



Computer prediction of drug resistance mutations in proteins

Zhi Wei Cao, Lian Yi Han, Chan Juan Zheng, Zhi Lang Ji, Xin Chen, Hong Huang Lin and Yu Zong Chen

Drug resistance is of increasing concern in the treatment of infectious diseases and cancer. Mutation in drug-interacting disease proteins is one of the primary causes for resistance particularly against anti-infectious drugs. Prediction of resistance mutations in these proteins is valuable both for the molecular dissection of drug resistance mechanisms and for predicting features that guide the design of new agents to counter resistant strains. Several protein structure- and sequence-based computer methods have been explored for mechanistic study and prediction of resistance mutations. These methods and their usefulness are reviewed here.

▶ Drug resistance is a global public health problem particularly for the treatment of infectious diseases [1,2] and cancer [3]. It has become increasingly possible for cross country transmission of drug-resistant organisms. There is an urgent need to develop resistance-evading drugs. Several mechanisms are responsible for drug resistance [4–8]. For infectious diseases, resistance is primarily mediated by mutations in the genes of infectious organisms that alter a drug's interaction with its corresponding target protein [4,5].

Considerable effort has been directed at the use of computational methods for studying the molecular mechanism of mutation-induced drug resistance and for developing predictive tools. The availability of the three-dimensional structure of drug targets involved in disease enables the use of molecular modeling and other structure-based approaches for evaluation of structural features, molecular interactions, solvation and dynamical properties of drug-protein binding and their correlation to resistance mutations [9–13]. Structure-derived binding energies [9,14–16] and binding-site volume-based fitness models [17] have also been used for facilitating the prediction of resistance mutations.

Structural information is available for a relatively small percentage of proteins. Thus methods for predicting resistance mutations directly from sequences are highly useful and are being developed. Interpretation programs have emerged for identifying and estimating the level of resistance [18,19]. Statistical learning methods such as neural networks [20,21], support vector machines (SVM) [22] and decision tree [23] have also shown promising potential for predicting resistance mutations.

Molecular modeling of drug resistance mutations

Structural analysis of proteins that contain resistance mutations indicated that mutations at drug-binding sites usually alter the tight packing between the binding drug and its receptor without substantial change in overall conformation [4,24]. Comparison of mutant and wild-type drug-receptor structures showed that, in the majority of cases, the only apparent change is in the pattern of local contact and hydrogen bonding at a mutation site [24,25]. Hydrophobic effects have been found to be important in several cases [10], but variation in local packing

Zhi Wei Cao
Lian Yi Han
Chan Juan Zheng
Xin Chen
Hong Huang Lin
Yu Zong Chen
 Bioinformatics and Drug Design Group,
 Department of Computational Science,
 National University of Singapore,
 BLK SOC 1,
 level 7,
 Singapore 117543
 *e-mail: cscyz@nus.edu.sg
Zhi Lang Ji
 Department of Biology,
 School of Life Sciences,
 Xiamen University,
 Xiamen 361000,
 Fujian Province,
 PR China

interactions appears to be a major factor for the reduced drug binding affinity leading to resistance [4]. Thus molecular mechanics, molecular dynamics simulation, and Monte Carlo simulation methods are expected to be useful for structural optimization and structure-based energetic analysis of these mutations.

Molecular mechanics methods use atom–atom interaction energies for structural optimization and energetic analysis of a ligand–protein complex [9,14,15]. Bonded interactions are modeled by bond stretch, angle bending, and torsion terms. Non-bonded interactions are modeled by van der Waals and electrostatic terms. Hydrogen bonds can either be modeled by a separate term [14] or they can be included in van der Waals and electrostatic terms [9,15]. Moreover, solvation and entropic effects may be considered by using a simple solvation free energy model [14,26] and side-chain entropy model [15] respectively.

Structural optimization involves the selection of low energy molecular structures. One approach is to search molecular conformations for identifying the structure of the lowest energy. The other is energy minimization such that molecular structure is varied towards the direction of lower energy until it reaches the local minimum energy configuration. In most cases, ligand-protein binding is determined by non-bonded, hydrogen bond, solvation and side-chain torsion interactions. Thus ligand-binding affinities are frequently estimated by using these terms [9,14,15].

Molecular dynamics simulation methods derive trajectories of atomic positions of molecular motions and dynamical fluctuations by solving Newton's equations governed by the same sets of bonded and non-bonded interactions used in molecular mechanics methods. Solvation effect is described by using either explicit water molecules or continuous medium models. Entropic effect is derived from statistical analysis of trajectories of atomic positions. These methods are useful for structural optimization, motional and energetic analysis, and free energy computation [10–13].

The free energy difference between wild-type and mutant systems, which is useful for indicating drug resistance mutations, can be derived by using Monte Carlo simulation [27] as well as molecular dynamics simulation [10–13]. Monte Carlo simulation methods generate a series of molecular structures that are randomly distributed in the molecular conformational space and conform to a certain distribution pattern governed by the laws of statistical mechanics. These randomly generated structures can then be used to derive free energy difference between the wild-type and mutant structure by using the free energy perturbation method [27].

The structure of a ligand-protein complex and its mutants often needs to be modeled from a template. Such a template is usually the structure of a different mutational variant of the same protein or its complex with a different ligand [9–16]. The modeled structures of both wild-type and

mutant HIV-1 protease–inhibitor complexes have been found to be consistent with the crystallographic structures, with root mean square differences ranging from 0.5Å to 1.2Å [9,14,15], which suggests that the quality of these modeled structures reaches the level useful for facilitating structure-based study and prediction of resistance mutations.

The ligand-protein binding energies computed from molecular mechanics and the free energies computed from molecular dynamics for several wild-type and mutant ligand–protein complexes showed significant correlation with the observed binding affinities [9–16,26], indicating that these energy functions are useful for facilitating energetic analysis and the prediction of resistance mutations.

Structure-based prediction of drug resistance mutations

Structure-base virtual screening has been widely used for designing new drugs [28,29]. To achieve high-speed screening of a large number of compounds, efficient computational procedures have been routinely applied to structural optimization and scoring of docked ligand-protein structures [30–34]. Some of these procedures are very similar to those used for the molecular study of drug resistance mutations [9,13–16], which raises the possibility of using virtual screening approaches to indicate possible drug resistance mutations [14]. While more accurate in modeling resistance mutations than those used in virtual screening studies [10–13], a full molecular dynamics procedure has yet to be employed in a general virtual screening process partly owing to its computationally intensive nature. Thus, procedures that either use molecular mechanics alone or molecular mechanics plus a small run of molecular dynamics for structural optimization has been the primary choice for structure-based prediction of resistance mutations in a virtual screening process [9,14].

Structural models and energy functions

Most models of mutant drug–protein complexes are based on a starting crystal structure of the wild-type protein complexed with the same drug [9,14,15]. Each mutation is introduced by stripping the amino acid down to the atom and replacing it by the side chain of the new amino acid. In some studies, the atom is also kept intact for mutations between amino acids R, K and Q, as these are relatively large amino acids and normally located at the protein surface [9]. The mutant structure is then optimized by conformation search for the local residues to release structural clash among them and the binding drug, which is conducted by variation of rotatable bonds of the drug and those of the side-chain of the amino acids in contact with the drug. This is followed by energy minimization to allow the mutant structure to find the local minimum energy conformation [9,15]. In some studies, molecular dynamics simulation is conducted to allow the mutant structure to reach a more appropriate local minimum energy conformation [14]. The procedure of this

modeling process and the subsequent prediction of resistance mutations from the derived mutant three-dimensional structure is illustrated in Figure 1.

Energy minimization is typically conducted for no less than a few hundred iterations, compared to the <300 steps of energy minimization used in the majority of virtual screening studies [34]. Molecular dynamics simulation is typically performed for picoseconds and at a temperature of 300K. The drug-protein interaction energy is computed by the following empirical potential energy function or its variations that includes non-bonded van der Waals and electrostatic interactions, atomic solvation free energy terms, side chain conformational entropy, and hydrogen bond energy (Equation 1)

$$V = \sum_{vdw} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right] + \sum_{electr} \left[\frac{q_i q_j}{\epsilon_r r_{ij}} + E(r_{ij}) \right] + \sum_{atoms\ i} \Delta\sigma_i A_i + \sum_{residue\ k} \alpha_k N_k + \sum_{H\ bonds} V_H$$

A_{ij} and B_{ij} are van der Waals parameters; ϵ_r is the dielectric constant, q_i and q_j are the partial charges of the i -th and j -th atoms, and r_{ij} is the distance between them; $E(r_{ij})$ is the electrostatic interaction correction term used in some studies [9], $\Delta\sigma_i$ is the atomic solvation parameter and is the solvent-accessible surface area of the i -th atom [35], N_k is the number of rotatable bonds and α_k the coefficient of the k -th residue [36], and V_H is the hydrogen-bond energy [14]. Various sets of potential energy parameters, which are called force fields, have been used for the prediction of resistance mutations, which includes AMBER [14], CHARMM [26], UFF [9], ECEPP/3 [15] and CVFF [16].

Apart from the molecular modeling approach, two other structure-based methods have also been used for predicting resistance mutations. One is the evolutionary simulation model that analyses resistance mutations that produces viable viruses using Michaelis-Menten kinetics derived from a simple measure of volume complementarity between a mutant and its binding drug [17]. The other is a neural network model that uses structural features generated from homology modeled mutant structures for classification of resistance mutations [21].

Prediction performance

Table 1 summarizes the results of structure-based prediction studies of resistance mutations in HIV-1 protease [9,13–16,37] and HIV-1 reverse transcriptase [14,27,38]. The performance of these studies is primarily measured by the correlation coefficient, the R value, between changes in the computed binding energy and those in the experimentally estimated binding affinities. Prediction accuracy for resistance and non-resistance mutations, the P value, has also been used in some studies [13,17,21]. The computed R values range from 0.31 to 0.90 [9,14–16]. The computed P values are in the range of 57%–86% [13] for the molecular modeling methods and 45%–70% for the other two structure-based methods [17,21].

Molecular modeling of HIV-1 protease consistently showed a good correlation with experimentally estimated binding affinities, indicating the usefulness of the molecular modeling approach for prediction of resistance mutations. The prediction accuracy for HIV-1 reverse transcriptase is substantially lower than those of HIV-1 protease. Significant conformational flexibility in this enzyme [39], which is inadequately modeled by molecular modeling methods, is likely a factor for the lower accuracy.

Several other factors have not been adequately described by molecular modeling methods. These include interactions with water, hydrophobic effect and electrostatic interactions involving aromatic rings, which likely lead to a certain degree of error in ligand-protein structural optimization and the scoring of resistance mutations. Proper modeling of these interactions and further refinement of currently

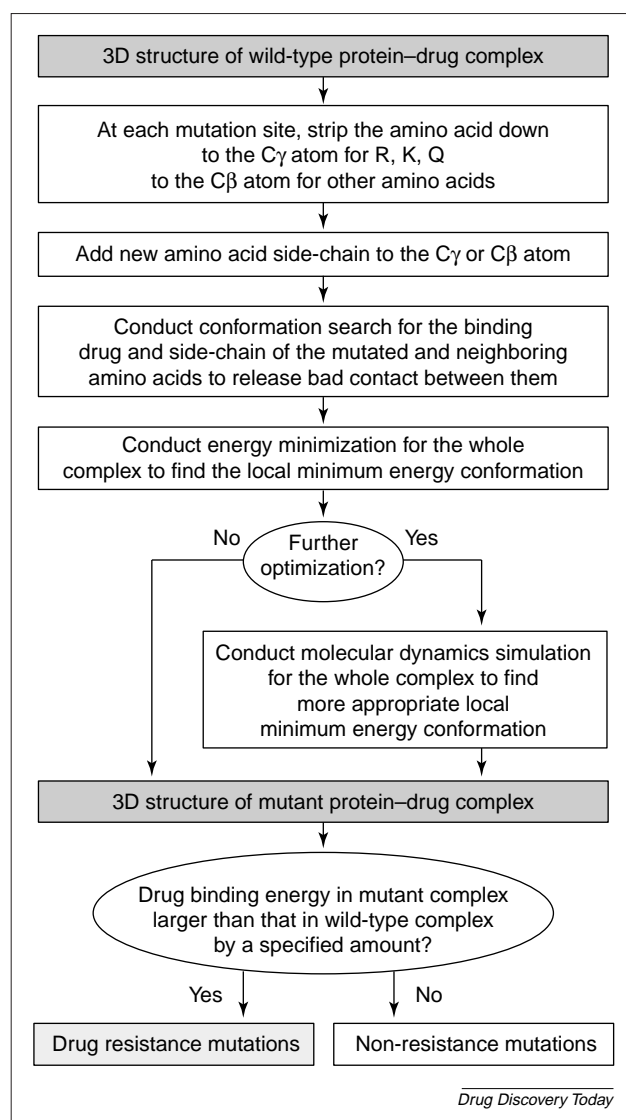


FIGURE 1

Schematic diagram illustrating the process of molecular modeling of the three-dimensional structure of a mutant protein-drug complex from that of the wild-type protein-drug complex, and the prediction of drug resistance mutations from the derived mutant three-dimensional structure.

TABLE 1
Structure-based prediction of drug resistance mutations

| Method and year of report | Protein | Drugs | Number of mutant-drug complexes | Reported prediction accuracy (R or P value) |
|---|-----------------------------|---|---------------------------------|--|
| Molecular modeling 1999 [9] | HIV-1 protease | Indinavir, Saquinavir | 17 | R = 0.68–0.76 |
| Molecular modeling 2001 [13] | HIV-1 protease | Amprenavir, Indinavir, Nelfinavir, Ritonavir, Saquinavir | 33 | P = 57%–86% |
| Molecular modeling with simple optimization scoring procedure 2001 [14] | HIV-1 protease | MK639, Saquinavir, SB203386, U89360E, VX478 | 22 | R = 0.64 |
| Molecular modeling 2002 [16] | HIV-1 protease | Ritonavir | 12 | R = 0.7 |
| Fitness evolution model 2002 [17] | HIV-1 protease | Indinavir, Nelfinavir, Ritonavir, Saquinavir | 11 | P = 45% (81% for resistance site prediction) |
| Molecular modeling 2003 [15] | HIV-1 protease | Amprenavir, Indinavir, Lopinavir, Nelfinavir, Ritonavir, Saquinavir | 750 | R = 0.87 for 360 complexes. P = 86% and 92% |
| Neural network model of homology modeled structures 2003 [21] | HIV-1 protease | Indinavir | 38 | P = 60%–70% |
| Molecular Dynamics simulation 2004 [37] | HIV-1 protease | Indinavir | 12 | R = 0.98 |
| Molecular modeling 1997 [38] | HIV-1 reverse transcriptase | 8-CL TIBO, α -APA | 8 | R = 0.80–0.97 |
| Monte Carlo Simulation 2000 [27] | HIV-1 reverse transcriptase | 8-CL TIBO | 2 | R = 0.78 |
| Molecular modeling with simple optimization scoring procedure 2001 [14] | HIV-1 reverse transcriptase | Nevirapine, TIBO R82913 | 13 | R = 0.31 |

Reported prediction accuracy is given by the R value (correlation coefficient between changes in the computed binding energy and those in the experimentally estimated binding affinities) or the P value (percentage of correctly predicted resistance and non-resistance mutations).

used optimization/scoring force field parameters may help improve the quality of modeled mutant structure and the accuracy of the scoring functions.

Sequence-based prediction of resistance mutations

Sequence-based methods predict resistance mutations by using the rules or classifiers derived from statistical analysis of the sequences of resistant and non-resistant samples. A straightforward approach is to identify known resistance mutations directly from the sequence, which is similar to the motif approach for protein functional analysis [40]. Genotype interpretation systems, such as Stanford HIVdb (HIVdb or HIV-SEQ), AntiRetroScan (ARS), and Visible Genetics/Bayer Diagnostics (VGI), have been developed for estimating the level of resistance mutations in HIV-1 protease and transcriptase based on this approach [18,19].

HIVdb scans a sequence for known resistance mutations and generates scores for each of the 16 anti-HIV drugs by computing the sum of the penalties retrieved from a table of resistance mutations of defined values of resistance level for each drug and its target protein (HIV-1 protease or reverse transcriptase). From the final score, one of the five susceptibility levels (susceptible, potential low-level resistance, low-level resistance, intermediate resistance, and high-level resistance) is assigned for each drug [18,19]. ARS is similar to HIVdb and uses the same penalty functions plus an additional set of 8–12 rules to determine the final scores [19]. VGI extracts known resistance mutations and compares them against a set of 85 rules for determining the level of resistance for each of the 16

anti-HIV drugs. The highest scored degree of resistance is used for assigning one of the three resistance levels (no evidence of resistance, possible resistance, and resistance) for each drug [19].

Statistical learning methods

Statistical learning methods such as neural networks (NN) [20,21], support vector machines (SVM) [22] and decision tree (DT) [23] have been used for developing classifiers and rules for predicting resistance mutations. In these approaches, protein sequence is represented by a feature vector x_i , with 20 binary bits (one for each type of amino acid) for each sequence position as its components. Samples of resistance and non-resistance mutations are used for training a statistical learning system to derive the classifiers and rules of resistance mutations [20–22]. These classifiers and rules can then be used to predict resistance mutations from the sequence of a mutant protein. The procedure for such a prediction is illustrated in Figure 2 where an example of SVM prediction of resistance mutations is provided.

NN trains a hidden-layer-containing network and uses its outcomes for pattern recognition and classification of the input feature vectors [20,21]. A classifier for NN is $y = g \sum_j w_{0j} h_j$, where w_{0j} is the output weight of a hidden node j to an output node, g is the output function, h_j is the value of a hidden layer node: $h_j = \sigma (\sum_i w_{ji} x_i + w_j)$, w_{ji} is the input weight from an input node i to a hidden node j , w_j is the threshold weight from an input node of value 1 to a hidden node j , and σ is a sigmoid function. Known

resistance and non-resistance samples are used for training a NN such that all the weights are determined, and the resulting classifier can be used for determining whether or not a new input sequence is resistant to a particular drug [20].

SVM constructs a hyperplane in a hyperspace for separating two groups of feature vectors with a maximum margin [22]. A classifier for a linear SVM is $y_i = (\mathbf{w} \cdot \mathbf{x}_i + b)$, where \mathbf{w} is a vector normal to the hyperplane, b is the perpendicular distance from the hyperplane to the origin with $\|\mathbf{w}\|$ as the Euclidean norm of \mathbf{w} , $y_i = +1$ for resistance and $y_i = -1$ for non-resistance mutations. The distribution functions of SVM can also be used for generating regression models [22]. Known resistance and non-resistance samples are used to train a SVM such that the hyperspace, distribution functions, and the parameters are determined. The resulting classifier can be used for predicting resistance and non-resistance mutations based on the value of y_i , and the trained regression model used for indicating the level of resistance from the regression model.

DT is an acyclic graph in which its interior vertices specify testing of a single attribute (sequence position) of a feature vector and its leaves indicate the classes of the attribute (resistance or non-resistance mutation) [23]. A

classifier of DT is constructed by recursively splitting the sample set, with each subset giving rise to one new vertex connected with an edge to its parent. This procedure continues until all samples at each leaf belong to the same class. Classification of a sample is achieved by running through the tree from the root to a leaf according to the values (amino acids) of the attributes of the protein sequence that appear on this path [23].

Prediction performance

Table 2 summarizes the results of sequence-based prediction studies of resistance mutations in HIV-1 protease and HIV-1 reverse transcriptase [20–23,41]. The performance of statistical learning method has been measured by the overall prediction accuracy for resistance and non-resistance mutations P , prediction accuracy for resistance mutations P_r , and prediction accuracy for non-resistance mutations P_n . The R value has been used in one study [22]. The number of samples in most of these studies is significantly higher than those in structure-based methods, reflecting the more extensive application range of sequence-based methods.

The computed P_r and P_n values are in the range of 58%–97% and 62%–97% respectively, and these are >90% in majority of the studies. The computed R values are in the range of 0.78–0.89 for HIV-1 protease, which are comparable to those obtained by structure-based methods, and 0.54–0.85 for HIV-1 reverse transcriptase, which are slightly better than those of structure-based methods. These results suggest that sequence-based methods are capable of equally accurate prediction of resistance mutations. As in the case of structure-based methods, the accuracies for HIV-1 reverse transcriptase appear to be substantially lower than those for HIV-1 protease. Conformational flexibility in HIV-1 reverse transcriptase [39], which is insufficiently represented in the feature vectors of statistical learning methods, is likely to be a major factor for the reduced accuracies.

The performance of sequence-based methods depends on the proper representation of mutant sequences as well as adequate training using a sufficiently diverse set of samples. While useful for distinguishing different mutations, it has been pointed out that the binary representation of amino acids used in most studies lacks biological basis and unnecessarily expands the input space such that it complicates the statistical learning task [42]. Thus, a more appropriate representation of amino acids is useful for further improving the performance of the sequence-based methods.

Prediction of resistance mutations by using simple rules

So far, both structure- and sequence-based methods for predicting resistance mutations have been primarily developed and tested on two proteins, HIV-1 protease and HIV-1 reverse transcriptase. This is because of the availability of

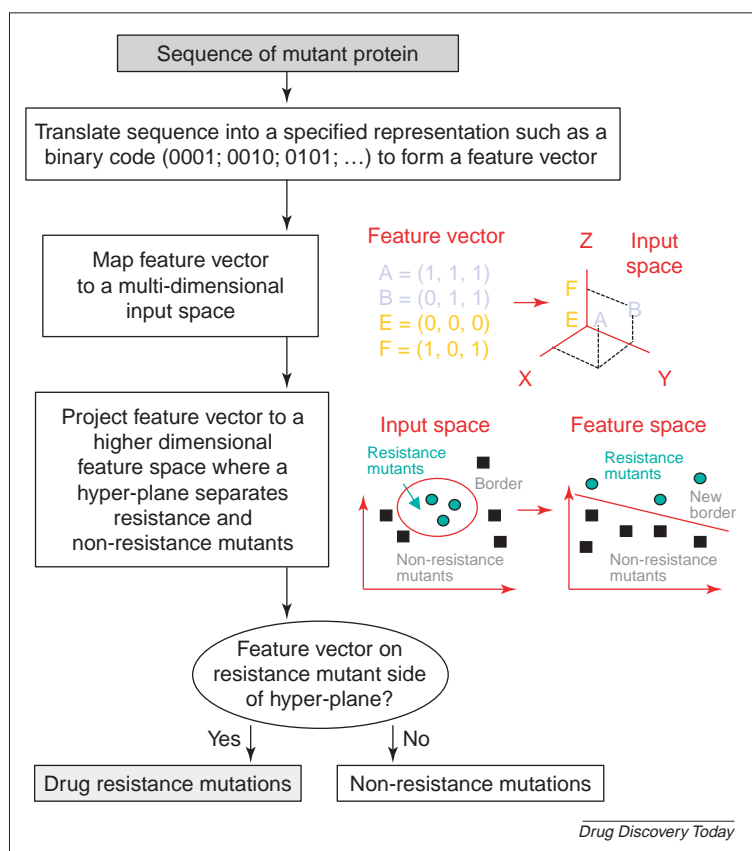


FIGURE 2

Schematic diagram illustrating the process of the prediction of resistance mutations from the sequence of a mutant by using a statistical learning method - support vector machines. A,B: feature vectors of non-resistance mutants; E,F: feature vectors of resistant mutants; black filled squares, non-resistant mutants; green circles, resistant mutants.

TABLE 2

Sequence-based prediction of drug resistance mutations

| Method and year of report | Protein | Drugs | Number of mutant-drug samples (N, N _r , N _n) | Reported prediction accuracy (P, P _r , P _n or R value) |
|----------------------------|-----------------------------|--|---|--|
| HIVdb (HIV SEQ) 2002 [41] | HIV-1 protease | Amprenavir, Indinavir, Nelfinavir, Ritonavir, Saquinavir | N = 529 | P = 79%–97% |
| HIVdb 2004 [19] | HIV-1 protease | Amprenavir, Indinavir, Nelfinavir, Ritonavir, Saquinavir | N = 2460 | P _r = 81%–95% P _n = 96%–97% |
| ARS 2004 [19] | HIV-1 protease | Amprenavir, Indinavir, Nelfinavir, Ritonavir, Saquinavir | N = 2460 | P _r = 87%–97% P _n = 94%–97% |
| VGI 2004 [19] | HIV-1 protease | Amprenavir, Indinavir, Nelfinavir, Ritonavir, Saquinavir | N = 2460 | P _r = 84%–97% P _n = 94%–97% |
| Decision Tree 2002 [23] | HIV-1 protease | Amprenavir, Indinavir, Nelfinavir, Ritonavir, Saquinavir | N = 2148 | P _r = 82%–90% P _n = 87%–90% |
| Neural Network 2003 [21] | HIV-1 protease | Saquinavir | N _r = 32 N _n = 811 | P = 78% |
| Neural Network 2003 [20] | HIV-1 protease | Lopinavir | N _r = 267 N _n = 1055 | P _r = 92% P _n = 93% |
| SVM (Geno2pheno) 2003 [22] | HIV-1 protease | Amprenavir, Atazanavir, Indinavir, Lopinavir, Nelfinavir, Ritonavir, Saquinavir | N = 3683 | R = 0.78–0.89 |
| HIVdb (HIV SEQ) 2002 [41] | HIV-1 reverse transcriptase | Zidovudine, didanosine, zalcitabine, stavudine, abacavir, lamivudine, delavirdine, efavirenz, nevirapine | N=954 | P = 53%–96% |
| HIVdb 2004 [19] | HIV-1 reverse transcriptase | AZT, abacavir, didanosine, lamivudine, stavudine, zalcitabine, delavirdine, efavirenz, nevirapine, tenofovir | N = 4780 | P _r = 90% P _n = 97% |
| ARS 2004 [19] | HIV-1 reverse transcriptase | AZT, abacavir, didanosine, lamivudine, stavudine, zalcitabine, delavirdine, efavirenz, nevirapine, tenofovir | N = 4780 | P _r = 95% P _n = 97% |
| VGI 2004 [19] | HIV-1 reverse transcriptase | AZT, abacavir, didanosine, lamivudine, stavudine, zalcitabine, delavirdine, efavirenz, nevirapine, tenofovir | N = 4780 | P _r = 93% P _n = 96% |
| Decision Tree 2002 [23] | HIV-1 reverse transcriptase | AZT, abacavir, didanosine, lamivudine, stavudine, zalcitabine, delavirdine, efavirenz, nevirapine | N = 4104 | P _r = 58%–92% P _n = 62%–97% |
| SVM 2003 [22] | HIV-1 reverse transcriptase | AZT, abacavir, didanosine, lamivudine, stavudine, zalcitabine, delavirdine, efavirenz, nevirapine, tenofovir | N = 6018 | R = 0.54–0.85 |

N, N_r, and N_n are the number of all, resistant, and non-resistant mutant-drug samples, respectively. P, P_r, P_n values are the percentage of correct predictions for all mutations, resistance mutations, and non-resistance mutations, respectively.

a larger amount of drug resistance mutation data and a higher number of high-resolution three-dimensional structures for these two proteins. For instance, comprehensive resistance data for these two proteins are collected in the HIVdb database [18]. There are 198 and 102 entries of the three-dimensional structure of HIV-1 protease- and reverse transcriptase-inhibitor complexes respectively in the PDB database [43]. Significantly less data and structural information is available for other proteins with known resistance mutations, which likely contributes to the lack of progress in the development and testing of computer prediction methods for these proteins.

Known resistance mutations in other proteins have been used in combination with experimental molecular techniques as markers for predicting drug resistance mutations [44–49]. These exploit the observation that spe-

cific mutations found in resistance strains are absent in susceptible strains. Based on the analysis of clinical samples and the results of mutagenesis, simple rules ranging from identification of a single point mutation to more complex mutation selection algorithm can be derived for the prediction of drug resistance [45]. Experimental techniques capable of differentiating between a single wild-type sequence and mutant sequences can then be used to detect these mutations and predict drug resistance strains [46].

The simple rules generated and applied in these experimental studies can be directly used for developing a computer prediction system in a similar fashion like the HIV resistant genotype interpretation systems HIVdb, ARS, and VGI [18,19]. To predict mutations of a protein resistant to a particular drug, the sequence of that protein can be scanned to identify mutations that match the simple

TABLE 3

Prediction of drug resistance mutations by using simple rules

| Protein | Drugs | Rule and Year of Report | Number of Samples (N _r , N _n) | Prediction Accuracy (P _r , P _n value) |
|--|--|--|--|---|
| <i>Mycobacterium tuberculosis</i> arabinosyltransferase B | Ethambutol | embB codon 306 mutation, 1997 [46,64] | N = 118 | P _r = 60%~68% |
| <i>Mycobacterium tuberculosis</i> catalase-peroxidase | Isoniazid | katG codon 315 mutation, 2000 [46,65] | N = 79 | P _r = 33%~45% |
| <i>Mycobacterium tuberculosis</i> β subunit of RNA polymerase | Rifampicin | Any mutation at rpoB codon 531/526/516, 2002 [46,66,67] | N = 20 | P = 90%~98% |
| <i>Plasmodium falciparum</i> multi-drug resistance protein 1 | Choloroquine | Any mutation at fmdr1 codon 86/1042/1246, 2000 [68] | N _r = 40 N _n = 17 | P _r = 86%~100% |
| <i>Plasmodium falciparum</i> dihydrofolate reductase | Sulfadoxine-pyrimethamine | Codon 59 mutation, 2003 [47] | N = 327 | P _r = 81% |
| Hepatitis C INF-sensitive-determining region of nonstructural 5A protein | Interferon | Amino acid 2218 to be H, 2001 [69] | N _r = 36 N _n = 14 | P _r = 58% P _n = 86% |
| <i>Helicobacter pylori</i> 23S rRNA gene | Clarithromycin | A2142G or A2143G mutation, 2001 [44] | N = 299 | P _r = 42% |
| <i>Neisseria Meningitidis</i> penicillin-binding protein 2 | Penicillin | I566V mutation, 2003 [49] | N _r = 30 | P _r = 51% |
| HIV-1 reverse transcriptase | Abacavir | L74V or NRTI MDR or any 3 mutations of 41/184/210/215, 2002 [45] | N = 307 | P _r = 88% P _n = 77% |
| Human matrix metalloproteinase 3 | 5-fluorouracil-cisplatin PKC412, SU5614, K-252a, D-64406, | Codon 1612 Adenine deletion, 2004 [70] | N = 148 | P _r = 45% |
| Human receptor tyrosine kinase FLT3 | D-65476, DQPPC, AGL2043, TMPPP, GTP-14564 (anti-leukemia) | G697R mutation, 2004 [48] | N _r = 9 | P _r = 100% |

rules associated with that protein and the corresponding drug. Table 3 gives the protein-drug systems that have known drug resistance mutations and sufficiently accurate prediction rules reported in the literature. The simple rules for these protein-drug systems and the reported prediction accuracy derived from these rules are also given in Table 3. Based on the test of the currently available samples, these rules are capable of predicting resistance and non-resistance mutations at accuracies of 42%–100% and 77%–86% respectively, which are compared to those of 81%–97% and 91%–97% from the HIV resistant genotype interpretation systems HIVdb, ARS and VGI [18,19]. This suggests that these simple rules have a certain capacity for facilitating the prediction of specific drug resistance mutations and they may be used as the basis for developing more sophisticated interpretation systems like those of HIVdb, ARS and VGI [18,19].

Conclusions and perspectives

Both structure-based and sequence-based methods consistently show a promising capability for predicting resistance mutations. Structure-based methods are particularly useful

for mechanistic study and the prediction of resistant protein variants with little or no preliminary knowledge of known resistance mutations. However, they depend on the availability of a structural template, which significantly limits their application range. Advances in structural genomics are expected to expand the application range of structure-based methods to more extensive sets of disease proteins and drugs. Inadequately described interactions need to be more properly modeled. For instance, algorithms for modeling main-chain and side-chain conformational flexibility [50], ligand-protein-solvent interactions [51], hydration effects [52] and electrostatic interactions involving aromatic rings [53] can be incorporated into structure-based methods to increase their prediction accuracies.

No structural template is required by sequence-based methods. However, a sufficiently diverse set of resistance and non-resistance samples is needed for training a statistical learning system. Thus these methods are not applicable for disease proteins and drugs with little or no resistance mutation data. Mining of the resistance and non-resistance mutation data from the literature [4,5,54,55] and other sources [56] is a key to more extensive exploration of

sequence-based and rule-based methods. Databases such as HIVdb [18] and a rational database of HIV protein sequences of AIDS patients [57] that provide resistance mutation data are useful resources for serving this purpose, and more such databases are desired. Studies of resistance mutation patterns [44–46,58] are also important for determining the phenomenon and mechanism of drug resistance. Moreover, the performance of sequence-based methods may be further improved by introducing a more biologically meaningful representation of amino acids [42].

Sequence-based methods are less effective in prediction of resistant mutations for new drugs, when only limited training datasets are available. By contrast, structure-based methods are capable of predicting mutations resistant to a new drug, if a high quality template structure of the drug-protein complex is available. Moreover, sequence-based methods are usually not intended for facilitating mechanistic study of resistance mutations. Some of these methods, such as DT [23,59,60], NN regression [61] and

SVM regression [22,62,63], have the capacity for providing the contribution of specific attributes (structural and physicochemical properties of amino acids) to a classification (resistance or non-resistance mutation). By introducing a more biologically meaningful representation of amino acids [42], this capacity may be explored for probing structural and physicochemical features contributing to resistance mutations.

Acknowledgements

The authors would like to thank the support from the EU-China Scientific Collaborative Project Shanghai Commission for Science and Technology (03DJ14011), China '863' National High-Tech Program (2003AA231010, 2004BA711A21), China '973' National Key Basic Research Program (2001CB510203, 2003CB715901), and Singapore NUS-ARF (R-151-000-031-112). Zhi Wei Cao and Yu Zong Chen are also affiliated with the Shanghai Center for Bioinformatics Technology.

References

- Chaix-Couturier, C. *et al.* (2000) HIV-1 drug resistance genotyping. A review of clinical and economic issues. *Pharmacoeconomics* 18, 425–433
- Liu, H.H. (1999) Antibiotic resistance in bacteria. A current and future problem. *Adv. Exp. Med. Biol.* 455, 387–396
- Lehne, G. *et al.* (1998) Challenging drug resistance in cancer therapy—review of the First Nordic Conference on Chemoresistance in Cancer Treatment, October 9th and 10th, 1997. *Acta Oncol* 37 (5), 431–439
- Erickson, J.W. and Burt, S.K. (1996) Structural mechanisms of HIV drug resistance. *Annu. Rev. Pharmacol. Toxicol.* 36, 545–571
- Blanchard, J.S. (1996) Molecular mechanisms of drug resistance in *Mycobacterium tuberculosis*. *Annu. Rev. Biochem.* 65, 215–239
- Borst, P. (1991) Genetic mechanisms of drug resistance. A review. *Acta Oncol.* 30, 87–105
- Remy, S. *et al.* (2003) A novel mechanism underlying drug resistance in chronic epilepsy. *Ann. Neurol.* 53, 469–479
- Schafer, J.M. *et al.* (2002) A mechanism of drug resistance to tamoxifen in breast cancer. *J. Steroid Biochem. Mol. Biol.* 83, 75–83
- Weber, I.T. and Harrison, R.W. (1999) Molecular mechanics analysis of drug-resistant mutants of HIV protease. *Protein Eng.* 12, 469–474
- Sussman, F. *et al.* (1997) Solvation effects are responsible for the reduced inhibitor affinity of some HIV-1 PR mutants. *Protein Sci.* 6, 1024–1030
- Rick, S.W. *et al.* (1998) Molecular mechanisms of resistance: free energy calculations of mutation effects on inhibitor binding to HIV-1 protease. *Protein Sci.* 7, 1750–1756
- Piana, S. *et al.* (2002) Drug resistance in HIV-1 protease: Flexibility-assisted mechanism of compensatory mutations. *Protein Sci.* 11, 2393–2402
- Wang, W. and Kollman, P.A. (2001) Computational study of protein specificity: the molecular basis of HIV-1 protease drug resistance. *Proc. Natl. Acad. Sci. U. S. A.* 98, 14937–14942
- Chen, Y.Z. *et al.* (2001) Can an optimization/scoring procedure in ligand-protein docking be employed to probe drug-resistant mutations in proteins? *J. Mol. Graph. Model.* 19, 560–570
- Shenderovich, M.D. *et al.* (2003) Structure-based phenotyping predicts HIV-1 protease inhibitor resistance. *Protein Sci.* 12, 1706–1718
- Nair, A.C. *et al.* (2002) 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, 171–179
- Stoffler, D. *et al.* (2002) Evolutionary analysis of HIV-1 protease inhibitors: Methods for design of inhibitors that evade resistance. *Proteins* 48, 63–74
- Rhee, S.Y. *et al.* (2003) Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* 31, 298–303
- Zazzi, M. *et al.* (2004) Comparative evaluation of three computerized algorithms for prediction of antiretroviral susceptibility from HIV type 1 genotype. *J. Antimicrob. Chemother.* 53, 356–360
- Wang, D. and Larder, B. (2003) Enhanced prediction of lopinavir resistance from genotype by use of artificial neural networks. *J. Infect. Dis.* 188, 653–660
- Draghici, S. and Potter, R.B. (2003) Predicting HIV drug resistance with neural networks. *Bioinformatics* 19, 98–107
- Beerenwinkel, N. *et al.* (2003) Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Res.* 31, 3850–3855
- Beerenwinkel, N. *et al.* (2002) Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. *Proc. Natl. Acad. Sci. U. S. A.* 99, 8271–8276
- Ala, P.J. *et al.* (1997) Molecular basis of HIV-1 protease drug resistance: structural analysis of mutant proteases complexed with cyclic urea inhibitors. *Biochemistry* 36, 1573–1580
- Rozwarski, D.A. *et al.* (1998) Modification of the NADH of the isoniazid target (InhA) from *Mycobacterium tuberculosis*. *Science* 279, 98–102
- Avram, S. *et al.* (2002) Comparative study of some energetic and steric parameters of the wild type and mutants HIV-1 protease: a way to explain the viral resistance. *J. Cell. Mol. Med.* 6, 251–260
- Smith, M.B. *et al.* (2000) Monte Carlo calculations on HIV-1 reverse transcriptase complexed with the non-nucleoside inhibitor 8-Cl TIBO: contribution of the L100I and Y181C variants to protein stability and biological activity. *Protein Eng.* 13, 413–421
- Kuntz, I.D. (1992) Structure-based strategies for drug design and discovery. *Science* 257, 1078–1082
- Blundell, T.L. (1996) Structure-based drug design. *Nature* 384 (Suppl. 6604), 23–26
- Rarey, M. *et al.* (1996) A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* 261, 470–489
- Jones, G. *et al.* (1997) Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* 267, 727–748
- Wang, J. *et al.* (1999) Flexible ligand docking: a multistep strategy approach. *Proteins* 36, 1–19
- Charifson, P.S. *et al.* (1999) Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* 42, 5100–5109
- Muegge, I. and Martin, Y.C. (1999) A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.* 42, 791–804
- Eisenberg, D. and McLachlan, A.D. (1986) Solvation energy in protein folding and binding. *Nature* 319, 199–203
- Wang, R. *et al.* (1998) SCORE: a new empirical method for estimating the bind affinity of a protein-ligand complex. *J. Mol. Model.* 4, 379–394
- Chen, X. *et al.* (2004) Molecular dynamics simulations of 14 HIV protease mutants in complexes with indinavir. *J. Mol. Model. (Online)* 10, 373–381
- Kroeger Smith, M.B. *et al.* (1997) Molecular modeling of HIV-1 reverse transcriptase drug-resistant mutant strains: implications for the mechanism of polymerase action. *Protein Eng.* 10, 1379–1383

- 39 Arnold, E. *et al.* (1996) Targeting HIV reverse transcriptase for anti-AIDS drug design: structural and biological considerations for chemotherapeutic strategies. *Drug Des Discov* 13, 29–47
- 40 Hu, J.S. *et al.* (2002) A flexible approach of measuring mRNA expression levels of any genes using a universal oligonucleotide microarray. *EASEB JOURNAL* 16, A543–A543
- 41 Puchhammer-Stockl, E. *et al.* (2002) Comparison of virtual phenotype and HIV-SEQ program (Stanford) interpretation for predicting drug resistance of HIV strains. *HIV Med.* 3, 200–206
- 42 Thomson, R. *et al.* (2003) Characterizing proteolytic cleavage site activity using bio-basis function neural networks. *Bioinformatics* 19, 1741–1747
- 43 Berman, H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242
- 44 van Doorn, L.J. *et al.* (2001) Accurate prediction of macrolide resistance in *Helicobacter pylori* by a PCR line probe assay for detection of mutations in the 23S rRNA gene: multicenter validation study. *Antimicrob. Agents Chemother.* 45, 1500–1504
- 45 Walter, H. *et al.* (2002) Prediction of abacavir resistance from genotypic data: impact of zidovudine and lamivudine resistance *in vitro* and *in vivo*. *Antimicrob. Agents Chemother.* 46, 89–94
- 46 Victor, T.C. *et al.* (2002) Prediction of drug resistance in *M. tuberculosis*: molecular mechanisms, tools, and applications. *IUBMB Life* 53, 231–237
- 47 Talisuna, A.O. *et al.* (2003) Population-based validation of dihydrofolate reductase gene mutations for the prediction of sulfadoxine-pyrimethamine resistance in Uganda. *Trans. R. Soc. Trop. Med. Hyg.* 97, 338–342
- 48 Cools, J. *et al.* (2004) Prediction of resistance to small molecule FLT3 inhibitors: implications for molecularly targeted therapy of acute leukemia. *Cancer Res.* 64, 6385–6389
- 49 Stefanelli, P. *et al.* (2003) Prediction of decreased susceptibility to penicillin of *Neisseria meningitidis* strains by real-time PCR. *J. Clin. Microbiol.* 41, 4666–4670
- 50 Zavodszky, M.I. *et al.* (2004) Modeling correlated main-chain motions in proteins for flexible molecular recognition. *Proteins* 57, 243–261
- 51 Floriano, W.B. *et al.* (2004) HierVLS hierarchical docking protocol for virtual ligand screening of large-molecule databases. *J. Med. Chem.* 47, 56–71
- 52 Liu, Y. and Beveridge, D.L. (2002) Exploratory studies of ab initio protein structure prediction: multiple copy simulated annealing, AMBER energy functions, and a generalized born/solvent accessibility solvation model. *Proteins* 46, 128–146
- 53 Burghardt, T.P. *et al.* (2002) Cation-pi interaction in a folded polypeptide. *Biopolymers* 63, 261–272
- 54 Cheung, P.K. *et al.* (2004) 2004: which HIV-1 drug resistance mutations are common in clinical practice? *AIDS Rev.* 6, 107–116
- 55 Anderson, T.J. (2004) Mapping drug resistance genes in *Plasmodium falciparum* by genome-wide association. *Curr. Drug Targets Infect. Disord.* 4, 65–78
- 56 Weinstein, J.N. *et al.* (1997) An information-intensive approach to the molecular pharmacology of cancer. *Science* 275, 343–349
- 57 Tamalet, C. *et al.* (2003) Resistance of HIV-1 to multiple antiretroviral drugs in France: a 6-year survey (1997–2002) based on an analysis of over 7000 genotypes. *AIDS* 17, 2383–2388
- 58 Yahi, N. *et al.* (1999) Mutation patterns of the reverse transcriptase and protease genes in human immunodeficiency virus type 1-infected patients undergoing combination therapy: survey of 787 sequences. *J. Clin. Microbiol.* 37, 4099–4106
- 59 Russell Stuart, J. and Norvig, P. (2003) *Artificial intelligence: a modern approach*, Prentice Hall
- 60 Quinlan, J.R. (1986) Induction of Decision Trees. *Mach. Learn.* 1, 81–106
- 61 Rutkowski, L. (2004) Generalized regression neural networks in time-varying environment. *IEEE Trans. Neural Netw.* 15, 576–596
- 62 Wang, Y.M. *et al.* (2003) Nonlinear estimation and modeling of fMRI data using spatio-temporal support vector regression. *Inf Process Med Imaging* 18, 647–659
- 63 Cherkassky, V. and Ma, Y. (2004) Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Netw.* 17, 113–126
- 64 Sreevatsan, S. *et al.* (1997) Ethambutol resistance in *Mycobacterium tuberculosis*: critical role of embB mutations. *Antimicrob. Agents Chemother.* 41, 1677–1681
- 65 Kiepiela, P. *et al.* (2000) Genomic mutations in the katG, inhA and aphC genes are useful for the prediction of isoniazid resistance in *Mycobacterium tuberculosis* isolates from Kwazulu Natal, South Africa. *Tuber. Lung Dis.* 80, 47–56
- 66 Whelen, A.C. *et al.* (1995) Direct genotypic detection of *Mycobacterium tuberculosis* rifampin resistance in clinical specimens by using single-tube heminested PCR. *J. Clin. Microbiol.* 33, 556–561
- 67 Williams, D.L. *et al.* (1998) Contribution of rpoB mutations to development of rifamycin cross-resistance in *Mycobacterium tuberculosis*. *Antimicrob. Agents Chemother.* 42, 1853–1857
- 68 Flueck, T.P. *et al.* (2000) Correlation of *in vivo* resistance to chloroquine and allelic polymorphisms in *Plasmodium falciparum* isolates from Uganda. *Trop. Med. Int. Health* 5, 174–178
- 69 Nishiguchi, S. *et al.* (2001) Method to detect substitutions in the interferon-sensitivity-determining region of hepatitis C virus 1b for prediction of response to interferon therapy. *Hepatology* 33, 241–247
- 70 Blons, H. *et al.* (2004) Matrix metalloproteinase 3 polymorphism: a predictive factor of response to neoadjuvant chemotherapy in head and neck squamous cell carcinoma. *Clin. Cancer Res.* 10, 2594–2599