An eigen-binding site based method for the analysis of anti-EGFR drug resistance in lung cancer treatment

Lichun Ma, *Student Member, IEEE,* Debby D. Wang, *Member, IEEE,* Bin Zou, *Student Member, IEEE,* and Hong Yan, *Fellow, IEEE*

Abstract—We explore the drug resistance mechanism in non-small cell lung cancer treatment by characterizing the drug-binding site of a protein mutant based on local surface and energy features. These features are transformed to an eigen-binding site space and used for drug resistance level prediction and analysis.

Index Terms—Epidermal growth factor receptor, non-small-cell lung carcinoma, tyrosine kinase inhibitor, gefitinib, binding free energy, binding site, alpha shape modeling.

1 Introduction

Lung cancer is a leading cause of cancer deaths worldwide [1], [2], [3], [4]. As a primary type of lung cancer, non-small-cell lung cancer (NSCLC) is an important research topic [5], [6], [7]. Clinically, reversible tyrosine kinase (TK) inhibitor (TKI) - gefitinib, which targets the kinase domain of epidermal growth factor receptor (EGFR), is widely used in the treatment of NSCLC [8], [9], [10]. This treatment is especially effective for those patients harboring activating mutations at the EGFR kinase domain [11]. However, the drug's effectiveness gradually decreases after a period of time, mostly due to a second EGFR mutation [12], [13], [14]. There is an urgent need to study this EGFR mutation-induced drug resistance and to explore its molecular mechanisms, which will benefit the development of new and specialized therapies.

Both experimental and computational methods have been developed to study the molecular mechanisms of EGFR mutation-induced drug resistance [15], [16], [17], [18]. These studies are primarily focused on decoding the EGFR-downstream signaling or EGFR-drug binding affinity. By observing lung cancer cell lines with MET amplification and those with inhibition of MET signaling, Engelman et al. [19] proposed that MET amplification could strengthen PI3K/Akt signaling by associating with ErbB-3, which leads to gefitinib-resistance in lung cancer treatments. Yun et al. [20] showed that a second mutation T790M in EGFR tyrosine kinase domain triggers gefitinib-resistance, as the mutation results in an increased affinity for this EGFR mutant and adenosine triphosphate (ATP). In recent years, computational methods have become very popular and have been effectively applied to the studies of drug resistance, with advantages of high efficiency and low cost [21]. Taking patient personal features into consideration, Wang et al. [22] employed binding free energies between EGFR mutants and inhibitors to predict EGFR mutation-induced drug resistance effectively. In our previous work [23], the molecular mechanisms of drug resistance was investigated from local surface geometric properties of EGFR, which showed a close relationship between the surface curvature of drug-binding pocket in EGFR and the progression-free survival (PFS) of a patient.

In this paper, we study the EGFR mutation-induced drug resistance, by analyzing the geometric properties of the drug-binding site on EGFR and the binding affinity between EGFR and a drug molecule (Fig. 1). Inspired by the concept of eigenface in human face recognition, in which original faces can be projected to the eigenface space to achieve feature dimension reduction, we employ the components obtained from principal component analysis (PCA) to represent the EGFR drug-binding site. In our method, Rosetta [24] was employed to generate the three-dimensional (3D) structures of EGFR mutants, using the crystal structure of wild-type (WT) EGFR as a template. Subsequently, we extracted the geometric properties of the drug-binding sites on these modeled mutants, based on alpha shape modeling [25], [26]. Amber [27] was used to carry out molecular dynamics (MD) simulations and to calculate binding free energies between our mutants and a drug molecule. The geometric properties of drug-binding site on EGFR coupled with the mutant-drug binding affinity were regarded as our principal features, to characterize an EGFR mutant. PCA [28] was used on the derived features, to project these original features to those in a new eigen-binding site space. Finally, the support vector machine (SVM) [29] was adopted to build a classification model that correlates the new binding-site features with the

- *L. Ma is with the Department of Electronic Engineering, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong, China. E-mail: lichunma2-c@my.cityu.edu.hk.*
- *D.D. Wang is with the Department of Electronic Engineering, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong, China. E-mail: debby.d.wang@gmail.com.*
- *B. Zou is with the Department of Electronic Engineering, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong, China. E-mail: binzou2-c@my.cityu.edu.hk.*
- *H. Yan is with Department of Electronic Engineering, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong, China. E-mail: h.yan@cityu.edu.hk.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2568184, IEEE/ACM Transactions on Computational Biology and Bioinformatics

response level to gefitinib for NSCLC patients.

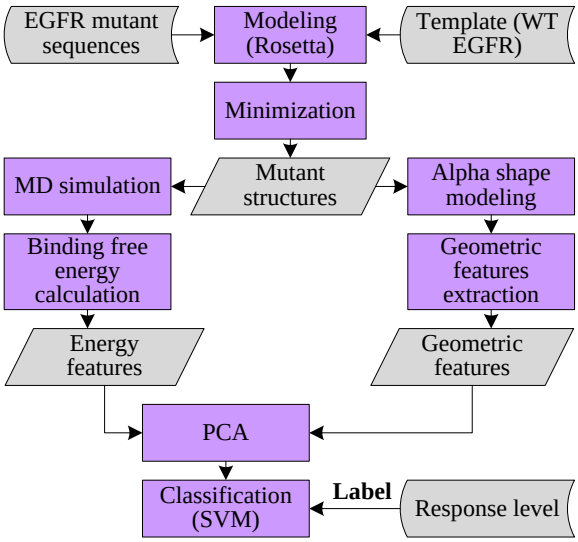

Fig. 1. Procedures used in the analysis of EGFR mutation-induced drug resistance based on eigen-binding site.

2 METHODS

2.1 Modeling of EGFR mutants

We produced the 3D structures of EGFR mutants based on the crystal structure of WT EGFR (2ITY), from the Protein Data Bank (PDB) [30]. We used the high-resolution ***ddg_monomer*** (HRDM) protocol in Rosetta software suite [31] to predict point mutation (amino acids substitution), and the ***comparative modeling*** (CM) protocol [32] to handle multi-point mutation (amino acids insertion, deletion, duplication and modification) [23].

Both protocols rely on a 3D template structure (WT EGFR) and a mutation sequence, for the modeling of each mutant. The Rosetta scoring function can be briefly summarized in the following equation [33]:

$$\begin{aligned} &P\left(structure \mid sequence\right) \\ &\propto P\left(sequence \mid structure\right) \times P\left(structure\right) \end{aligned} \tag{1}$$

The HRDM protocol allows protein backbones to fluctuate to a certain degree, and eliminates the range using Cα-Cα distance restraints. The CM protocol was implemented based on three steps. First, ***ClustalW*** [34] was applied to align the mutant sequence to the template sequence. Subsequently, the 3D mutant structure was predicted by copying the template structure of the well-aligned regions and reconstructing the remaining regions using loop modeling [33]. Finally, a full-atom refinement [35] was carried out to improve the predicted model, and the structure with the lowest energy was selected as the final output.

We further implemented a minimization step using Amber to optimize the predicted structure. This optimized structure was aligned to the template (2ITY) using UCSF Chimera [36] to form a complex with the drug molecule (gefitinib) to calculate the binding free energy between them.

2.2 Alpha shape modeling and solid angle calculation

Alpha shape modeling was employed to approximate the molecular surface of an EGFR mutant with a computational geometric shape. In this study, weighted alpha shapes [25], [26] were used to represent the surfaces of EGFR mutants. Each atom of the protein structure corresponds to a weighted point in 3D space, and the weight of a point is defined as the squared Van der Waals (VDW) radius. Suppose two atoms can be described as $\mathbf{a_1} = (\boldsymbol{p_1}, w_1)$ and $\mathbf{a_2} = (\boldsymbol{p_2}, w_2)$, where $\boldsymbol{p_1}$ and $\boldsymbol{p_2}$ stand for the locations of these two atoms in the 3D space, and w_1 and w_2 are the weights of these two points respectively. The two atoms can be defined to be orthogonal ($\perp$) or sub-orthogonal ($\perp_S$) in the following equation:

$$\begin{cases} if \mid \boldsymbol{p_1 p_2} \mid = w_1 + w_2, \ then \ \mathbf{a_1} \perp \mathbf{a_2} \\ if \mid \boldsymbol{p_1 p_2} \mid > w_1 + w_2, \ then \ \mathbf{a_1} \perp_S \mathbf{a_2} \end{cases} \tag{2}$$

For a given α, the weighed alpha shape of a protein can be generated with the simplices where there exists one weighed point that satisfies orthogonal condition to the weighed points of the simplex and satisfies sub-orthogonal condition to those of the others. The Computational Geometry Algorithms Library (CGAL) [37] was adopted to compute the alpha shape of each EGFR mutant.

After deriving the alpha shape of an EGFR mutant, we used the solid angle [38] of each surface atom to describe the local surface geometric properties. The solid angle Ω_i at vertex **P** in a tetrahedron **PABC** can be calculated as:

$$\Omega_i = \varphi_{ab} + \varphi_{bc} + \varphi_{ac} - \pi \tag{3}$$

where φ_{ab}, φ_{bc} and φ_{ac} represent the dihedral angles between **PAC** and **PBC**, **PAB** and **PAC**, **PAB** and **PBC**, respectively. By summing up the solid angles of all the tetrahedrons originated from the same atom, we can average them to obtain its solid angle Ω. Then it is transformed to the range of [-1, 1] by $\Omega' = \cos(\Omega/4)$. If $\Omega' > 0$, it is a convex shape. Otherwise, it is a concave shape.

2.3 Molecular dynamics (MD) simulations and binding free energy calculation

Before calculating the binding free energies between the EGFR mutants and gefitinib, we carried out MD simulations using Amber to optimize and equilibrate the mutant-drug systems. MD simulations are based on the Newton's second law of motion. Relying on this motion equation, the trajectories of positions, velocities and accelerations of each atom can be obtained.

We performed Amber simulations as follows. First, we generated a TIP3P water box for each mutant-drug complex using *tleap* program, as the simulations are conducted in a computational solvent environment. Then the generally-used ***ff99SB*** force field was employed. The total energy in the force field consists of bonded terms, related to bond stretching, angle bending and torsion terms of covalent bonds, and non-bonded terms, represented by long-range electrostatic forces and van der Waals (VDW) forces:

$$\begin{aligned} E_{total} = E_{stretch} + E_{bend} + E_{torsion} \\ + E_{electrostatic} + E_{vdw} \end{aligned} \tag{4}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2568184, IEEE/ACM Transactions on Computational Biology and Bioinformatics

After the optimization and equilibration of each system, a production MD simulation was implemented to obtain the motion trajectories of each mutant-drug complex.

Using these trajectories, the binding free energy in the corresponding mutant-drug system can be calculated. The MM-GBSA protocol in Amber tools was adopted for the calculations. In this protocol, the thermodynamic cycle of the solvent and vacuum environments is the core step, which avoids massive computations between bound and unbound states of two solvated molecules, and simply expresses the binding free energy in the solvent environment ($\triangle G_{bind,solv}$) as follows:

$$\begin{aligned}\triangle G_{bind,solv} = \triangle G_{bind,vacuum} + \triangle G_{solv,complex} \\ - (\triangle G_{solv,ligand} + \triangle G_{solv,receptor})\end{aligned} \quad (5)$$

where $\triangle G_{bind,vacuum}$ is the binding free energy for a receptor-ligand system in the vacuum environment. $\triangle G_{solv,receptor}$, $\triangle G_{solv,ligand}$ and $\triangle G_{solv,complex}$ represent the free energy differences between the solvent and vacuum environments, respectively for the receptor, ligand and complex.

2.4 Principal component analysis (PCA)

PCA is widely used in many studies for dimension reduction, through converting a set of possibly correlated features into a set of linearly uncorrelated variables called principal components (PCs) [39], [40]. In other words, the original feature space is transformed into a new PC space. Given a data set with real numbers, $\boldsymbol{X} = \{\boldsymbol{x^{(1)}}, \boldsymbol{x^{(2)}}, \ldots, \boldsymbol{x^{(m)}}\}$ where $\boldsymbol{x^{(i)}}$ is an n-dimensional vector, a normalization process (zero mean and unit variance) is first implemented to obtain the normalized vectors $\{\boldsymbol{z^{(1)}}, \boldsymbol{z^{(2)}}, \ldots, \boldsymbol{z^{(m)}}\}$. Then the covariance matrix can be computed using the following equation:

$$\boldsymbol{\Sigma} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{z^{(i)}}\boldsymbol{z^{(i)T}} \quad (6)$$

The PC space is constructed by the corresponding eigenvectors or PCs $(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \ldots, \boldsymbol{\mu_k})$ of the largest k eigenvalues of the covariance matrix $\boldsymbol{\Sigma}$. The new vectors $\boldsymbol{p^{(i)}}$ with k ($k < n$) features can be derived by projecting the original features to the PC space:

$$\boldsymbol{p^{(i)}} = (\boldsymbol{\mu_1^T z^{(i)}}, \boldsymbol{\mu_2^T z^{(i)}}, \ldots, \boldsymbol{\mu_k^T z^{(i)}}) \quad (7)$$

In our work, the previously extracted mutant features, namely the geometric properties of drug-binding site and the mutant-drug binding free energy, represent an original feature space. Using PCA, this feature space can be projected to a new PC space, called the eigen-binding site space. This eigen-binding site space can reveal important properties of our sampled mutants. Accordingly, based on the new-space features of our mutants, a further analysis can be implemented.

2.5 Classification based on support vector machines (SVMs)

After deriving new features of our mutants in the eign-binding site space, we built a classification model using SVM to group these mutant samples.

In decades, SVMs have been used effectively to solve both linear and non-linear classification problems [41], [42]. Given a training data set $\boldsymbol{D} = \{(\boldsymbol{x_i}, y_i) \mid \boldsymbol{x_i} \in R^d, y_i \in \{-1, 1\}, i = 1, 2, \ldots, N\}$ and a decision function $y_i[\boldsymbol{\omega}^T\boldsymbol{x_i} + b] \geq 1$, a two-class SVM searches for an optimal hyperplane that separates the training samples and maximizes the margin between the support vectors (SVs) located nearest to this hyperplane. By defining the Lagrange multiplier $\boldsymbol{\alpha}$, the problem can finally be reduced to maximizing $L(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$, as shown in the following equation:

$$\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j \boldsymbol{x_i^T x_j} \quad (8)$$

For nonlinear classification, kernel functions can be used. In this work, the Gaussian radial basis function $k(\boldsymbol{x_i}, \boldsymbol{x_j}) = exp(-g\|\boldsymbol{x_i} - \boldsymbol{x_j}\|^2)$, which can be rewritten into $k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \phi(\boldsymbol{x_i})^T\phi(\boldsymbol{x_j})$, was employed. In this regard, the optimization problem is transformed into the following one:

$$\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j \phi(\boldsymbol{x_i})^T\phi(\boldsymbol{x_j}) \quad (9)$$

3 Results and Discussion

3.1 Data analysis

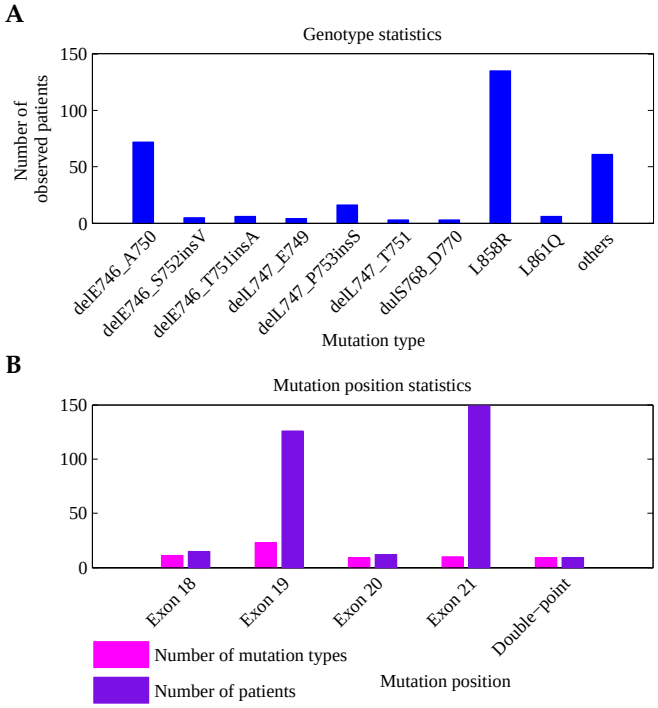

Fig. 2. (A) Statistics of the 62 mutation types occurred in 311 observed patients. (B) Distributions of the mutation types as well as the number of patients at each exon.

The data used in this paper were obtained from a combination of various literatures [22], [23], [43], [44], [45]. Specifically, there are 311 NSCLC patients harboring a total of 62 mutation types (Fig. 2A). Gefitinib was used for all these patients during their treatments. As shown in Fig. 2A, L858R and delE746_A750 are the most common mutations in these patients. The naming rules for mutations are established according to the differences of amino acid sequences

between the corresponding mutants and the WT EGFR. All these EGFR mutations occur in the EGFR TK domain, corresponding to exons 18 to 21 at the gene level. The distributions of patients and mutation types along the exon positions are shown in Fig. 2B. Exon 19 holds the largest number of mutation types, while most of patients harbor mutations at exon 21. A double-point mutation implies substitutions of amino acids at two positions of the WT EGFR sequence. Based on the sequence information of these EGFR mutants, we can generate their 3D structures based on the structure of WT EGFR PDB: 2ITY).

3.2 Prediction of EGFR mutants and calculation of binding free energy

EGFR mutant structures were generated using Rosetta based on the 3D structure of WT EGFR and the mutated sequences. We employed the scoring function with full-atom energy to evaluate those structures, and the one with the lowest energy was selected for each mutant. Then Amber was used to conduct 1000 minimization steps to refine the structures. Fig. 3 shows the WT EGFR and several examples of our predicted mutant structures (displayed using UCSF Chimera).

Posterior to the generation of mutant structures, we aligned each of them to the template (PDB: 2ITY) to form a complex with gefitinib. Then the complex was computationally solvated into a TIP3P water box, with a 10.0-angstrom (Å) buffer around the complex. A series of equilibration steps were subsequently carried out for each solvated complex. Specifically, 1000 circles of energy minimization, 50 ps of heating, 50 ps of density equilibration and 500 ps of constant pressure equilibration were conducted. To verify the equilibration of the systems, we checked the curves of temperature, density, total energy and backbone root-mean-square deviation (RMSD) for each system, and Fig. 4 presents an example. In this figure, the density, temperature or total energy of the delL747_K754insSR-gefitinib system converges in the equilibration period, and the backbone RMSD is acceptably stabilized.

After the equilibration period, a production MD simulation of 2 ns was implemented, with the trajectories of the involved complex recorded. Similarly, temperature, density, total energy and backbone RMSD were checked to validate the stabilization of the system. Based on these produced trajectories, we can further calculate the binding free energy of each mutant-drug complex, using the MM-GBSA protocol in Amber. This binding free energy, coupled with a number of local geometric properties of the mutant, plays an important role in characterizing the drug-binding site of a mutant.

3.3 Characterization of the drug-binding site on an EGFR mutant

Protein function sites control molecular interactions in biological processes, and thus are crucial for protein function annotation and rational drug design. Many studies have focused on analyzing and predicting these functionally important sites, providing various methods for their representation. These methods commonly fall into two categories, sequence- and structure-based approaches. Sequence-based methods are easily implemented by employing features of the amino acid residues in the functional important regions without using any structural information [46], [47]. Murakami and Mizuguchi used predicted accessibility and position-specific scoring matrix as sequence features to represent the potential function site and achieved a best prediction accuracy of 66.4% with different interface definitions [48]. Ofran and Rost employed features such as evolutionary profiles, predicted solvent accessibility, secondary structure and amino acid composition to conduct prediction and obtained an accuracy of 68% [49]. The limitation that only sequence information is used to represent the protein function site makes it difficult to improve the prediction accuracy. Structure-based methods usually describe the function site using surface geometric properties of the function sites as well as physicochemical features, such as hydrogen bond, hydrophobicity of the side-chain and electrostatic potential [50], [51], [52], [53]. Porollo and Meller [54] applied structure features coupled with sequence ones (features based on protein tertiary structure, single sequence-based attributes, evolutionary profiles and relative solvent accessibility) to describe the function site and yielded an overall classification accuracy of about 74%. Zhou and Yan computed alpha shape model of the protein and extracted features such as residue index (the percentage of each residue type at the function site), curvature (cleft or knob level) and connectivity (the connection between surface atoms) of the function site [55]. Using these features, they carried out protein function site prediction and achieved an accuracy of 68.8%.

Considering the specific problem of representing the EGFR drug-binding site, when choosing features, we should select those that can distinguish different mutations. In this work, we finally used 14 features to represent an EGFR drug-binding site, including surface curvature, connectivity, atom index and binding free energy.

Surface curvature

We extracted local geometric properties of each mutant, based on alpha shape modeling. Curvatures of surface atoms belonging to the drug-binding site were calculated (Figs. 5A and 5B). In this process, the CGAL was employed to build the alpha shape of each mutant (Fig. 5C). Specifically, 14 amino acids residues of 102 atoms are located at the binding site of WT EGFR. Depending on the derived alpha shape of each mutant, we calculated the solid angles of the binding-site surface atoms. Finally, the curvature of a binding pocket is represented by seven terms (named as *curvature indices* 1∼7): the number of convex atoms having a solid angle belonging to [0.5, 1], [0.61, 1] and [0.71, 1], the average and variance of convex atoms, and the average and variance of concave atoms. The average and variance of convex atoms stand for the knob level of the binding pocket. Similarly, the average and variance of concave atoms represent the cleft level. In [10], the number of convex atoms with a solid angle in [0.5, 1], [0.61, 1] and [0.71, 1] are highly correlated with the PFS of a patient, therefore we adopted them to characterize the binding sites.

Connectivity

The connectivity of a surface atom corresponds to the number of edges related to this atom in the alpha shape [56]. It is a parameter to represent the connection between one atom and other surface atoms. We calculated the average

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2568184, IEEE/ACM Transactions on Computational Biology and Bioinformatics

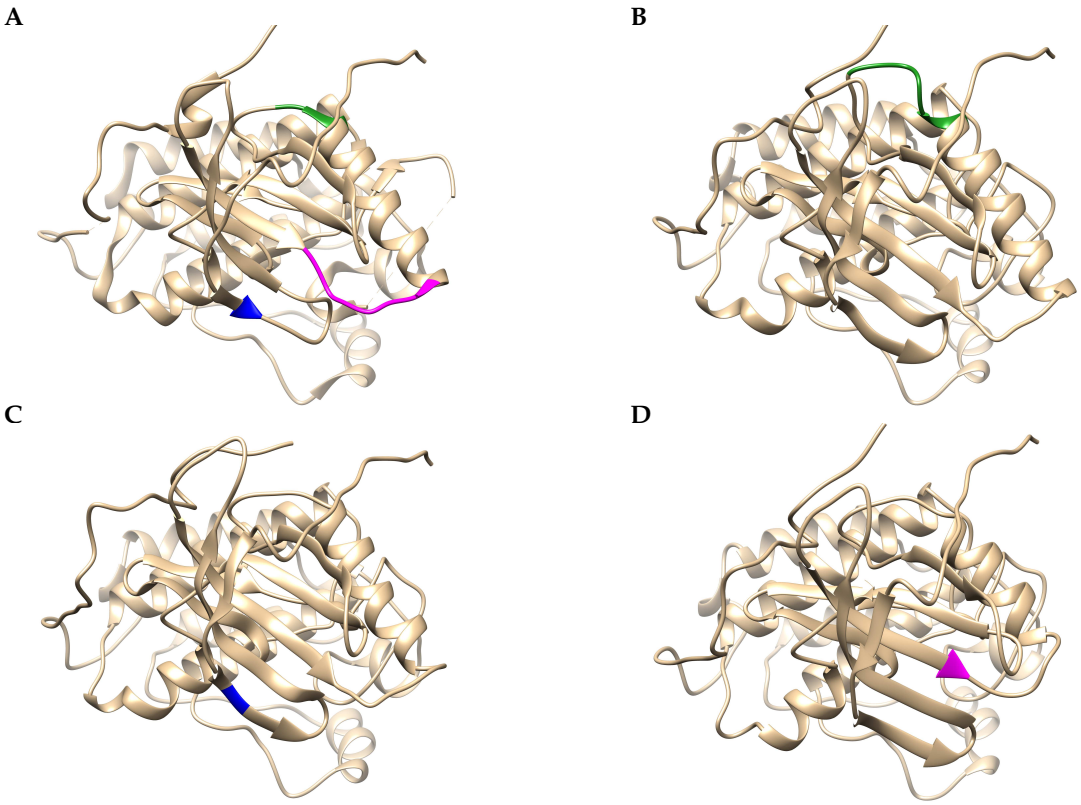

Fig. 3. Crystal structure of (A) WT EGFR and the predicted mutant structures of (B) dulS768_D770, (C) G23S, and (D) delE746_S752insV. The original sites and the corresponding mutant sites (dulS768_D770, G23S, delE746_S752insV) are shown in green, blue and magenta, respectively.

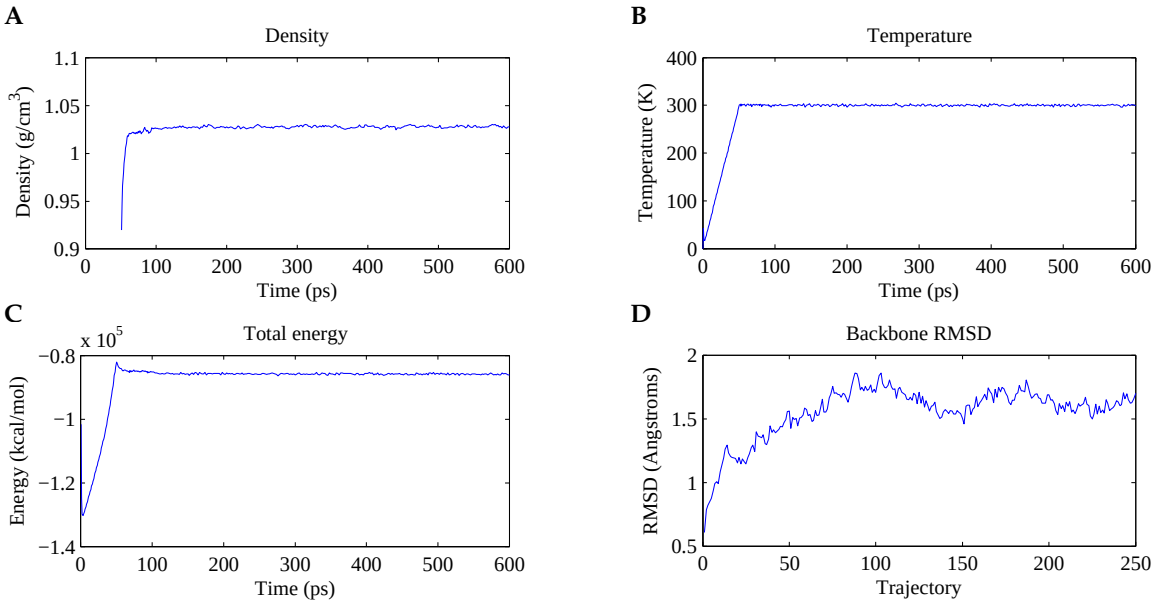

Fig. 4. The density (A), temperature (B), total energy (C) and backbone RMSD (D) curves of the system involving mutant delL747_K754insSR and gefitinib.

and variance of the connectivity of the binding-site surface atoms to represent the mean connectivity and its deviation. These two features are defined as the *mean connectivity* and the *connectivity variance*.

Atom index

Atom index is defined as the proportion of surface atoms in all the binding-site atoms. It can reflect the influence of a local mutation to the surface of a drug-binding pocket.

Binding free energy

We used binding free energy to measure the binding affinity of an EGFR mutant and gefitinib. Binding free energy is composed of Van der Waals force (VDWAALS), electrostatic energy (EEL), the electrostatic contributions (EGB) and nonpolar (ESURF) terms. Accordingly, these energy features were defined as *VDWAALS*, *EEL*, *EGB* and *ESURF*. To normalize these features, we subtracted the corresponding

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2568184, IEEE/ACM Transactions on Computational Biology and Bioinformatics

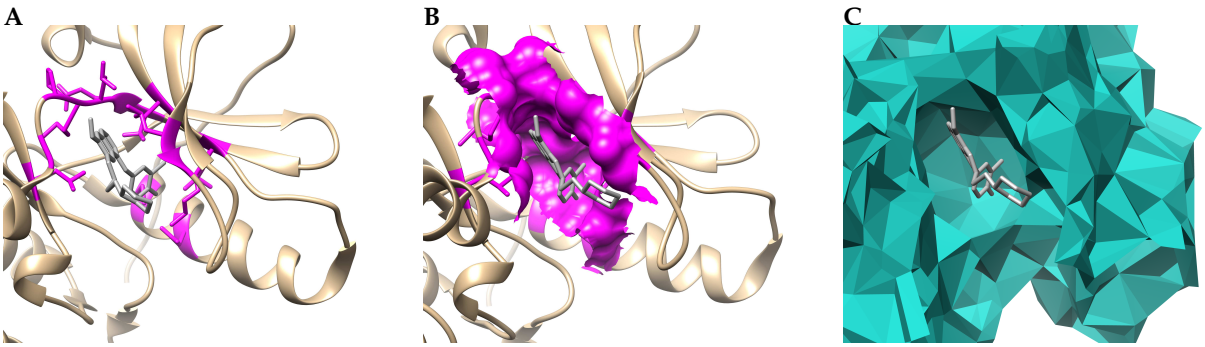

Fig. 5. The drug-binding site of WT EGFR, with gefitinib shown (dark gray). (A) and (B) show the amino acid residues and solvent-excluded molecular surface of the binding site, colored magenta. (C) The drug-binding pocket represented by an alpha shape.

energy terms of the WT EGFR-drug complex from them.

Overall, we extracted 14 features to characterize each mutant, and the distributions of these features are shown in Fig. 6A. In order to compensate the value differences, we normalized each feature into the range [-1, 1] (Fig. 6B).

The biological importance of these features can be demonstrated by a case study. L858R is a common mutant responding to gefitinib in NSCLC patients while the emergence of a second mutation T790M (L858R_T790M) causes drug resistance. The drug response level (*RL*) of the two mutants are 2 (L858R, *partial* response) and 3 (L858R_T790M, *stable disease*) respectively, obtained by calculating the median value of the patients sharing the same EGFR mutation type. Table 1 lists all the 14 features of the two mutants. *Curvature indices* $1 \sim 7$ represent the number of convex atoms with solid angle value belonging to [0.5, 1], [0.61, 1] and [0.71, 1], the average and variance of convex atoms, and the average and variance of concave atoms. For the mutant L858R_T790M, the number of atoms with solid angle value in [0.5, 1], [0.61, 1] and [0.71, 1] are more than that of L858R. In addition, the average solid angle values of convex and concave atoms are larger, even with a greater convex variance. These curvature features show the geometric changes (a higher knob level and a lower cleft level) of the drug-binding pocket of the mutant L858R_T790M compared with L858R, accounting for the drug resistance of the mutant L858R_T790M. Although the *connectivity* and *atom index* cannot demonstrate a direct relationship with the drug response level, they partly reflect the surface changes of a drug-binding pocket. The energy terms (before normalizing using the corresponding terms of the WT EGFR-drug complex) *VDWAALS*, *EEL* and *ESURF* of L858R-gefitinib are all lower than that of L858R_T790M-gefitinib complex. Even with a high *EGB*, still, there is a lower total binding free energy (summation of the four energy terms) for L858R-gefitinib, indicating a tighter binding status for L858R-gefitinib than for the L858R_T790M-gefitinib complex.

TABLE 1
COMPARISON OF THE FEATURES FOR THE MUTANTS L858R AND L858R_T790M.

Feature	L858R (*RL*=2)	L858R_T790M(*RL*=3)
Curvature index 1	6	9
Curvature index 2	4	8
Curvature index 3	4	6
Curvature index 4	0.4700	0.5252
Curvature index 5	0.0876	0.0941
Curvature index 6	-0.6128	-0.5639
Curvature index 7	0.0754	0.0728
Mean connectivity	6.1148	6.4576
Connectivity variance	4.5934	4.2821
Atom index	0.5980	0.5728
VDWAALS	-51.6461	-47.1378
EEL	-26.1769	-9.8236
EGB	38.4677	29.7224
ESURF	-6.6548	-5.9291

3.4 Transformation of a drug-binding site and analysis of drug resistance based on an eigen-binding site

In the treatments of NSCLC patients, the potency of drug can be evaluated by PFS or response level to the drug. Response levels (*RL*) include *complete response*, *partial response*, *stable disease* and *progressive disease*, mapping to four degrees of 1, 2, 3 and 4. Since a group of patients may have the same mutation in their EGFR TK domains, we took the median value in this group to represent the response level of this mutant. For all of our 62 mutation types, the corresponding normalized binding-site features were examined.

Before analyzing drug resistance using the extracted 14 features, we evaluated the performance of each feature via leave-one-out cross validation based on an SVM classification model. SVMs are supervised learning algorithms with several distinct advantages. The kernel strategy used in SVMs solves non-linear classification problems and provides flexibility to choose threshold. A unique solution can be obtained as the optimization problem is a convex one. The classification algorithm can be an out-of-sample generalization by choosing appropriate parameters C and r [57]. LIBSVM 3.18 [58] was used in this procedure. For simplicity, we combined the groups of $RL = 1$ and $RL = 2$ to form a larger group of *Response* group, and similarly build a *No-response* group containing those of $RL = 3$ and $RL = 4$. The performance of the features is shown in Table 2. All the features contribute to the drug response level prediction, especially the curvature indices. Specifically, curvature index 3 performs well in both the two-group and four response

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2568184, IEEE/ACM Transactions on Computational Biology and Bioinformatics

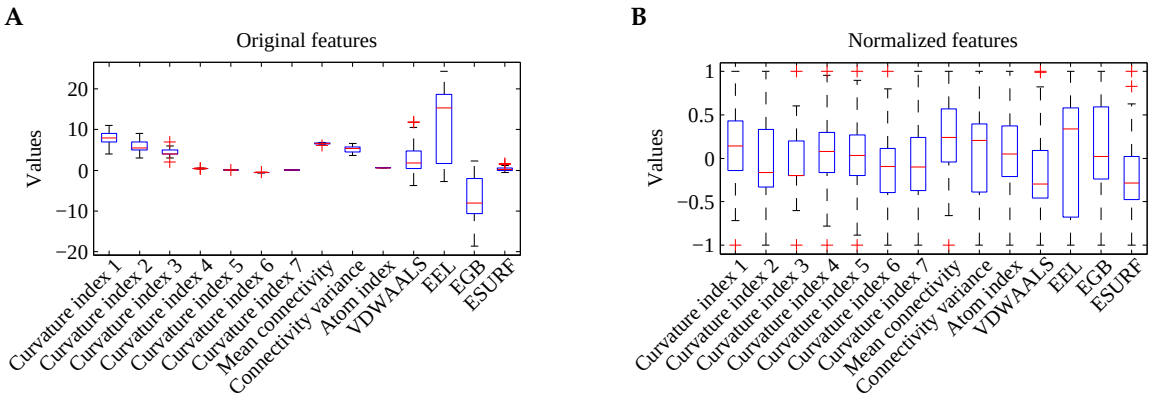

Fig. 6. Distributions of (A) the original features and (B) the normalized features.

levels situations with the highest prediction accuracies.

TABLE 2
PERFORMANCE OF EACH OF THE 14 FEATURES IN DRUG RESPONSE LEVEL PREDICTION.

Feature	Accuracy (four drug resistance levels)	Accuracy (two levels of *Response* or *No-response*)
Curvature index 1	62.90%	72.58%
Curvature index 2	62.90%	77.42%
Curvature index 3	69.35%	80.65%
Curvature index 4	62.90%	69.35%
Curvature index 5	62.90%	67.74%
Curvature index 6	64.52%	70.97%
Curvature index 7	62.90%	66.13%
Mean connectivity	62.90%	66.13%
Connectivity variance	62.90%	67.74%
Atom index	62.90%	66.13%
VDWAALS	62.90%	66.13%
EEL	62.90%	70.97%
EGB	62.90%	66.13%
ESURF	62.90%	66.13%

Different from other studies in which original features are used to represent drug the binding site directly, we employed PCA to transform the original binding site to an eigen-binding site, based on the extracted features. This eigen-binding site space is formed by the first k eigenvectors of the mutation-feature covariance matrix (*Method/Principal component analysis (PCA)* Section). Each new derived feature is a projection of the original features to the eigenvector direction of the corresponding eigenvalue of the covariance matrix. Considering two largest eigenvalues and the corresponding eigenvectors, in our work, λ_1 = 0.81, λ_2 = 0.68, $\boldsymbol{\mu_1}^T$ = [0.34, 0.47, 0.49, 0.29, 0.27, -0.03, -0.18, 0.04, 0.01, 0.15, 0.08, 0.34, -0.28, 0.10], $\boldsymbol{\mu_2}^T$ = [0.17, 0.22, 0.24, 0.08, 0.17, 0.05, -0.10, 0.07, 0.04, 0.20, -0.25, -0.63, 0.48, -0.27], then the first two new features can be obtained with the following equations, where f_i is the ith original feature.

$$\begin{aligned} PC_1 = {} & 0.34f_1 + 0.47f_2 + 0.49f_3 + 0.29f_4 + 0.27f_5 \\ & - 0.03f_6 - 0.18f_7 + 0.04f_8 + 0.01f_9 + 0.15f_{10} \\ & + 0.08f_{11} + 0.34f_{12} - 0.28f_{13} + 0.10f_{14} \end{aligned} \quad (10)$$

$$\begin{aligned} PC_2 = {} & 0.17f_1 + 0.22f_2 + 0.24f_3 + 0.08f_4 + 0.17f_5 \\ & + 0.05f_6 - 0.10f_7 + 0.07f_8 + 0.04f_9 + 0.20f_{10} \\ & - 0.25f_{11} - 0.63f_{12} + 0.48f_{13} - 0.27f_{14} \end{aligned} \quad (11)$$

The new feature PC_1 is the most important one among all the new derived features, as it has the largest variance (corresponding to λ_1) for all the samples. Similarly, PC_2 is the second most important component with the variance equal to λ_2. Fig. 7A shows the distribution of the samples in the PC_1-PC_2 space. Apparently, the centroids of the four response-level groups are separated. Moreover, the mutants corresponding to complete (RL = 1) or partial (RL = 2) response are wide apart, compared to those corresponding to stable (RL = 3) or progressive (RL = 4) disease. In addition, the importance of the original features can be obtained from the new derived ones, according to the significance of the new features and the contributions of the original ones. For example, PC_1 is the most important new feature. From (10), we can find that curvature indices f_3 and f_2 have the largest and second largest contributions to PC_1. Similarly, energy terms f_{12} and f_{13} make the largest and second largest contributions to PC_2, according to (11). Thus, Fig. 7A corresponds to a curvature-energy space, in which samples of different drug response levels are separated. This is consistent with our knowledge that surface and energy properties are key factors in molecular interactions. The eigen-binding site based framework proposed here provides a systematic method to study these properties. We further explored the distribution of the features projected to the first three PCs. For simplicity, we only show the 3D features and labels of the two groups (*Response* and *No-response* groups) of mutants in Fig. 7B. These two groups are then fitted to two surfaces respectively, using the gradient information (Figs. 7C and 7D). As shown in the figures, the feature trends of the two groups of mutants are apparently different.

With the features of each mutant projected to the first several PCs, we employed SVM to build a classification model to predict the drug response condition of each mutant. The classifier was built for 62 times, each time with 61

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2568184, IEEE/ACM Transactions on Computational Biology and Bioinformatics

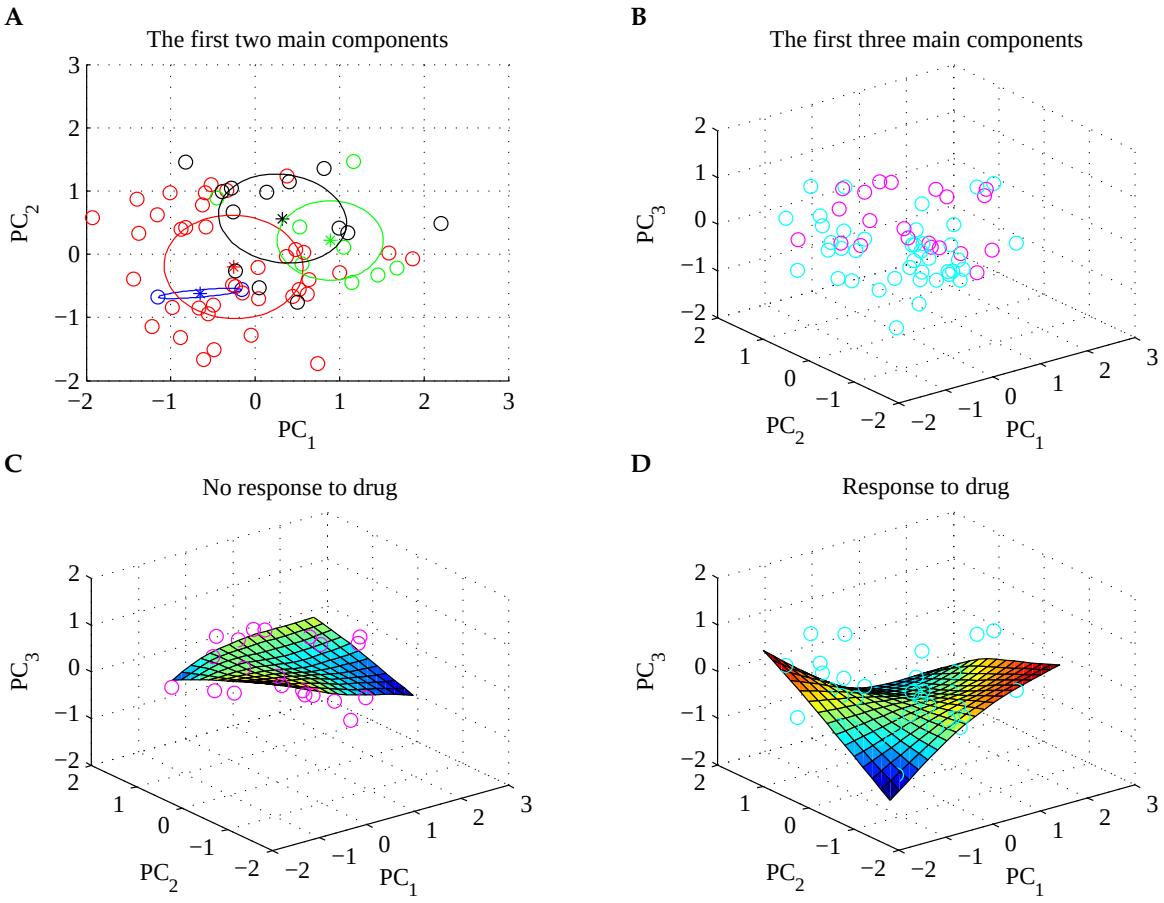

Fig. 7. Distributions of mutant samples described with the first several (k) PCs. (A) Projection of the mutant features to the first two PCs (k = 2). Here blue, red, green and dark circles represent mutant groups that correspond to drug response levels from 1 to 4 respectively, and '*' stands for the centroid of each group. (B) Mutant features projected to the first three PCs, with the data set condensed to Response (blue circles) and No-response (magenta circles) groups. (C) and (D) show the feature trends of the two groups of mutants.

TABLE 3
THE RESULTS OF SVM CLASSIFICATION.

Contribution rate threshold	0.91	0.93	0.95	0.97	0.99
The number of PCs	8	8	9	10	11
Accuracy (four drug resistance levels)	64.52%	64.52%	66.13%	69.35%	69.35%
Accuracy (two levels of *Response* or *No-response*)	79.03%	79.03%	80.65%	80.65%	85.48%

samples selected for training and one for testing. The results are now presented in Table 3. Contribution rate represents the proportion of the selected eigenvalues in the sum of all the eigenvalues. In the two-group situation (*Response* or *No-response*), the best accuracy (85.48%) achieves at contribution rate = 0.99 and PC numbers = 11. For a scenario where four resistance levels were considered (RL = 1, 2, 3 or 4), the highest accuracy is 69.35%. The response level to drug is not only be related to the geometric properties of drug-binding pocket of EGFR TK domain and the TK-drug binding affinity, but also be affected by personal features of each patient, such as smoke history, age and gender. Moreover, more patient data can be collected in future studies, to further refine our model.

4 Conclusion

In this study, we have proposed an eigen-binding site based method to analyze the EGFR mutation-induced drug resistance. Instead of using all the original features of the drug-binding site, we project these original features to an eigen-binding site space to obtain new features. Then we built a classification model based on SVM to correlate the new derived features with the response level to gefitinib for NSCLC patients.

With a group of EGFR mutants generated by Rosetta, we employed alpha shape modeling to extract the geometric properties of the drug-binding site for each mutant. Amber was used to implement MD simulations for a mutant-drug complex, and the binding free energy was calculated using the MM-GBSA protocol. The derived local surface geometric properties of a mutant, coupled with the mutant-drug binding free energy, were used as major features to characterize

a mutant. We built an eigen-binding site space using PCA, after which the original features were projected to this space. When the first two or three PCs were considered, the mutants with different response levels were separated. Finally, SVM was employed to build a classification model and to predict whether a mutant responds or not to gefinib. The best accuracy of the leave-one-out cross-validation mechanism reaches 85.48%.

As one of our future works, more clinical data will be collected to refine our model. When sufficient patients have the same mutation type, relatively-accurate drug response levels of the corresponding mutants can be obtained. In addition, more features of the drug-binding sites can be extracted to describe each binding site in more detail. These further studies will benefit cancer drug discovery and personalized therapy design.

ACKNOWLEDGMENT

This work is supported by the Health and Medical Research Fund (HMRF) of Hong Kong (Project 01121986) and the Hong Kong Research Grants Council (Project CityU 11200715).

REFERENCES

[1] M. G. Krebs, R. Sloane, L. Priest, L. Lancashire, J. M. Hou, A. Greystoke, T. H. Ward, R. Ferraldeschi, A. Hughes, G. Clack, *et al.*, "Evaluation and prognostic significance of circulating tumor cells in patients with non–small-cell lung cancer," *J. Clin. Oncol.*, vol. 29, no. 12, pp. 1556–1563, 2011.

[2] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2015," *CA: Cancer J. Clin.*, vol. 65, no. 1, pp. 5–29, 2015.

[3] S. V. Sharma, D. W. Bell, J. Settleman, and D. A. Haber, "Epidermal growth factor receptor mutations in lung cancer," *Nat. Rev. Cancer*, vol. 7, no. 3, pp. 169–181, 2007.

[4] R. S. Herbst, D. Prager, R. Hermann, L. Fehrenbacher, B. E. Johnson, A. Sandler, M. G. Kris, H. T. Tran, P. Klein, X. Li, *et al.*, "TRIBUTE: A phase III trial of erlotinib hydrochloride (OSI-774) combined with carboplatin and paclitaxel chemotherapy in advanced non–small-cell lung cancer," *J. Clin. Oncol.*, vol. 23, no. 25, pp. 5892–5899, 2005.

[5] J. Bar and A. Onn, "Overcoming molecular mechanisms of resistance to first-generation epidermal growth factor receptor tyrosine kinase inhibitors," *Clin. Lung Cancer*, vol. 13, no. 4, pp. 267–279, 2012.

[6] M. Soda, Y. L. Choi, M. Enomoto, S. Takada, Y. Yamashita, S. Ishikawa, S.-i. Fujiwara, H. Watanabe, K. Kurashina, H. Hatanaka, *et al.*, "Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer," *Nat.*, vol. 448, no. 7153, pp. 561–566, 2007.

[7] R. Govindan, L. Ding, M. Griffith, J. Subramanian, N. D. Dees, K. L. Kanchi, C. A. Maher, R. Fulton, L. Fulton, J. Wallis, *et al.*, "Genomic landscape of non-small cell lung cancer in smokers and never-smokers," *Cell*, vol. 150, no. 6, pp. 1121–1134, 2012.

[8] M. G. Kris, R. B. Natale, R. S. Herbst, T. J. Lynch Jr, D. Prager, C. P. Belani, J. H. Schiller, K. Kelly, H. Spiridonidis, A. Sandler, *et al.*, "Efficacy of gefitinib, an inhibitor of the epidermal growth factor receptor tyrosine kinase, in symptomatic patients with non–small cell lung cancer: a randomized trial," *Jama*, vol. 290, no. 16, pp. 2149–2158, 2003.

[9] F. Cappuzzo, F. R. Hirsch, E. Rossi, S. Bartolini, G. L. Ceresoli, L. Bemis, J. Haney, S. Witta, K. Danenberg, I. Domenichini, *et al.*, "Epidermal growth factor receptor gene and protein and gefitinib sensitivity in non–small-cell lung cancer," *J. Natl. Cancer Inst.*, vol. 97, no. 9, pp. 643–655, 2005.

[10] V. D. Cataldo, D. L. Gibbons, R. Pérez-Soler, and A. Quintás-Cardama, "Treatment of non–small-cell lung cancer with erlotinib or gefitinib," *New Engl. J. Med.*, vol. 364, no. 10, pp. 947–955, 2011.

[11] R. Sordella, D. W. Bell, D. A. Haber, and J. Settleman, "Gefitinib-sensitizing EGFR mutations in lung cancer activate anti-apoptotic pathways," *Sci.*, vol. 305, no. 5687, pp. 1163–1167, 2004.

[12] W. Pao, V. A. Miller, K. A. Politi, G. Ricly, R. Somwar, M. F. Zakowski, M. G. Kris, and H. Varmus, "Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain," *PLoS Med.*, vol. 2, no. 3, p. 225, 2005.

[13] L. V. Sequist, B. A. Waltman, D. Dias-Santagata, S. Digumarthy, A. B. Turke, P. Fidias, K. Bergethon, A. T. Shaw, S. Gettinger, A. K. Cosper, *et al.*, "Genotypic and histological evolution of lung cancers acquiring resistance to EGFR inhibitors," *Sci. Transl. Med.*, vol. 3, no. 75, pp. 75ra26–75ra26, 2011.

[14] A. Y. Helena, M. E. Arcila, N. Rekhtman, C. S. Sima, M. F. Zakowski, W. Pao, M. G. Kris, V. A. Miller, M. Ladanyi, and G. J. Riely, "Analysis of tumor specimens at the time of acquired resistance to EGFR-TKI therapy in 155 patients with EGFR-mutant lung cancers," *Clin. Cancer Res.*, vol. 19, no. 8, pp. 2240–2247, 2013.

[15] G. Hao, G. Yang, and C. Zhan, "Structure-based methods for predicting target mutation-induced drug resistance and rational drug design to overcome the problem," *Drug Discov. Today*, vol. 17, no. 19, pp. 1121–1126, 2012.

[16] S. Kobayashi, T. J. Boggon, T. Dayaram, P. A. Jänne, O. Kocher, M. Meyerson, B. E. Johnson, M. J. Eck, D. G. Tenen, and B. Halmos, "Egfr mutation and resistance of non–small-cell lung cancer to gefitinib," *New Engl. J. Med.*, vol. 352, no. 8, pp. 786–792, 2005.

[17] L. Regales, Y. Gong, R. Shen, E. de Stanchina, I. Vivanco, A. Goel, J. A. Koutcher, M. Spassova, O. Ouerfelli, I. K. Mellinghoff, *et al.*, "Dual targeting of EGFR can overcome a major drug resistance mutation in mouse models of EGFR mutant lung cancer," *J. Clin. Investing.*, vol. 119, no. 10, p. 3000, 2009.

[18] E. L. Kwak, R. Sordella, D. W. Bell, N. Godin-Heymann, R. A. Okimoto, B. W. Brannigan, P. L. Harris, D. R. Driscoll, P. Fidias, T. J. Lynch, *et al.*, "Irreversible inhibitors of the EGF receptor may circumvent acquired resistance to gefitinib," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, no. 21, pp. 7665–7670, 2005.

[19] J. A. Engelman, K. Zejnullahu, T. Mitsudomi, Y. Song, C. Hyland, J. O. Park, N. Lindeman, C. M. Gale, X. Zhao, J. Christensen, *et al.*, "MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling," *Sci.*, vol. 316, no. 5827, pp. 1039–1043, 2007.

[20] C. H. Yun, K. E. Mengwasser, A. V. Toms, M. S. Woo, H. Greulich, K. K. Wong, M. Meyerson, and M. J. Eck, "The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP," *Proc. Natl. Acad. Sci.*, vol. 105, no. 6, pp. 2070–2075, 2008.

[21] Z. P. Liu, L. Y. Wu, Y. Wang, X. S. Zhang, and L. Chen, "Bridging protein local structures and protein functions," *Amino Acids*, vol. 35, no. 3, pp. 627–650, 2008.

[22] D. D. Wang, W. Zhou, H. Yan, M. Wong, and V. Lee, "Personalized prediction of EGFR mutation-induced drug resistance in lung cancer," *Sci. Rep.*, vol. 3, 2013.

[23] L. Ma, D. D. Wang, Y. Huang, M. P. Wong, V. H. Lee, and H. Yan, "Decoding the EGFR mutation-induced drug resistance in lung cancer treatment by local surface geometric properties," *Comput. Boil. Med.*, 2014.

[24] A. Leaver-Fay, M. Tyka, S. M. Lewis, O. F. Lange, J. Thompson, R. Jacak, K. Kaufman, P. D. Renfrew, C. A. Smith, W. Sheffler, *et al.*, "ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules," *Methods Enzymol*, vol. 487, p. 545, 2011.

[25] H. Edelsbrunner and E. P. Mücke, "Three-dimensional alpha shapes," *ACM Trans. Gr. (TOG)*, vol. 13, no. 1, pp. 43–72, 1994.

[26] H. Edelsbrunner, *Weighted alpha shapes*. University of Illinois at Urbana-Champaign, Department of Computer Science, 1992.

[27] D. A. Case and et al., "AMBER 12." University of California, San Francisco, 2012.

[28] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *J. Educ. Psychol.*, vol. 24, no. 6, p. 417, 1933.

[29] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. learning*, vol. 20, no. 3, pp. 273–297, 1995.

[30] "The Protein Data Bank." http://www.rcsb.org.

[31] E. H. Kellogg, A. Leaver-Fay, and D. Baker, "Role of conformational sampling in computing mutation-induced changes in protein structure and stability," *Proteins: Struct. Funct. Bioinform.*, vol. 79, no. 3, pp. 830–838, 2011.

[32] M. A. Martí-Renom, A. C. Stuart, A. Fiser, R. Sánchez, F. Melo, and A. Šali, "Comparative protein structure modeling of genes and genomes," *Annu. Rev. Biophys. Biomo. Struct.*, vol. 29, no. 1, pp. 291–325, 2000.

[33] R. Das and D. Baker, "Macromolecular modeling with rosetta," *Annu. Rev. Biochem.*, vol. 77, pp. 363–382, 2008.
[34] J. D. Thompson, T. Gibson, D. G. Higgins, *et al.*, "Multiple sequence alignment using ClustalW and ClustalX," *Curr. Protoc. Bioinform.*, pp. 2–3, 2002.
[35] K. M. Misura, D. Chivian, C. A. Rohl, D. E. Kim, and D. Baker, "Physically realistic homology models built with ROSETTA can be more accurate than their templates," *Proc. Natl. Acad. Sci.*, vol. 103, no. 14, pp. 5361–5366, 2006.
[36] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "UCSF Chimeraa visualization system for exploratory research and analysis," *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, 2004.
[37] "Computational Geometry Algorithms Library." https://www.cgal.org/.
[38] W. Zhou, H. Yan, and Q. Hao, "Analysis of surface structures of hydrogen bonding in protein–ligand interactions using the alpha shape model," *Chem. Phys. Lett.*, vol. 545, pp. 125–131, 2012.
[39] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemom. Intel. Lab. Syst.*, vol. 2, no. 1, pp. 37–52, 1987.
[40] G. G. Maisuradze, A. Liwo, and H. A. Scheraga, "Principal component analysis for protein folding dynamics," *J. Mol. Boil.*, vol. 385, no. 1, pp. 312–329, 2009.
[41] Y. Bazi and F. Melgani, "Toward an optimal SVM classification system for hyperspectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3374–3385, 2006.
[42] H. Li, Y. Liang, Q. Xu, D. Cao, B. Tan, B. Deng, and C. Lin, "Recipe for uncovering predictive genes using support vector machines based on model population analysis," *IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB)*, vol. 8, no. 6, pp. 1633–1641, 2011.
[43] V. H. Lee, V. P. Tin, T. Choy, K. Lam, C. Choi, L. Chung, J. W. Tsang, P. P. Ho, D. K. Leung, E. S. Ma, *et al.*, "Association of exon 19 and 21 EGFR mutation patterns with treatment outcome after first-line tyrosine kinase inhibitor in metastatic non–small-cell lung cancer," *J. Thorac. Oncol.*, vol. 8, no. 9, pp. 1148–1155, 2013.
[44] "EGFR Mutation Database." http://www.cityofhope.org/egfr-mutation-database.
[45] D. Gu, W. A. Scaringe, K. Li, J. Saldivar, K. A. Hill, Z. Chen, K. D. Gonzalez, and S. S. Sommer, "Database of somatic mutations in EGFR with analyses revealing indel hotspots but no smoking-associated signature," *Hum. Mutat.*, vol. 28, no. 8, pp. 760–770, 2007.
[46] S. Chakrabarti and C. J. Lanczycki, "Analysis and prediction of functionally important sites in proteins," *Protein Sci.*, vol. 16, no. 1, pp. 4–13, 2007.
[47] F. Glaser, T. Pupko, I. Paz, R. E. Bell, D. Bechor-Shental, E. Martz, and N. Ben-Tal, "ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information," *Bioinform.*, vol. 19, no. 1, pp. 163–164, 2003.
[48] Y. Murakami and K. Mizuguchi, "Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites," *Bioinform.*, vol. 26, no. 15, pp. 1841–1848, 2010.
[49] Y. Ofran and B. Rost, "ISIS: interaction sites identified from sequence," *Bioinform.*, vol. 23, no. 2, pp. e13–e16, 2007.
[50] K. Kinoshita, Y. Murakami, and H. Nakamura, "eF-seek: prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape," *Nucleic Acids Res.*, vol. 35, no. suppl 2, pp. W398–W402, 2007.
[51] K. Kinoshita, J. Furui, and H. Nakamura, "Identification of protein functions from a molecular surface database, eF-site," *J. Struct. Funct. Genom.*, vol. 2, no. 1, pp. 9–22, 2002.
[52] A. Shulman-Peleg, R. Nussinov, and H. J. Wolfson, "Recognition of functional sites in protein structures," *J. Mol. Boil.*, vol. 339, no. 3, pp. 607–633, 2004.
[53] W. Zhou and H. Yan, "A discriminatory function for prediction of protein–DNA interactions based on alpha shape modeling," *Bioinform.*, vol. 26, no. 20, pp. 2541–2548, 2010.
[54] A. Porollo and J. Meller, "Prediction-based fingerprints of protein–protein interactions," *Proteins: Struct. Funct. Bioinform.*, vol. 66, no. 3, pp. 630–645, 2007.
[55] W. Zhou and H. Yan, "Prediction of DNA-binding protein based on statistical and geometric features and support vector machines," *Proteome Sci.*, vol. 9, no. 12, pp. 1–6, 2011.
[56] W. Zhou, H. Yan, X. Fan, and Q. Hao, "Prediction of protein-protein interactions using alpha shape modeling," in *2011 International Symposium on Computational Models for Life Sciences(CMLS11)*, vol. 1371, pp. 244–252, AIP Publishing, 2011.
[57] L. Auria and R. A. Moro, "Support vector machines (SVM) as a technique for solvency analysis," 2008.
[58] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol. (TIST)*, vol. 2, no. 3, p. 27, 2011.

Lichun Ma received the M.S. degree from Chinese Academy of Sciences, Beijing, in 2013. She is currently pursuing the Ph.D. degree in the Department of Electronic Engineering, City University of Hong Kong. Her research interests include bioinformatics, computational biology and drug resistance analysis.

Debby D. Wang received the Ph.D. degree in Electronic Engineering from City University of Hong Kong in 2014. Her research interests include bioinformatics, computational biology and machine learning.

Bin Zou received the M.S. degree from Wuhan University, Wuhan, in 2014. He is currently working toward the Ph.D. degree in the Department of Electronic Engineering, City University of Hong Kong. His research interests include computational biology and drug resistance.

Hong Yan received the Ph.D. degree in Electronic Engineering from Yale University, New haven, Connecticut, in 1989. He was professor of imaging science at the University of Sydney and is currently professor of computer engineering at City University of Hong Kong. His research interests include bioinformatics, image processing and pattern recognition. He is a fellow of the IEEE and IAPR.