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.

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...
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)

\[ V = \sum_{i<j} \left( \frac{A_{ij}}{r_{ij}^2} - \frac{B_{ij}}{r_{ij}} \right) + \sum_{\text{elec}} \left( \frac{q_i q_j}{r_{ij}} + E(r_{ij}) \right) + \sum_{\text{atoms}} \sigma_i A_i + \sum_{\text{residue}} \alpha_k N_k + \sum_{\text{H bonds}} V_H \]

Where \( A_{ij} \) and \( B_{ij} \) are van der Waals parameters; \( \varepsilon_i \) 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 \( \alpha_k \) 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.
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. A highly accurate scoring system can be achieved by combining known resistance mutations with rules to predict resistance mutations. The pattern of scoring functions provides a high degree of resistance for each drug.

**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 \( x_p \) 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 1 where an example of SVM prediction of resistance mutations is provided.

**TABLE 1**

<table>
<thead>
<tr>
<th>Method and year of report</th>
<th>Protein</th>
<th>Drugs</th>
<th>Number of mutant-drug complexes</th>
<th>Reported prediction accuracy (R or P value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular modeling 1999 [9]</td>
<td>HIV-1 protease</td>
<td>Indinavir, Saquinavir</td>
<td>17</td>
<td>R = 0.68–0.76</td>
</tr>
<tr>
<td>Molecular modeling 2001 [13]</td>
<td>HIV-1 protease</td>
<td>Amprenavir, Indinavir, Nelfinavir, Ritonavir, Saquinavir</td>
<td>33</td>
<td>P = 57%–86%</td>
</tr>
<tr>
<td>Molecular modeling with simple optimization scoring procedure 2001 [14]</td>
<td>HIV-1 protease</td>
<td>MK639, Saquinavir, SB203386, U99360E, VX478</td>
<td>22</td>
<td>R = 0.64</td>
</tr>
<tr>
<td>Molecular modeling 2002 [16]</td>
<td>HIV-1 protease</td>
<td>Ritonavir</td>
<td>12</td>
<td>R = 0.7</td>
</tr>
<tr>
<td>Fitness evolution model 2002 [17]</td>
<td>HIV-1 protease</td>
<td>Indinavir, Nelfinavir, Ritonavir, Saquinavir</td>
<td>11</td>
<td>P = 45% (81% for resistance site prediction)</td>
</tr>
<tr>
<td>Molecular modeling 2003 [15]</td>
<td>HIV-1 protease</td>
<td>Indinavir, Lopinavir, Nelfinavir, Ritonavir, Saquinavir</td>
<td>750</td>
<td>R = 0.87 for 360 complexes. P = 86% and 92%</td>
</tr>
<tr>
<td>Neural network model of homology modeled structures 2003 [21]</td>
<td>HIV-1 protease</td>
<td>Indinavir</td>
<td>38</td>
<td>P = 60%–70%</td>
</tr>
<tr>
<td>Molecular Dynamics simulation 2004 [37]</td>
<td>HIV-1 protease</td>
<td>Indinavir</td>
<td>12</td>
<td>R = 0.98</td>
</tr>
<tr>
<td>Molecular modeling 1997 [38]</td>
<td>HIV-1 reverse transcriptase</td>
<td>8-CL TIBO, α-APA</td>
<td>8</td>
<td>R = 0.80–0.97</td>
</tr>
<tr>
<td>Monte Carlo Simulation 2000 [27]</td>
<td>HIV-1 reverse transcriptase</td>
<td>8-CL TIBO</td>
<td>2</td>
<td>R = 0.78</td>
</tr>
<tr>
<td>Molecular modeling with simple optimization scoring procedure 2001 [14]</td>
<td>HIV-1 reverse transcriptase</td>
<td>Nevirapine, TIBO R82913</td>
<td>13</td>
<td>R = 0.31</td>
</tr>
</tbody>
</table>

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).
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 $y_i = (w \cdot x_i + b)$, where $w$ is a vector normal to the hyperplane, is the perpendicular distance from the hyperplane to the origin with $\|w\|$ as the Euclidean norm of $w$, $y_i = +1$ 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 $P$ 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 amino acids 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
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 specific 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>
<thead>
<tr>
<th>Method and year of report</th>
<th>Protein</th>
<th>Drugs</th>
<th>Number of mutant-drug samples (N, N_r, N_n)</th>
<th>Reported prediction accuracy (P, P_r, P_n or R value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIVdb (HIV SEQ) 2002 [41]</td>
<td>HIV-1 protease</td>
<td>Amprenavir, Indinavir, Nelfinavir, Ritonavir, Saquinavir</td>
<td>N = 529</td>
<td>P = 79%–97%</td>
</tr>
<tr>
<td>HIVdb 2004 [19]</td>
<td>HIV-1 protease</td>
<td>Amprenavir, Indinavir, Nelfinavir, Ritonavir, Saquinavir</td>
<td>N = 2460</td>
<td>P_r = 81%–95%</td>
</tr>
<tr>
<td>ARS 2004 [19]</td>
<td>HIV-1 protease</td>
<td>Amprenavir, Indinavir, Nelfinavir, Ritonavir, Saquinavir</td>
<td>N = 2460</td>
<td>P_r = 87%–97%</td>
</tr>
<tr>
<td>VGI 2004 [19]</td>
<td>HIV-1 protease</td>
<td>Amprenavir, Indinavir, Nelfinavir, Ritonavir, Saquinavir</td>
<td>N = 2460</td>
<td>P_r = 94%–97%</td>
</tr>
<tr>
<td>Decision Tree 2002 [23]</td>
<td>HIV-1 protease</td>
<td>Amprenavir, Indinavir, Nelfinavir, Ritonavir, Saquinavir</td>
<td>N = 2148</td>
<td>P_r = 82%–90%</td>
</tr>
<tr>
<td>Neural Network 2003 [21]</td>
<td>HIV-1 protease</td>
<td>Saquinavir</td>
<td>N_r = 32</td>
<td>P = 78%</td>
</tr>
<tr>
<td>Neural Network 2003 [20]</td>
<td>HIV-1 protease</td>
<td>Lopinavir</td>
<td>N_r = 267</td>
<td>P = 92%</td>
</tr>
<tr>
<td>SVM (Geno2pheno) 2003 [22]</td>
<td>HIV-1 protease</td>
<td>Amprenavir, Atazanavir, Indinavir, Lopinavir, Nelfinavir, Ritonavir, Saquinavir</td>
<td>N = 3683</td>
<td>R = 0.78–0.89</td>
</tr>
<tr>
<td>HIVdb (HIV SEQ) 2002 [41]</td>
<td>HIV-1 reverse transcriptase</td>
<td>Zidovudine, didanosine, zalcitabine, stavudine, abacavir, lamivudine, delavirdine, efavirenz, nevirapine</td>
<td>N = 954</td>
<td>P = 53%–96%</td>
</tr>
<tr>
<td>HIVdb 2004 [19]</td>
<td>HIV-1 reverse transcriptase</td>
<td>AZT, abacavir, didanosine, lamivudine, stavudine, zalcitabine, delavirdine, efavirenz, nevirapine, tenofovir</td>
<td>N = 4780</td>
<td>P_r = 90%</td>
</tr>
<tr>
<td>ARS 2004 [19]</td>
<td>HIV-1 reverse transcriptase</td>
<td>AZT, abacavir, didanosine, lamivudine, stavudine, zalcitabine, delavirdine, efavirenz, nevirapine, tenofovir</td>
<td>N = 4780</td>
<td>P_r = 95%</td>
</tr>
<tr>
<td>VGI 2004 [19]</td>
<td>HIV-1 reverse transcriptase</td>
<td>AZT, abacavir, didanosine, lamivudine, stavudine, zalcitabine, delavirdine, efavirenz, nevirapine, tenofovir</td>
<td>N = 4780</td>
<td>P_r = 93%</td>
</tr>
<tr>
<td>Decision Tree 2002 [23]</td>
<td>HIV-1 reverse transcriptase</td>
<td>AZT, abacavir, didanosine, lamivudine, stavudine, zalcitabine, delavirdine, efavirenz, nevirapine</td>
<td>N = 4104</td>
<td>P_r = 58%–92%</td>
</tr>
<tr>
<td>SVM 2003 [22]</td>
<td>HIV-1 reverse transcriptase</td>
<td>AZT, abacavir, didanosine, lamivudine, stavudine, zalcitabine, delavirdine, efavirenz, nevirapine, tenofovir</td>
<td>N = 6018</td>
<td>R = 0.54–0.85</td>
</tr>
</tbody>
</table>

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.
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

| TABLE 3 | Prediction of drug resistance mutations by using simple rules |
|-----------------|-----------------|-----------------|-----------------|
| Protein | Drugs | Rule and Year of Report | Number of Samples (N, N_r, N_n) | Prediction Accuracy (P, P_r, P_n value) |
| Mycobacterium tuberculosis arabinosyltransferase B | Ethambutol | embB codon 306 mutation, 1997 [46,64] | N = 118 | P_r = 60%–68% |
| Mycobacterium tuberculosis catalase-peroxidase | Isoniazid | katG codon 315 mutation, 2000 [46,65] | N = 79 | P_r = 33%–45% |
| Mycobacterium tuberculosis β subunit of RNA | Rifampicin | Any mutation at rpoB codon 531/526/516, 2000 [46,66,67] | N = 20 | P_r = 90%–98% |
| Plasmodium falciparum multi-drug resistance protein 1 | Choloroquine | Any mutation at fmdr1 codon 86/1042/1246, 2000 [68] | N_r = 40 | P_r = 86%–100% |
| Plasmodium falciparum 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 | P_r = 58% |
| Helicobacter pylori 23S rRNA gene | Clarithromycin | A2142G or A2143G mutation, 2001 [44] | N = 299 | P_r = 42% |
| Neisseria Meningitidis penicillin-binding protein 2 | Penicillin | IS660V 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% |
| Human matrix metalloproteinase 3 | S-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% |
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 factors 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

10 Sussman, E. et al. (1997) Solvation effects are responsible for the reduced inhibitor affinity of some HIV-1 PR mutants. Protein Sci. 6, 1024–1030
27 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