performance and suggest that inclusion of additional RT mutants into the training set, based on future availability of more phenotyping data, may further improve classification capabilities of the models. Finally, predictions of the sensitivity or resistance to NVP for all the remaining uncharacterized mutants of RT obtained from patient samples are obtained from the trained SVM and RF models and display 90% concordance.

Acknowledgements

We thank T. Taylor for help with the statistical potential, A. Carr for creating the HIV-1 RT tessellation image, and M. Barenboim and J. Ganju for fruitful discussions. One of us (V.R.) is grateful to colleagues at NCI for helpful discussions and work-schedule related accommodations while conducting his graduate research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.bpc.2010.11.004.

References

[1] G. Barbaro, A. Scozzafava, A. Mastrolorenzo, C.T. Supuran, Highly active antiretroviral therapy: current state of the art, new agents and their pharmacological interactions useful for improving therapeutic outcome, Curr. Pharm. Des. 11 (2005) 1805–1843.

[2] J.D. Reeves, A.J. Piefer, Emerging drug targets for antiretroviral therapy, Drugs 65 (2005) 1747–1766.

[3] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T. E. Ferrin, UCSF Chimera — a visualization system for exploratory research and analysis, J. Comput. Chem. 25 (2004) 1605–1612.

[4] J.A. Wrobel, S.F. Chao, M.J. Conrad, J.D. Merker, R. Swanstrom, G.J. Pielak, C.A. Hutchison III, A genetic approach for identifying critical residues in the fingers and palm subdomains of HIV-1 reverse transcriptase, Proc. Natl Acad. Sci. USA 95 (1998) 638–645.

[5] J. Ding, K. Das, Y. Hsiou, S.G. Sarafianos, A.D. Clark Jr., A. Jacobo-Molina, C. Tantillo, S.H. Hughes, E. Arnold, Structure and functional implications of the polymerase active site region in a complex of HIV-1 RT with a double-stranded DNA template-primer and an antibody Fab fragment at 2.8 Å resolution, J. Mol. Biol. 284 (1998) 1095–1111.

[6] J. Lingner, T.R. Hughes, A. Shevchenko, M. Mann, V. Lundblad, T.R. Cech, Reverse transcriptase motifs in the catalytic subunit of telomerase, Science 276 (1997) 561–567.

[7] T.M. Nakamura, G.B. Morin, K.B. Chapman, S.L. Weinrich, W.H. Andrews, J. Lingner, C.B. Harley, T.R. Cech, Telomerase catalytic subunit homologs from fission yeast and human, Science 277 (1997) 955–959.

[8] J. Ren, R. Esnouf, E. Garman, D. Somers, C. Ross, I. Kirby, J. Keeling, G. Darby, Y. Jones, D. Stuart, et al., High resolution structures of HIV-1 RT from four RT-inhibitor complexes, Nat. Struct. Biol. 2 (1995) 293–302.

[9] D.W. Rodgers, S.J. Gamblin, B.A. Harris, S. Ray, J.S. Culp, B. Hellmig, D.J. Woolf, C. Debouck, S.C. Harrison, The structure of unliganded reverse transcriptase from the human immunodeficiency virus type 1, Proc. Natl Acad. Sci. USA 92 (1995) 1222–1226.

[10] Y. Hsiou, J. Ding, K. Das, A.D. Clark Jr., P.L. Boyer, P. Lewi, P.A. Janssen, J.P. Kleim, M. Rosner, S.H. Hughes, E. Arnold, The Lys103Asn mutation of HIV-1 RT: a novel mechanism of drug resistance, J. Mol. Biol. 309 (2001) 437–445.

[11] D.P. Wang, R.C. Rizzo, J. Tirado-Rives, W.L. Jorgensen, Antiviral drug design: computational analyses of the effects of the L100I mutation for HIV-RT on the binding of NNRTIs, Bioorg. Med. Chem. Lett. 11 (2001) 2799–2802.

[12] Y.Z. Chen, X.L. Gu, Z.W. Cao, Can an optimization/scoring procedure in ligand-protein docking be employed to probe drug-resistant mutations in proteins? J. Mol. Graph. Model. 19 (2001) 560–570.

[13] J. Sebastian, H. Faruki, Update on HIV resistance and resistance testing, Med. Res. Rev. 24 (2004) 115–125.

[14] M. Zazzi, L. Romano, G. Venturi, R.W. Shafer, C. Reid, F. Dal Bello, C. Parolin, G. Palu, P.E. Valensin, Comparative evaluation of three computerized algorithms for prediction of antiretroviral susceptibility from HIV type 1 genotype, J. Antimicrob. Chemother. 53 (2004) 356–360.

[15] A.G. DiRienzo, V. DeGruttola, B. Larder, K. Hertogs, Non-parametric methods to predict HIV drug susceptibility phenotype from genotype, Stat. Med. 22 (2003) 2785–2798.

[16] B. Schmidt, H. Walter, B. Moschik, C. Paatz, K. van Vaerenbergh, A.M. Vandamme, M. Schmitt, T. Harrer, K. Uberla, K. Korn, Simple algorithm derived from a geno-/phenotypic database to predict HIV-1 protease inhibitor resistance, AIDS 14 (2000) 1731–1738.

[17] Z.W. Cao, L.Y. Han, C.J. Zheng, Z.L. Ji, X. Chen, H.H. Lin, Y.Z. Chen, Computer prediction of drug resistance mutations in proteins, Drug Discov. Today 10 (2005) 521–529.

[18] E. Puchhammer-Stockl, C. Steininger, E. Geringer, F.X. Heinz, Comparison of virtual phenotype and HIV-SEQ program (Stanford) interpretation for predicting drug resistance of HIV strains, HIV Med. 3 (2002) 200–206.

[19] D. Wang, B. Larder, Enhanced prediction of lopinavir resistance from genotype by use of artificial neural networks, J. Infect. Dis. 188 (2003) 653–660.

[20] N. Beerenwinkel, B. Schmidt, H. Walter, R. Kaiser, T. Lengauer, D. Hoffmann, K. Korn, J. Selbig, Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype, Proc. Natl Acad. Sci. USA 99 (2002) 8271–8276.

[21] N. Beerenwinkel, M. Daumer, M. Oette, K. Korn, D. Hoffmann, R. Kaiser, T. Lengauer, J. Selbig, H. Walter, Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes, Nucleic Acids Res. 31 (2003) 3850–3855.

[22] S. Draghici, R.B. Potter, Predicting HIV drug resistance with neural networks, Bioinformatics 19 (2003) 98–107.

[23] I.T. Weber, R.W. Harrison, Molecular mechanics analysis of drug-resistant mutants of HIV protease, Protein Eng. 12 (1999) 469–474.

[24] W. Wang, P.A. Kollman, Computational study of protein specificity: the molecular basis of HIV-1 protease drug resistance, Proc. Natl Acad. Sci. USA 98 (2001) 14937–14942.

[25] M.D. Shenderovich, R.M. Kagan, P.N. Heseltine, K. Ramnarayan, Structure-based phenotyping predicts HIV-1 protease inhibitor resistance, Protein Sci. 12 (2003) 1706–1718.

[26] A.C. Nair, I. Bonin, A. Tossi, W.J. Wels, S. Miertus, Computational studies of the resistance patterns of mutant HIV-1 aspartic proteases towards ABT-538 (ritonavir) and design of new derivatives, J. Mol. Graph. Model. 21 (2002) 171–179.

[27] D. Stoffler, M.F. Sanner, G.M. Morris, A.J. Olson, D.S. Goodsell, Evolutionary analysis of HIV-1 protease inhibitors: methods for design of inhibitors that evade resistance, Proteins 48 (2002) 63–74.

[28] X. Chen, I.T. Weber, R.W. Harrison, Molecular dynamics simulations of 14 HIV protease mutants in complexes with indinavir, J. Mol. Model (Online) 10 (2004) 373–381.

[29] J. Platt, Fast training of support vector machines using sequential minimal optimization, in: B. Scholkopf, C. Burges, A. Smola (Eds.), Advances in Kernel Methods — Support Vector Learning, MIT Press, Cambridge, MA, 1998, pp. 185–208.

[30] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[31] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.

[32] R.K. Singh, A. Tropsha, I.I. Vaisman, Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues, J. Comput. Biol. 3 (1996) 213–221.

[33] C.B. Barber, D.P. Dobkin, H.T. Huhdanpaa, The quickhull algorithm for convex hulls, ACM Trans. Math. Softw. 22 (1996) 469–483.

[34] R. Esnouf, J. Ren, C. Ross, Y. Jones, D. Stammers, D. Stuart, Mechanism of inhibition of HIV-1 reverse transcriptase by non-nucleoside inhibitors, Nat. Struct. Biol. 2 (1995) 303–308.

[35] M. Masso, I.I. Vaisman, Comprehensive mutagenesis of HIV-1 protease: a computational geometry approach, Biochem. Biophys. Res. Commun. 305 (2003) 322–326.

[36] S.Y. Rhee, T. Liu, J. Ravela, M.J. Gonzales, R.W. Shafer, Distribution of human immunodeficiency virus type 1 protease and reverse transcriptase mutation patterns in 4, 183 persons undergoing genotypic resistance testing, Antimicrob. Agents Chemother. 48 (2004) 3122–3126.

[37] C.J. Petropoulos, N.T. Parkin, K.L. Limoli, Y.S. Lie, T. Wrin, W. Huang, H. Tian, D. Smith, G.A. Winslow, D.J. Capon, J.M. Whitcomb, A novel phenotypic drug susceptibility assay for human immunodeficiency virus type 1, Antimicrob. Agents Chemother. 44 (2000) 920–928.

[38] N.T. Parkin, N.S. Hellmann, J.M. Whitcomb, L. Kiss, C. Chappey, C.J. Petropoulos, Natural variation of drug susceptibility in wild-type human immunodeficiency virus type 1, Antimicrob. Agents Chemother. 48 (2004) 437–443.

[39] E. Frank, M. Hall, L. Trigg, G. Holmes, I.H. Witten, Data mining in bioinformatics using Weka, Bioinformatics 20 (2004) 2479–2481.

[40] J. Vingerhoets, L. Tambuyzer, H. Azijn, A. Hoogstoel, S. Nijs, M. Peeters, M.P. de Bethune, G. De Smedt, B. Woodfall, G. Picchio, Resistance profile of etravirine: combined analysis of baseline genotypic and phenotypic data from the randomized, controlled Phase III clinical studies, AIDS 24 (2010) 503–514.

[41] M.C. Prosperi, A. Altmann, M. Rosen-Zvi, E. Aharoni, G. Borgulya, F. Bazso, A. Sonnerborg, E. Schulter, D. Struck, G. Ulivi, A.M. Vandamme, J. Vercauteren, M. Zazzi, Investigation of expert rule bases, logistic regression, and non-linear machine learning techniques for predicting response to antiretroviral treatment, Antivir. Ther. 14 (2009) 433–442.

[42] T. Hou, W. Zhang, J. Wang, W. Wang, Predicting drug resistance of the HIV-1 protease using molecular interaction energy components, Proteins 74 (2009) 837–846.

[43] M. Masso, I.I. Vaisman, A novel sequence–structure approach for accurate prediction of resistance to HIV-1 protease inhibitors, Proc. IEEE Bioinform. Bioeng. 2 (2007) 952–958.

[44] P. Baldi, S. Brunak, Y. Chauvin, C.A. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (2000) 412–424.