Mutation Research 780 (2015) 24–34

Contents lists available at ScienceDirect

Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis

journal homepage: www.elsevier.com/locate/molmut
Community address: www.elsevier.com/locate/mutres

Discrimination of driver and passenger mutations in epidermal growth factor receptor in cancer

P. Anoosha [a], Liang-Tsung Huang [b], R. Sakthivel [a], D. Karunagaran [a], M. Michael Gromiha [a,*]

[a] Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai 600 036, Tamil Nadu, India
[b] Department of Medical Informatics, Tzu Chi University, Hualien 970, Taiwan

ARTICLE INFO

Article history:
Received 18 February 2015
Received in revised form 21 May 2015
Accepted 7 July 2015
Available online 20 July 2015

Keywords:
Driver mutation
Passenger mutation
EGFR
Machine learning

ABSTRACT

Cancer is one of the most life-threatening diseases and mutations in several genes are the vital cause in tumorigenesis. Protein kinases play essential roles in cancer progression and specifically, epidermal growth factor receptor (EGFR) is an important target for cancer therapy. In this work, we have developed a method to classify single amino acid polymorphisms (SAPs) in EGFR into disease-causing (driver) and neutral (passenger) mutations using both sequence and structure based features of the mutation site by machine learning approaches. We compiled a set of 222 features and selected a set of 21 properties utilizing feature selection methods, for maximizing the prediction performance. In a set of 540 mutants, we obtained an overall classification accuracy of 67.8% with 10 fold cross validation using support vector machines. Further, the mutations have been grouped into four sets based on secondary structure and accessible surface area, which enhanced the overall classification accuracy to 80.2%, 81.9%, 77.9% and 75.1% for helix, strand, coil-buried and coil-exposed mutants, respectively. The method was tested with a blind dataset of 60 mutations, which showed an average accuracy of 85.4%. These accuracy levels are superior to other methods available in the literature for EGFR mutants, with an increase of more than 30%. Moreover, we have screened all possible single amino acid polymorphisms (SAPs) in EGFR and suggested the probable driver and passenger mutations, which would help in the development of mutation specific drugs for cancer treatment.

1. Introduction

Epidermal Growth Factor Receptor (EGFR) is a receptor tyrosine kinase protein present on the cell surface, mainly expressed in epithelial cells. EGFR protein mainly consists of three domains such as an extracellular ligand-binding region, a single transmembrane helix, and an intra-cellular tyrosine kinase domain. It gets activated upon binding to its specific ligands such as EGF (epidermal growth factor), TGF-α (transforming growth factor-α) and induces down-stream signaling pathways that ultimately lead to cell differentiation, proliferation, migration or motility, adhesion, protection from apoptosis, enhanced survival and gene transcription [1,2]. When mutations assemble in a precancerous cell, some of them confer a selective growth advantage by contributing to tumorigenic functions (termed as 'drivers'), whereas others are competently neutral (termed as 'passengers'). Passenger mutations are mostly observed in mature cancer cells and are not known for any pathogenic characteristics [3]. EGFR is a well-known oncogene, its mutations and over expression are frequently observed in many of the cancer types. Several somatic mutations observed in cancer samples within the kinase domain of the EGFR gene (Exons 18–21) have been reported in the literature [4]. Hence, it has become an important target for cancer therapeutics [5]. Many studies reported that deleterious mutations are mostly observed in non-small cell lung cancer and are highly sensitive to EGFR-Tyrosine Kinase inhibitors such as erlotinib and gefitinib [6]. The continuous treatment with erlotinib and gefitinib lead to acquired resistance of the cancer patients due to the formation of new mutations during tumor development and progression [7]. Recently, Wang et al. reported the prediction of EGFR mutation-induced drug resistance in lung cancer, which is a major problem to be resolved in cancer treatment [8].

As the mutations in EGFR are increasing rapidly, it is very difficult to assess each mutation experimentally. Hence, it is necessary to develop computational methods for efficiently identifying deleterious mutations (drivers), which would help to design

* Corresponding author. Tel.: +91 2257 4138; fax: +91 2257 4102.
E-mail address: gromiha@iitm.ac.in (M.M. Gromiha).

http://dx.doi.org/10.1016/j.mrfmmm.2015.07.005
0027-5107/

experiments. For the past one decade, several computational tools have been developed for predicting the functional mutations in proteins, which are mainly based on sequence, structural and evolutionary features [9–16]. All these methods have been trained with a set of mutants from many target proteins which are not specific to EGFR, and hence the driver and passenger mutations have been distinguished with very low accuracy. In addition, other aspects such as (i) unbiased training sets with experimentally known annotation of mutations, and (ii) protein specific features to efficiently identify driver mutations in a protein of interest, should be considered for better classification accuracy.

In the current study, we mainly focused on EGFR point mutations which are more frequently observed in different cancer types. A dataset of 600 missense mutations comprising of 242 drivers and 298 passengers is collected from COSMIC database. We have analysed the distribution of residues along the protein sequence and the characteristic features of amino acid residues in both wild-type and mutant protein structures, and derived a set of 222 features (attributes). Utilizing feature selection methods, a set of 21 features has been identified, which could classify the driver and passenger mutations with an accuracy of 67.8% in 10-fold cross validation using support vector machines in a dataset of 540 mutants. Further, we have grouped the mutations based on their secondary structure and accessible surface area, which increased the classification accuracy to 80.2%, 81.8%, 77.9% and 75.1% for helix, strand, coil-buried and coil-exposed mutants, respectively. An independent test set of 60 mutations is used to assess our method, which showed the accuracy of 80.0%, 85.0%, 90.9% and 85.7%, respectively. Further, we have utilized our prediction method and identified the most probable driver and passenger mutations in EGFR. We suggest that these prediction results would serve as a useful tool in the development of mutation specific drugs for cancer therapy.

2. Materials and methods

2.1. Dataset

We derived a dataset of 600 EGFR somatic missense mutations, which comprises of 266 drivers and 334 passengers from COSMIC database [17] (version v71) for the present study. We have randomly chosen 540 mutations as a training set and 60 test set. The classification criteria of mutations as driver or passenger, is based on the information available in the literature [18–22], Swiss-Prot Variant database [23], dbSNP database [24] and recurrent mutations in cosmic database. Mutations occurring recurrently (count >1, i.e., mutations observed in more than one sample in cosmic database) are termed as drivers and observed only once in cancer samples are termed as passengers [11,25]. In addition, we have taken into account of treatment status for all the driver mutations. We found that only four mutants (T790M, L747S, D761Y and T854A) are secondary site mutations. However they are also observed prior to the treatment with tyrosine kinase inhibitors in rare cases [26–28].

2.2. Attribute construction

We calculated a wide array of attributes focusing at the mutation position using both amino acid sequence and three-dimensional structure of EGFR. Sequence based features include physico-chemical properties of amino acid residues [29], amino acid matrices and contact potentials from AAIndex database [30], and neighbouring residue information at different window lengths. Structure based features include accessible surface area [31], long range contacts [32], surrounding hydrophobicity [33], secondary structure element of amino acids [34] and number of hydrogen bonds [35]. We have obtained the 3D structure of the wild-type protein from the PDB, Protein Data Bank [36] and modelled the mutant protein using Swiss model [37].

2.3. Computation of sequence and structure based features

2.3.1. Physico-chemical properties

We have considered 49 properties belonging to physical, chemical, energetic and conformational parameters in this study [29]. The change in property between wild type and mutant residue is computed as:

$$\Delta P_{(\text{mutation})} = P_{(\text{Snp})} - P_{(\text{wild-type})} \quad (1)$$

where $P_{(\text{wild-type})}$ and $P_{(\text{Snp})}$ are the property values of wild type and mutant residues, respectively and $\Delta P_{(\text{mutation})}$ is the change in property due to mutation.

2.3.2. AAindex database matrices

Amino acid mutation matrices are collected from AAIndex2 database [30] and the value is substituted for each mutation. Pair-wise contact potential matrices are collected from AAIndex3 database [30] and difference of amino acid contact potential for a mutation is obtained by subtracting contact potential value of N-/C-neighbour of mutation position to wild type residue from N-/C-neighbour to mutant residue.

2.3.3. Neighbouring residue information, solvent accessibility and Conservation

We have considered the preference of neighbouring residues (N- and C-termini) of the mutation position in the sequence at window length of 3–13 residues (1–6 residues towards N- and C- termini) as additional features. The twenty amino acid residues are grouped into aliphatic, sulphur-containing, aromatic, polar, positively and negatively charged categories, based on their physico-chemical nature of their side chains and computed the number of residues in each group in the neighbouring regions of mutation site for different window lengths. Accessible surface area of the residues at mutation site was calculated using SABLE Program [38]. Evolutionary conservation of amino acid positions in protein is estimated using Consurf server [39], which is a bioinformatics tool based on the phylogenetic relations between homologous sequences.

2.3.4. Structure based features

We have computed long range order [40] and surrounding hydrophobicity [33] of residues at each mutant position using the procedure described in the literature. Secondary structures of amino acid residues were obtained using DSSP [41] and the hydrogen bonds were calculated using HBPLUS [42] program.

2.4. Attribute selection and classification

We have utilized various feature selection and classification methods available in Weka [43] for distinguishing between driver and passenger mutations. Based upon the performance of all the methods on our dataset, we have chosen the combination of SVM attribute evaluator and Ranker search method for feature selection, which was reported to be an efficient approach [44]. SMO (Sequential Minimal Optimization) algorithm, which is a SVM based method, has been employed for the classification of mutants into drivers and passengers. SVM is a widely used machine learning technique which provides the probability estimates for the predicted class. Using feature selection methods, we derived a set of 14, 18, 8 and 13 features for the mutations in helix, strand, coil-buried and coil-exposed mutants, respectively.

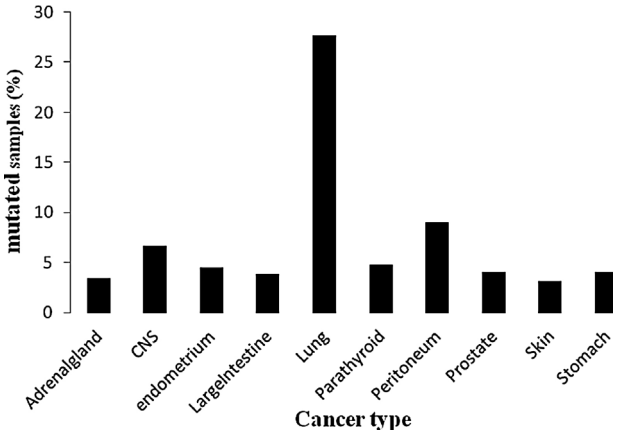

Fig. 1. Occurrence of EGFR mutations in topmost ten cancer types.

2.5. Performance evaluation

The model is evaluated using n-fold cross validation in which n-1 data are used to train the classifier and rest of them are used for testing. The classification performance of the model has been assessed by following measures:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (4)$$

where TP, TN, FP and FN refer to the number of true positives, true negatives, false positives and false negatives, respectively. Driver mutations are considered as positive class and passenger as negative class. In addition, we have estimated Area under the ROC (receiver operating characteristic) curve, which is a plot of true positive rate against false positive rate to estimate the trade-off between sensitivity and specificity at different thresholds.

3. Results and discussion

3.1. Occurrence of missense mutations in various cancer types

EGFR mutations consisting of insertions, deletions, missense and nonsense substitutions collected from cosmic database are observed in 30 different cancer types. Among them, more than 80% of mutations in cancer samples are missense substitutions. Hence, we have analysed the occurrence of missense substitutions in different cancer types and observed that lung cancer is dominated with 28% of the mutations followed by Peritoneum (9%), Central nervous system (6.7%), Parathyroid (4.8%) and Endometrium cancer (4.5%). Occurrence of missense mutations in topmost ten cancer types is given in Fig. 1. Considering the number of mutations in various cancer types, EGFR is at second position in lung cancer and among topmost five in the other cancer types. This suggests the fact that the number of mutations observed in EGFR is dominant in most of the cancer types when compared with other proteins. Hence, there is an immense need to carry out extensive analyses and develop prediction methods to deepen our understanding about the role of EGFR in cancer development and progression.

3.2. Distribution of missense mutations in EGFR structure

We have mapped all the mutations on the three-dimensional structure of EGFR and their locations are shown in Fig. 2. We

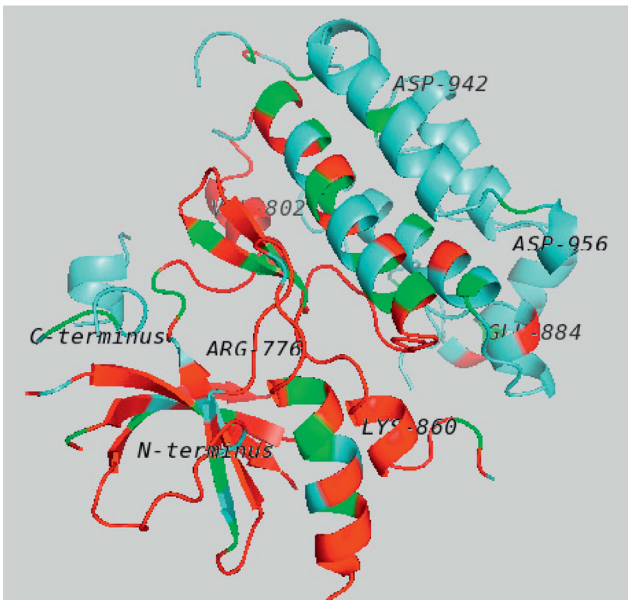

Fig. 2. Distribution of driver (red) and passenger (green) mutations in the intracellular domain of EGFR (N- and C- terminus, and few residues are labeled).

observed that majority of the mutations are accumulated in the intracellular region (669–1210 residues), which contains tyrosine kinase and cytoplasmic domains, known to have important role in activating downstream signalling and regulate its function [45]. We have studied the location of amino acid substitutions in EGFR and observed that 78% of the total mutations are present in the intracellular region of which 38% of them are drivers. We further analysed the distribution of mutations in different secondary structures such as helix, strand and coil which showed that driver mutations are enriched in strands and coils. We also observed that most of the glycine and valine substitutions present in strands and coils are drivers whereas, cysteine residue mutations are observed as passengers. Further, we have analysed the importance of these residues using the experimental data available in the literature and a typical example, G719S presented in a β-strand, is discussed below. We have retrieved the wild-type and G719S mutant structures of EGFR from Protein Data Bank, PDB [36] and superimposed them using PyMol. It is interesting to note that the single G to S substitution caused a root mean square deviation (RMSD) of 1.8 Å (Fig. 3) and destabilized the kinase domain by distorting a set of hydrophobic interactions observed in the wild type structure. In addition, G719S mutation activates the kinase by disrupting autoinhibitory interactions [46].

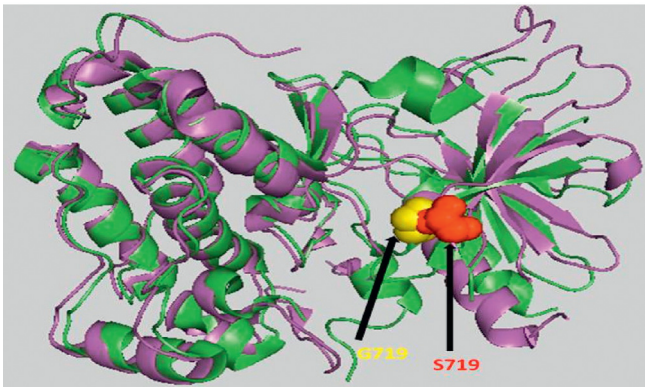

Fig. 3. Superposition of wild-type (green) and G719S mutant (magenta) structures of EGFR. The PDB codes are 2RGP and 2EB2, respectively and the RMSD is 1.8 Å.

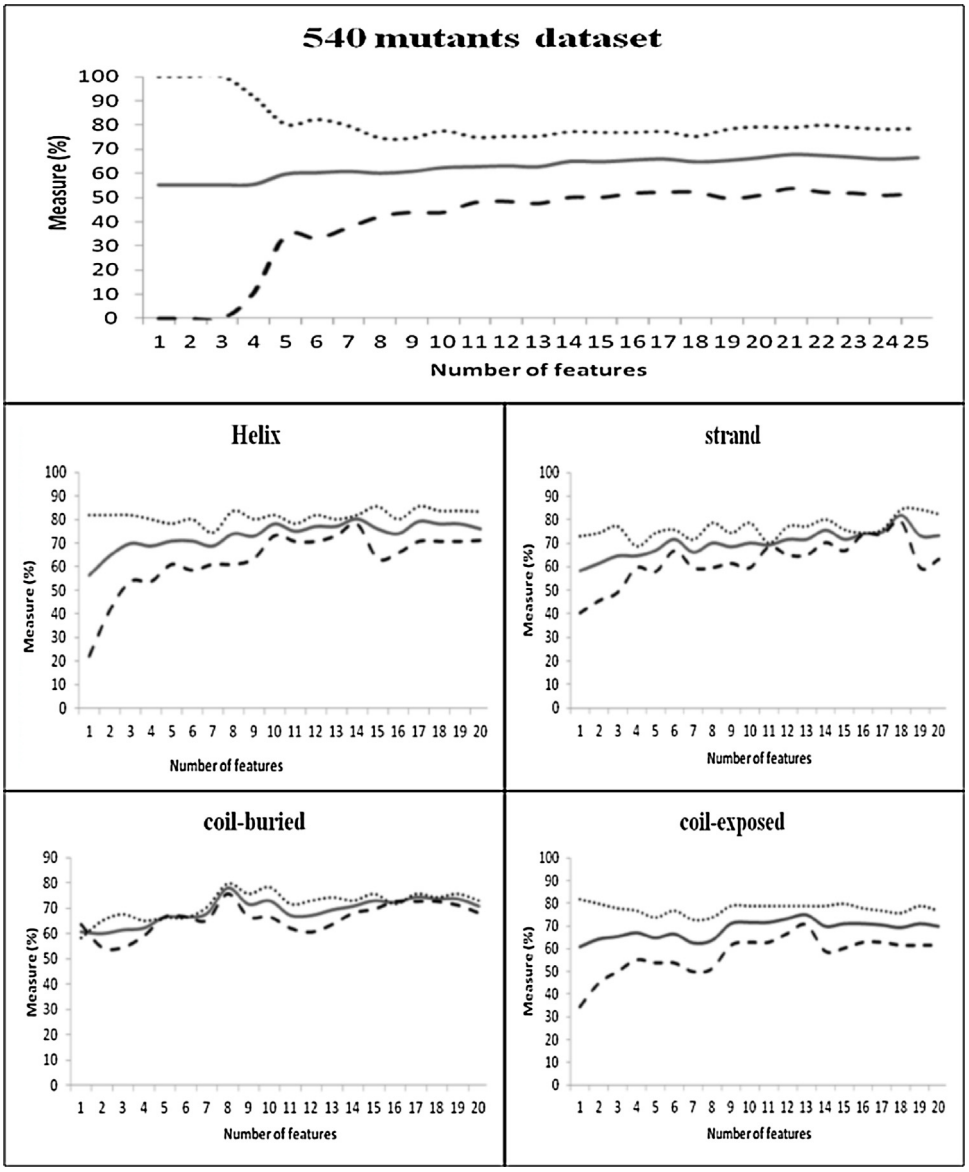

Fig. 4. Variation of accuracy, sensitivity and specificity for the classification of mutants in helix, strand, coil-buried and coil-exposed models. ▬▬ Accuracy; ▬ ▬ Sensitivity; ▪▪▪▪▪▪ Specificity

3.3. Combination of features

Firstly, we classified the data set of 540 mutations into drivers and passengers using 222 individual features by SMO classifier. The classification accuracy ranges between 55% and 60%, in which the statistical contact potential feature showed a maximum accuracy of 59.8% with 40.5% and 74.8% of sensitivity and specificity, respectively. These results suggest that the information of individual feature is not sufficient to discriminate driver and passenger mutations. Further, we combined the features using feature selection method and identified a set of 21 best features, which showed an accuracy, sensitivity and specificity of 67.8%, 53.7% and 79.2%, respectively in 10-fold cross validation. These selected features include amino acid substitution matrices and amino acid contact potentials derived from the sequence segment containing wild-type, mutant, N-terminal and C-terminal neighbouring residues. The accuracy, sensitivity and specificity of the classification obtained by the combination of features up to 25 are shown in Fig. 4. The maximum accuracy of 67.8% is obtained at 21 feature combination, and it decreased with the addition of new feature for classification. This analysis suggests the necessity of classifying mutants based on their location and physico-chemical behaviour, which would biologically play an important role in the discrimination of mutants.

3.4. Grouping the mutations based on secondary structure and solvent accessibility

Earlier studies showed that protein secondary structure and solvent accessibility play important roles to understand the folding, stability and function of protein as well as the diseases [47,48]. Hence, we grouped the 540 mutations into four classes of mutants in helix, strand, coil-buried and coil-exposed regions based on their secondary structure and accessible surface area and performed feature selection. Accuracy, sensitivity and specificity of different combinations of up to 20 features for all the models are shown in Fig. 4. The maximum accuracy is obtained for a set of 14, 18, 8 and 13 features for the helix, strand, coil-buried

Table 1
Performance of classification models on different datasets based on secondary structure and solvent accessibility.

Performance	Helix (11)	Strand (14)	Coil-buried (15)	Coil-exposed (20)
Accuracy (%)	80.2 (90.9)	81.9 (85.7)	77.9 (80.0)	75.1 (85.0)
Sensitivity (%)	78.0 (80.0)	78.9 (83.3)	75.8 (85.7)	70.5 (83.3)
Specificity (%)	81.8 (100)	84.3 (87.5)	79.7 (75.0)	78.8 (85.7)
No. of mutants	96 (11)	127 (14)	140 (15)	177 (20)

The results were obtained with 10-fold cross validation. The test set results are given in parenthesis.

Table 2
List of selected features for four different classification models.

Helix	Strand	Coil-buried	Coil-exposed
Snp residue	Snp residue	Polar	Positive
C + 1 residue	**Polar**	ALTS910101	Aliphatic
Polar	Sulphur-containing	CSEM940101	P
pK′	**pK′**	**MOOG990101_dN**	BENS940101
KOLA920101	ΔC_{ph}	**SKOJ000101_dN**	MIYT790101
LUTR910104	LUTR910102	SIMK990102_dN	RUSR970103
MCLA710101	**QU_C930103**	SIMK990104_dN	NGPC000101
NIEK910101	TUDE900101	KOLA930101_dC	VENM980101_dN
QU_C930102	RUSR970101		KESO980102_dN
QU_C930103	MUET010101		ZHAC000104_dN
RISJ880101	**MOOG990101_dN**		VENM980101_dC
KAPO950101	**SKOJ000101_dN**		BONM030103_dC
TANS760101_dN	SIMK990101_dN		**Conservation**
TANS760102_dC	THOP960101_dC		
	BONM030104_dC		
	ZHAC000104_dC		
	Conservation		

The common features are shown in bold; _dC and _dN are C- and N- terminal contact potentials, respectively. Snp, Mutant residue; C + 1 residue, C-terminal neighbour of mutant position; Conservation, conservation score of the mutation site; Polar, number of neighbouring polar residues within window length 13; Sulphur-contain, number of neighbouring sulphur containing residues within window length 13; Positive, number of neighbouring positively charged residues within window length 13; Aliphatic, number of neighbouring aliphatic residues within window length 13; P, Polarity; pK′, Equilibrium constant with reference to the ionization property of COOH group; ΔC_{ph}, Unfolding hydration heat capacity change; KOLA920101, Conformational similarity weight matrix; LUTR910104, Structure-based comparison table for inside α-class; MCLA710101, The similarity of pairs of amino acids; NIEK910101, Structure-derived correlation matrix; QU_C930102, Cross-correlation coefficients of spatial preferences of side chain; QU_C930103, The mutant distance based on spatial preference factor; RISJ880101, Distance scoring matrix; KAPO950101, extent of amino acid substitution and its correlation with variation in globin sequence volume; LUTR910102, Structure-based comparison table for coil/turn class; NIEK910102, Structure-derived correlation matrix; TUDE900101, Isomorphicity of replacements; RUSR970101, Substitution matrix based on structural alignments of analogous proteins; MUET010101, Non-symmetric substitution matrix for detection of homologous transmembrane proteins; ALTS910101, The PAM-120 matrix; CSEM940101, Residue replace ability matrix; BENS940101, Log-odds scoring matrix; MIYT790101, Amino acid pair distance; RUSR970103, Substitution matrix based on structural alignments of analogous and remote homologous proteins; NGPC000101, Substitution matrix built from hydrophobic and transmembrane regions of the Blocks database; TANS760101, Statistical contact potential derived from 25 X-ray protein structures; TANS760102, Number of contacts between side chains derived from 25 X-ray protein structures; MOOG990101, Quasichemical potential derived from interfacial regions of protein-protein complexes; SKOJ000101, Statistical quasichemical potential with the partially composition-corrected pair scale; SIMK990101, Distance-dependent statistical potential (contacts within 0–5 Å); SIMK990102, Distance-dependent statistical potential (contacts within 5–7.5 Å); SIMK990104, Distance-dependent statistical potential (contacts within 10–12 Å); VENM980101, Statistical potential derived by the maximization of the perceptron criterion; KESO980102, Quasichemical energy in an average protein environment derived from interfacial regions of protein–protein complexes; ZHAC000104, Environment-dependent residue contact energies; THOP960101, Mixed quasichemical and optimization-based protein contact potential; BONM030104, Distances between centres of interacting side chains in the anti-parallel orientation; KOLA930101, Statistical potential derived by the quasichemical approximation; BONM030103, Quasichemical statistical potential for the parallel orientation of interacting side groups;.

and coil-exposed mutants, respectively. Different classification models are developed for each category of mutants and the performance results are presented in Table 1. The selected features in different categories of mutants are given in Table 2. Interestingly, the features LUTR910104 (structure-based comparison table for α-class) and ZHAC000104 (matrix based on residue contact energies in strand) are specific to α-helices and β-strands, and these properties are selected as important features to discriminate mutations in helical and strand regions, respectively. These results showed that the information of secondary structure enhanced the prediction performance significantly on EGFR mutation dataset, especially in helical and strand structures. However, the improvement in coil-buried and coil-exposed mutants is less than 8% compared to the whole dataset, which agrees with the earlier report that the mutations within the coil structure are more difficult to predict than those within helix or strand structures [49]. We have also plotted the area under the ROC curve for all the classification models in Fig. 5. ROC analysis yielded the area under the curve (AUC) of 0.80, 0.84, 0.81 and 0.78 for the mutants in helix, strand, coil-buried and coil-exposed regions, respectively.

3.4.1. Mutations in helix

Totally, 96 mutations are observed in helix structures in which 41 and 55 are driver and passenger mutations, respectively. A set of 14 features is selected using feature selection method and we

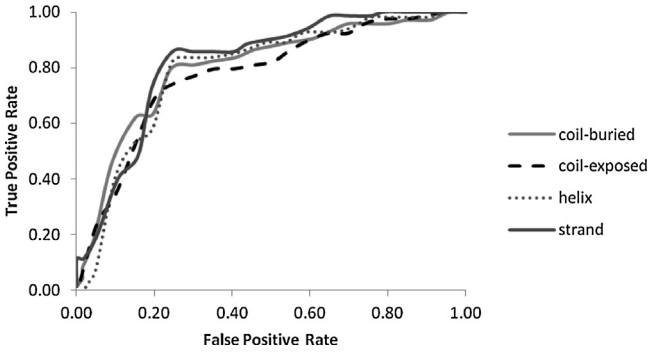

Fig. 5. ROC curve for the classification models for helix, strand, coil-buried and coil-exposed mutants.

Table 3a
Distribution of lung cancer mutations.

Mutation type	Helix	Strand	Coil-buried	Coil-exposed
Drivers	39 (85%)	55 (87%)	49 (67%)	56 (67%)
Passengers	31 (51%)	33 (42%)	36 (44%)	37 (33%)
Total	70 (65%)	88 (62%)	85 (55%)	93 (47%)

Proportion of lung cancer mutations in the dataset is given in parenthesis.

Table 3b
Performance of the model on lung cancer mutations.

Performance	Helix	Strand	Coil-buried	Coil-exposed
Accuracy (%)	84.3	81.8	76.5	76.3
Sensitivity (%)	82.1	80.0	77.6	73.2
Specificity (%)	87.1	84.9	75.0	81.1

achieved a classification accuracy of 80.2% in 10-fold cross validation with the sensitivity and specificity of 78.0% and 81.8%, respectively. The selected features include neighbouring residue information from sequence, protein structure based mutation matrices and statistical contact potential matrices. Driver mutations are dominated with valine and leucine substitutions whereas, passengers with lysine and isoleucine substitutions in helix region.

3.4.2. Mutations in strand

Strand contains 127 mutations in which 57 are drivers and 70 are passengers. We have identified a set of 17 best features using feature selection method and achieved a classification accuracy of 81.9% in 10-fold cross validation with sensitivity and specificity of 78.9% and 84.3%, respectively. The training set is dominated with glycine and leucine amino acid substitutions as drivers and passengers, respectively. The selected features are mutant residue, polar and sulphur containing amino acids in the neighbouring residues of mutation site within a window length of 13 in the sequence, structure based mutation matrices and evolutionary conservation of the mutation position.

3.4.3. Mutations in coil

Among 540 mutations, 317 are observed in coil, which comprises of majority of mutations in EGFR. The classification accuracy of these mutants is below 70%. Hence, the mutants are further grouped into two classes based on ASA as the coil structures are highly flexible and complex. The mutated residues with ASA $\leq$30% and >30% are grouped separately into coil-buried and coil-exposed, respectively. Using feature selection method, 8 and 13 best features are selected in these classes of mutants, which include neighbouring residue information from sequence, mutation matrices and contact potentials. The classification accuracy of the mutants in these two groups is 77.9% and 75.1% in 10-fold cross validation with the sensitivity and specificity of 75.8%, 79.7% and 70.5%, 78.8%, respectively.

3.5. Performance of the model on blind set

Further, we have evaluated all four classification models using a test set of 60 mutations and the results are presented in Table 1. Ten percent of the original (60 mutants) are chosen randomly for evaluating the performance of the method. Mutations in the helix region are classified with the highest accuracy of 90.9%, with sensitivity and specificity of 80.0% and 100%, respectively. The accuracy of strand, coil-buried and coil-exposed mutations are 85.7%, 80.0% and 85.0%, respectively and the average accuracy is 85.4%. The results suggest that our method could be used as an efficient tool to predict driver mutations in EGFR protein and to understand the deleterious nature of the mutants in causing cancer. Hence, we utilized the method for identifying the most prominent drivers and passengers among all possible 22401 mutations in EGFR.

3.6. Performance of the model on lung cancer mutations

We observed that majority of the mutations (55%) in our dataset belong to lung cancer samples. The distribution of them in different secondary structures is shown in Table 3a and the percentage of mutants in helical, strand, coil-buried and coil-exposed are 65%, 52%, 55% and 47%, respectively. Further, we have evaluated the performance of the present model on these mutations and the results are presented in Table 3b. The current method discriminated the driver and passenger mutations in lung cancer with an accuracy of 84.3%, 81.8%, 76.5% and 76.3%, respectively in helix, strand, coil-buried and coil-exposed regions. These results are comparable to those obtained with the whole data set (Table 1), which suggests that the current method could distinguish between driver and passenger mutations on one of the most common forms of cancer in EGFR.

3.7. Importance of the selected features

We have analysed the importance of selected features by removing each feature from the respective models and calculated the accuracy, sensitivity and specificity. The results are presented in Table 4. For mutations located in helix region, the elimination of three among fourteen features (LUTR910104, QU_C930102 and TANS760102_dC) reduced the classification accuracy by 8%. Interestingly, LUTR910104 (structure-based comparison table for α-class) is specific to α-helix and other two features are based on the preferred contacts between side chains. The mutant residue is selected as an important feature to discriminate driver and passenger mutations and this shows the preference of specific residues to act as drivers compared to the rest. Further, we have verified the preference of various amino acids to be drivers/passengers for "helix" set and observed that Arg, Lys and Thr prefer to be snp residue in driver mutations and Val, Gly and Asn in passenger mutations. We also observed that neighbouring residues, especially adjacent C-terminal neighbour (C + 1) and number of polar residues within a window length of 13 residues in the sequence, which would be important in maintaining the stability of α-helix as discussed in the earlier reports [50] might influence the cancer causing effect of mutations occurring in this region of EGFR.

For the mutations in strand model, the number of sulphur containing amino acids in neighbouring residues within a window length of 13 residues along the sequence and ZHAC000104 (matrix based on residue contact energies in strand) are selected as important features, suggesting the importance of disulphide bridges (cysteine) and hydrophobic nature (methionine) in maintaining proper structural orientation of the native structure of EGFR protein. These feature when removed from the model reduced the classification accuracy and sensitivity by 12% and 12% to 14%, respectively, showing its importance in predicting the actual driver mutations. We propose that mutations in strand might cause conformational changes in adjacent regions there by disrupting disulphide bonds and other interactions that play important roles in maintaining the normal function of the protein.

For discriminating the mutations in coil region into drivers and passengers, the nature of neighbouring residues as well as contact potentials play a vital role. However, classification of coil mutations based on the solvent accessibility (buried and exposed) revealed few important properties that are specific for the two categories. We observed that the number of polar amino acids in neighbouring residues within a window length of 13 residues in sequence is a key feature to differentiate drivers from passengers for coil mutations in the buried regions. This might be attributed with the fact the

Table 4
Importance of selected features in individual models.

Feature	Accuracy (%)	Sensitivity (%)	Specificity (%)
Helix			
Snp	77.1	73.2	80.0
C-terminal residue	77.1	73.2	80.0
polar	74.0	65.9	80.0
pK	76.0	70.7	80.0
KOLA920101	78.1	73.2	81.1
LUTR910104	71.9	65.9	76.4
MCLA710101	75.0	68.3	80.0
NIEK910101	79.2	75.6	81.8
QU_C930102	71.9	61.0	80.0
QU_C930103	79.2	75.6	81.8
RISJ880101	75.0	65.9	81.8
KAPO950101	77.1	68.3	83.6
TANS760102_dC	71.9	65.9	76.4
TANS760101_dN	74.0	65.9	80.0
All	80.2	78.0	81.8
Strand			
Snp	75.6	70.2	80.0
Polar	77.2	73.7	80.0
Sulphur-contain	69.3	66.7	71.4
pK	80.3	78.9	81.4
ΔC_{ph}	74.8	71.9	77.1
LUTR910102	76.4	71.9	80.0
NIEK910102	72.4	71.9	72.9
QU_C930103	71.7	66.7	75.7
TUDE900101	74.8	70.2	78.6
RUSR970101	73.2	70.2	75.7
MUET010101	76.4	75.4	77.1
THOP960101_dC	79.5	78.9	80.0
BONM030104_dC	78.7	73.7	82.9
ZHAC000104_dC	69.3	64.9	72.9
MOOG990101_dN	76.4	71.9	80.0
SKOJ000101_dN	78.0	77.2	78.6
SIMK990101_dN	74.8	71.9	77.1
Conservation score	81.1	78.9	82.9
All	81.9	78.9	84.3
Coil-buried			
Polar	72.1	66.7	77.0
ALTS910101	67.1	60.6	73.0
CSEM940101	72.1	69.7	74.3
KOLA930101_dC	72.9	71.2	74.3
MOOG990101_dN	70.7	68.2	73.0
SKOJ000101_dN	67.9	65.2	70.3
SIMK990102_dN	69.3	65.2	73.0
SIMK990104_dN	73.6	72.7	74.3
All	77.9	75.8	79.7
Coil-exposed			
Positive	70.1	62.5	75.8
Aliphatic	68.4	65.4	70.7
P	69.5	62.8	74.7
BENS940101	70.6	62.8	76.8
MIYT790101	70.1	60.3	77.8
RUSR970103	70.1	60.3	77.8
NGPC000101	68.9	62.8	73.7
VENM980101_dC	67.8	60.3	73.7
BONM030103_dC	68.4	61.5	73.7
VENM980101_dN	68.4	59.0	75.8
KESO980102_dN	68.9	59.0	76.8
ZHAC000104_dN	71.2	67.9	73.7
Conservation score	73.5	66.7	78.8
All	75.1	70.5	78.8

The accuracy, sensitivity and specificity are obtained by removing the respective feature from the models, the last row with "All" represents the model with all the selected features.

accessibility of a specific residue to the solvent is often influenced by the polar nature of the residues in the surrounding environment [51]. The final model for the coil mutations in solvent exposed regions, consist of features such as polar nature of the mutated position and number of aliphatic and charged amino acids in the neighbouring residues within a window length of 13 residues in the sequence and conservation score of the mutation position. These selected features suggest that mutations, which disrupt some of the key interactions especially hydrophobic and salt bridges in the neighbouring regions of their occurrence, might alter the native function of the protein and act as drivers.

3.8. Analysis based on different counts of driver mutations in COSMIC database

We have developed different models by considering the number of recurring mutations in COSMIC database to assign the driver mutations and the results obtained with three cut-off values (>1, >2 and >3) are presented in Table 5. We obtained an average prediction accuracy of 78.8%, 85.0% and 79.9% for the models obtained with the counts of >1, >2 and >3, respectively. This analysis shows that the present method could be used to predict the driver mutation at different counts for driver mutations.

3.9. Performance of the model using different training and test sets of data

We have further evaluated the performance of the model using different proportions of training and test sets of data. The 10-fold cross-validation results obtained with 90%, 80%, 70% and 60% of training data are presented in Table 6. We noticed that the average accuracy lies in the range of 77–79% using four different sets of data. It is noteworthy that the results are consistent in all the cases. The test set results with the remaining set of mutants, 10%, 20%, 30% and 40% are also included in the table. The average accuracy for the test set of data varies between 4% and 11% among different sets. This analysis reveals that the present method could be effectively used to identify the driver mutations in EGFR.

3.10. Prediction of known drivers

We have collected a list of known drivers from the literature and checked the prediction performance of the present model. Experimentally known driver mutations are L861Q [52], R108K, T263P, A289V, A289D, A289T, G598V [53], T751I, R748K, E804G, F856L, A839V, G863D, V851I [54], N700D, E709A, E709V, E709K, E709G, G719A, G719C, G719S, S720F, L747S, V765M, S768I, R776C, R776H, G779F, K806E, L814P, L833V, R836C, L838P, A839T, F856L, G857R, L858R, L861R, G863D [55], T847I, L688P, P694L, P694S, L730F, P733L, G735S, V742A, E746K, D761N, S784F, G810S [18], E884K, H835L, W731L, C797Y, Y801H, R831H, E868G [56], H870R and Q787R [57]. Interestingly, our method could correctly predict 84% of the mutants as drivers. This dataset includes four secondary site mutants, T790M, L747S, D761Y and T854A, which are mostly observed after treatment with tyrosine kinase inhibitors (TKI) [58,59]. However, reports are available in the literature that they are present even prior to the treatment [26–28].

3.11. Comparison with other methods

We compared the results of our model with different available methods such as SIFT, Polyphen-2, Mutation Assessor using the same dataset of 540 mutations. SIFT is one of the most widely used tools to predict the amino acid substitutions that affects protein function based on sequence homology and physical properties of amino acids. Polyphen-2 is a tool for prediction of possible impact of an amino acid substitution on the structure and function of a human protein based on sequence, structural and phylogenetic features that characterize the substitution. Mutation Assessor predicts the functional impact of amino acid substitutions in proteins involved in cancer as well as missense polymorphisms based on evolutionary conservation of the affected amino acid in protein homologs.

Table 5
Performance of the models using different counts (>1, >2 and >3) to assign driver mutations.

Model	Measure	10-fold cross validation			Test set		
		>1	>2	>3	>1	>2	>3
Helix	No. of drivers, Passengers	41, 55	22, 74	11, 85	5, 6	3, 8	3, 8
	Accuracy (%)	80.2	85.9	89.7	90.9	90.9	83.8
	Sensitivity (%)	78.0	60.0	85.7	80.0	75.0	66.7
	Specificity (%)	81.8	93.9	90.3	100	100	100
	ROC	0.80	0.75	0.87	0.87	0.82	0.80
Strand	No. of drivers, Passengers	57, 70	29, 98	21, 106	6, 8	7, 7	4, 10
	Accuracy (%)	81.9	85.1	81.6	85.7	85.7	85.7
	Sensitivity (%)	78.9	66.7	80.0	83.3	80.0	75.0
	Specificity (%)	84.3	91.4	81.9	87.5	88.9	90.0
	ROC	0.83	0.83	0.83	0.83	0.91	0.85
Coil-buried	No. of drivers, Passengers	66, 74	42, 98	31, 109	7, 8	7, 8	4, 11
	Accuracy (%)	77.9	85.2	77.4	80.0	80.0	73.3
	Sensitivity (%)	75.8	77.6	80.0	85.7	77.8	66.7
	Specificity (%)	79.7	88.7	76.7	75.0	83.3	75.0
	ROC	0.81	0.84	0.82	0.82	0.82	0.86
Coil-exposed	No. of drivers, Passengers	78, 99	39, 138	24, 153	6, 14	5, 15	5, 15
	Accuracy (%)	75.1	83.8	71.1	85.0	90.0	70.0
	Sensitivity (%)	70.5	70.5	72.4	83.3	80.0	60.0
	Specificity (%)	78.8	87.6	70.8	85.7	93.3	81.8
	ROC	0.77	0.80	0.75	0.89	0.90	0.68

>1, >2 and >3 represent the models obtained with driver counts >1, >2 and >3, respectively.

These methods have been developed for predicting protein disruption (both gain and loss of function). Hence, it is not appropriate to directly compare the predictive performance of the present method with other methods. However, the comparison would provide the information that utilizing the method specific to EGFR, in which driver mutations lead to gain of function, could enhance the prediction performance. The accuracy, sensitivity, specificity of different methods is reported in Table 7. We observed that the performance of these three methods on EGFR mutation resulted in a very low classification accuracy. SIFT method could classify only ~50% of the EGFR mutants correctly whereas the other tools polyphen-2 and mutation assessor classified mutants with low accuracy of 49.0% and specificity of 18.0% and 50.0%, respectively. Comparatively, the current method showed better results with high accuracy and sensitivity. This analysis suggests that target dependent method is necessary for identifying driver mutations as represented in the case of protein folding and stability, and recognition mechanism of protein-RNA complexes [60,61]. The present method developed for EGFR could correctly distinguish between driver and passenger mutations with an accuracy of 85.4%, which is 36% higher than that of other methods in the literature.

3.12. Screening of all possible point mutants

Further, we have generated all possible point mutations for each position in EGFR and screened them using the developed models for helix, strand, coil-buried and coil-exposed mutants. Among the possible mutants in helix, strand and coil structures, the topmost 10 mutants predicted with the highest probability of being a driver or passenger are listed in Tables 8a and 8b. The results for all mutants

Table 6
Performance of the model on different training and test sets.

Model	Measure	10-fold cross validation				Test set			
		90%	80%	70%	60%	10%	20%	30%	40%
Helix	Accuracy (%)	80.2	76.7	76.0	78.5	90.9	90.5	87.5	78.6
	Sensitivity (%)	78.0	70.3	68.8	77.8	80.0	77.8	78.6	69.0
	Specificity (%)	81.8	81.6	81.4	78.9	100	100	94.4	87.0
	ROC	0.80	0.75	0.74	0.83	0.87	0.94	0.90	0.79
Strand	Accuracy (%)	81.9	80.5	81.8	77.7	85.7	78.6	78.6	71.4
	Sensitivity (%)	78.9	80.0	79.5	73.7	83.3	70.0	68.4	72.0
	Specificity (%)	84.3	81.0	83.6	80.9	87.5	86.7	87.0	71.0
	ROC	0.83	0.84	0.83	0.77	0.83	0.87	0.83	0.77
Coil-buried	Accuracy (%)	77.9	79.0	75.8	77.4	80.0	74.2	80.4	75.8
	Sensitivity (%)	75.8	74.6	74.6	75.0	85.7	71.4	76.2	75.9
	Specificity (%)	79.7	83.1	76.9	73.5	75.0	76.5	84.0	75.8
	ROC	0.81	0.82	0.81	0.79	0.82	0.81	0.83	0.83
Coil-exposed	Accuracy (%)	75.1	75.3	74.6	77.9	85.0	82.1	74.6	73.4
	Sensitivity (%)	70.5	70.6	70.0	73.5	83.3	75.0	70.0	60.0
	Specificity (%)	78.8	78.9	78.8	81.2	85.7	87.0	78.8	84.1
	ROC	0.77	0.81	0.79	0.82	0.89	0.79	0.79	0.75
Average	Accuracy (%)	78.3	77.8	77.0	77.8	85.4	81.4	80.3	75.0
	Sensitivity (%)	75.8	73.9	73.2	75.0	83.1	73.6	73.3	72.6
	Specificity (%)	81.2	81.2	80.2	78.6	87.1	87.6	86.1	79.5
	ROC	0.80	0.81	0.79	0.80	0.85	0.85	0.84	0.78

90%, 80%, 70% and 60% are different proportions of training sets and 10%, 20%, 30% and 40% are test sets, respectively.

Table 7
Comparison with different methods.

Dataset	SIFT			Polyphen-2			Mutation assessor			Current method		
	Ac	Sn	Sp	Ac	Sn	Sp	Ac	Sn	Sp	Ac	Sn	Sp
(A) Cross validation dataset of 540 mutants												
Helix	44.8	78.1	20.0	46.9	92.7	12.7	48.9	41.5	54.6	80.2	78	81.8
Strand	53.5	91.2	22.9	53.5	91.2	22.9	50.4	52.6	48.6	81.1	78.9	82.9
Coil-buried	47.9	77.3	21.6	49.3	89.4	13.5	52.1	63.6	41.9	77.9	75.8	79.7
Coil-exposed	51.4	64.1	41.4	47.5	78.2	23.2	50.3	37.2	60.6	73.5	70.5	75.8
(B) Test set of 60 mutants												
Helix	45.5	80.0	16.7	45.5	80.0	16.7	45.5	20.0	66.7	90.9	80.0	100
Strand	57.1	100	25	57.1	100	25	50.0	100	12.5	85.7	83.3	87.5
Coil-buried	40.0	85.7	0	46.7	100	0	40.0	42.9	37.5	80.0	85.7	75.0
Coil-exposed	50.0	83.3	35.7	50.0	83.3	35.7	50.0	33.3	57.1	85.0	83.3	85.7

The results for current method are obtained by 10-fold cross validation. Accuracy (Ac), Sensitivity (Sn) and Specificity (Sp) are given in %.

Table 8a
Topmost 10 mutants predicted with high probability as driver.

No.	Helix	Strand	Coil-buried	Coil-exposed
1	**Q684F**	**G724F**	**M137K**	I214K
2	**Q684I**	**G729F**	**V250E**	I673K
3	**A840I**	**G341I**	R836W	T710D
4	**A955I**	**G729E**	**R836V**	A702D
5	**A763I**	**G729I**	V250D	**T710P**
6	**A661I**	**G724I**	V250N	*V674K*
7	**A1000I**	**G729D**	**C295K**	I673E
8	**A375I**	G729K	**N422E**	T710E
9	A661C	**P741I**	**V250K**	T710C
10	**A840F**	**G719F**	**V1010D**	T363E

Mutations predicted as drivers with all the three models are shown in bold. Predicted driver mutations with two models are underlined.

are available at http://www.iitm.ac.in/bioinfo/EGFR_Driver/. In the predicted drivers, most of them are Alanine and glycine substitutions in helix and strand structures, respectively. Most of these mutations predicted as drivers are present in intracellular tyrosine kinase domain close to the known driver mutants (D761Y, T790M, T854A, L858R) which play a key role in affecting its activity and leading to over expression [62].

Further, we have analysed the predicted drivers to understand their role in causing cancer. To identify the residues which are functionally important, we have obtained the evolutionary conservation scores of the mutation site using Consurf server. The degree to which an amino acid position is evolutionarily conserved is strongly dependent on its structural and functional importance. Among the predicted drivers at different positions in the protein sequence, 50% of the mutation sites in the topmost 10 drivers in Table 5 are highly conserved which reveal their importance in maintaining the native structure and function of the protein. On the other hand, 90% of the predicted passenger mutations are located in the extracellular cellular domain of EGFR and most of these positions are not conserved. This analysis emphasizes the importance of conservation score, which is a useful feature in discriminating the driver and passenger mutants in strand and coil-exposed regions. Further, in strand mutants, we observed that 9 out of topmost 10 predicted drivers constitute Glycine substitutions occurring at ATP binding region. Interestingly, most of them are being replaced by hydrophobic amino acids with bulky side chains. This suggest the importance of Glycine residues to provide flexibility to the structure and we hypothesize that the hydrophobic mutant residues cause distortion of native structure by causing steric hindrance and alter the crystal packing, that might in turn affect the native function of EGFR protein.

Table 8B
Topmost 10 mutants predicted with high probability as passenger.

No.	Helix	Strand	Coil-buried	Coil-exposed
1	**W898E**	**Y113W**	**L156C**	**N540C**
2	**W898G**	**C231Y**	**W200G**	**N540L**
3	**W898S**	**Y270W**	**L448C**	**F1023W**
4	**W164D**	**V308Q**	**N516G**	**K189Y**
5	**W164E**	**K328S**	**L633C**	**S198W**
6	**W164G**	**I556E**	**E634M**	**K226Y**
7	**W164S**	I556K	**I643C**	**N540Y**
8	**G197W**	I556N	**L38G**	**D179C**
9	**E344W**	**I556Q**	**I1060N**	**K189M**
10	**H358G**	**I556R**	**V1147Y**	**N280F**

Mutations predicted as drivers with all the three models are shown in bold. Predicted driver mutations with two models are underlined.

4. Conclusion

We have analysed a set of 266 driver and 334 passenger mutations in EGFR using sequence based features such as wild-type, mutant residue, number of neighbouring residues at different window lengths. Driver mutations are dominated with leucine and glycine substitutions in helix and strand, whereas coil-buried and coil-exposed mutants are dominated with substitutions in charged residues Arginine and Glutamic acid, respectively. Different machine learning algorithms have been tested to discriminate driver and passenger mutations and we achieved an accuracy of 67.8% with SVM based classifier, in 10-fold cross validation on a dataset of 540 mutations. Further, we grouped the mutations into four classes based on secondary structure and accessible surface area, which enhanced the classification accuracy to 80.2%, 81.8%, 77.9% and 75.1% for helix, strand, coil-buried and coil-exposed mutants, respectively. An independent test set of 60 mutations is used to assess our method which showed an accuracy of 90.9%, 85.7%, 80.0% and 85.0%, respectively. In comparison with other available methods, this model performed well on EGFR mutation dataset. We have screened all possible point mutants in EGFR using this method and provided the prediction results. We suggest that the prediction results could help in the development of mutation specific drugs for cancer therapy.

Conflict of interest

There is no conflict of interest.

Acknowledgments

MMG thanks Professor Y-h. Taguchi, Chuo University, Japan for fruitful discussions. We acknowledge the reviewers for their constructive comments on our manuscript. We thank the Bioinformatics facility and Indian Institute of Technology Madras for

computational facilities. PA thanks Department of Science and Technology (DST), India for providing research fellowship.

Grant sponsor: Department of Science and Technology, India.

References

[1] D. Karunagaran, E. Tzahar, N. Liu, D. Wen, Y. Yarden, Neu Differentiation Factor inhibits EGF Binding: a model for trans-regulation within the erbb family of receptor tyrosine kinases, J. Biol. Chem. 270 (1995) 9982–9990.
[2] Y. Yarden, J. Schlessinger, Epidermal growth factor induces rapid reversible aggregation of the purified epidermal growth factor receptor, Biochemistry 26 (1987) 1443–1451.
[3] C. Greenman, P. Stephens, R. Smith, G.L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, S. Edkins, S. O'Meara, et al., Patterns of somatic mutation in human cancer genomes, Nature 446 (2007) 153–158.
[4] T.J. Lynch, D.W. Bell, R. Sordella, S. Gurubhagavatula, R.A. Okimoto, B.W. Brannigan, P.L. Harris, S.M. Haserlat, J.G. Supko, F.G. Haluska, D.N. Louis, D.C. Christiani, et al., Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib, N. Engl. J. Med. 350 (2004) 2129–2139.
[5] B.F. El-Rayes, P.M. LoRusso, Targeting the epidermal growth factor receptor, Br. J. Cancer 91 (2004) 418–424.
[6] W. Pao, V. Miller, M. Zakowski, J. Doherty, K. Politi, I. Sarkaria, B. Singh, R. Heelan, V. Rusch, L. Fulton, E. Mardis, D. Kupfer, et al., EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib, Proc. Natl. Acad. Sci. U.S.A. 101 (2004) 13306–13311.
[7] H. Greulich, T.H. Chen, W. Feng, P.A. Janne, J.V. Alvarez, S.E. Bulmer, M. Zappaterra, D.A. Frank, W.C. Hahn, W.R. Sellers, M. Mayerson, Oncogenic transformation by inhibitor-sensitive and resistant EGFR mutations, PLoS Med. 2 (2005) e313.
[8] D.D. Wang, W. Zhou, H. Yan, M. Wong, V. Lee, Personalized prediction of EGFR mutation-induced drug resistance in lung cancer, Sci. Rep. 3 (2013) 2855.
[9] P.C. Ng, S. Henikoff, SIFT: predicting amino acid changes that affect protein function, Nucl. Acids Res. 31 (2003) 3812–3814.
[10] I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, A.S. Kondrashov, S.R. Sunyaev, A method and server for predicting damaging missense mutations, Nat. Methods 7 (2010) 248–249.
[11] B. Reva, Y. Antipin, C. Sander, Predicting the functional impact of protein mutations: application to cancer genomics, Nucl. Acids Res. 39 (2011) e118.
[12] P.D. Thomas, A. Kejariwal, N. Guo, H. Mi, M.J. Campbell, A. Muruganujan, B. Lazareva-Ulitsky, Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools, Nucl. Acids Res. 34 (2006) 645–650.
[13] C. Ferrer-Costa, J.L. Gelpí, L. Zamakola, I. Parraga, X. De La Cruz, M. Orozco, PMUT: a web-based tool for the annotation of pathological mutations on proteins, Bioinformatics 21 (2005) 3176–3178.
[14] J.S. Kaminker, Y. Zhang, C. Watanabe, Z. Zhang, CanPredict: a computational tool for predicting cancer-associated missense mutations, Nucl. Acids Res. 35 (2007) 595–598.
[15] H. Carter, S. Chen, L. Isik, S. Tyekucheva, V.E. Velculescu, K.W. Kinzler, B. Vogelstein, R. Karchin, Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations, Cancer Res. 69 (2009) 6660–6667.
[16] B. Li, V.G. Krishnan, M.E. Mort, F. Xin, K.K. Kamati, D.N. Cooper, S.D. Mooney, P. Radivojac, Automated inference of molecular mechanisms of disease from amino acid substitutions, Bioinformatics 25 (2009) 2744–2750.
[17] S.A. Forbes, G. Bharma, S. Bamford, E. Dawson, M.R. Stratton, The catalogue of somatic mutations in cancer (COSMIC), Curr. Protoc. Hum. Genet. 57 (2008) 20–26.
[18] R.K. Kancha, N. von Bubnoff, C. Peschel, J. Duyster, Functional analysis of epidermal growth factor receptor (EGFR) mutations and potential implications for EGFR targeted therapy, Clin. Cancer Res. 15 (2009) 460–467.
[19] T.M. Gilmer, L. Cable, K. Alligood, D. Rusnak, G. Spehar, K.T. Gallagher, E. Woldu, H.L. Carter, A.T. Truesdale, L. Shewchuk, E.R. Wood, Impact of common epidermal growth factor receptor and HER2 variants on receptor activity and inhibition by lapatinib, Cancer Res. 68 (2008) 571–579.
[20] A. Tatematsu, J. Shimizu, Y. Murakami, Y. Horio, S. Nakamura, T. Hida, T. Mitsudomi, Y. Yatabe, Epidermal growth factor receptor mutations in small cell lung cancer, Clin. Cancer Res. 14 (2008) 6092–6096.
[21] E. Avizienyte, R.A. Ward, A.P. Garner, Comparison of the EGFR resistance mutation profiles generated by EGFR targeted tyrosine kinase inhibitors and the impact of drug combinations, Biochem. J. 415 (2008) 197–206.
[22] S.V. Sharma, D.W. Bell, J. Settleman, D.A. Haber, Epidermal growth factor receptor mutations in lung cancer, Nat. Rev. Cancer 7 (2007) 169–181.
[23] A. Mottaz, F.P. David, A.L. Veuthey, Y.L. Yip, Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar, Bioinformatics 26 (2010) 851–852.
[24] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, K. Sirotkin, dbSNP: the NCBI database of genetic variation, Nucl. Acids Res. 29 (2001) 308–311.
[25] A. Torkamani, N.J. Schork, Prediction of cancer driver mutations in protein kinases, Cancer Res. 68 (2008) 1675–1682.
[26] M. Inukai, S. Toyooka, S. Ito, H. Asano, S. Ichihara, J. Soh, H. Date, Presence of epidermal growth factor receptor gene T790M mutation as a minor clone in non-small cell lung cancer, Cancer Res. 66 (2006) 7854–7858.
[27] T. Kozuki, A. Hisamoto, M. Tabata, N. Takigawa, K. Kiura, Y. Segawa, M. Tanimoto, Mutation of the epidermal growth factor receptor gene in the development of adenocarcinoma of the lung, Lung Cancer 58 (2007) 30–35.
[28] G. Mathur, D. Ma, Coexistence of tyrosine kinase inhibitor-sensitizing and resistant EGFR Mutations in an untreated lung adenocarcinoma patient and response to erlotinib, J. Thorac Oncol. 9 (2014) e55–e57.
[29] M.M. Gromiha, A Statistical model for predicting protein folding rates from amino acid sequence with structural class information, J. Chem. Inf. Model. 45 (2005) 494–501.
[30] S. Kawashima, M. Kanehisa, AAindex: amino acid index database, Nucl. Acids Res. 28 (2000) 374.
[31] B. Lee, F.M. Richards, The interpretation of protein structures: estimation of static accessibility, J. Mol. Biol. 55 (1971) 379–384.
[32] M.M. Gromiha, S. Selvaraj, Influence of medium and long range interactions in different structural classes of globular proteins, J. Biol. Phys. 23 (1997) 151–162.
[33] P. Manavalan, P.K. Ponnuswamy, Hydrophobic character of amino acid residues in globular proteins, Nature 275 (1978) 673–674.
[34] J.K. Myers, T.G. Oas, Preorganized secondary structure as an important determinant of fast protein folding, Nat. Struct. Biol. 8 (2001) 552–558.
[35] A. Ben-Naim, The role of hydrogen bonds in protein folding and protein association, J. Phys. Chem. 95 (1991) 1437–1444.
[36] P.W. Rose, C. Bi, W.F. Bluhm, C.H. Christie, D. Dimitropoulos, S. Dutta, P.E. Bourne, The RCSB Protein Data Bank: new resources for research and education, Nucl. Acids Res. 41 (2013) 475–482.
[37] K. Arnold, L. Bordoli, J. Kopp, T. Schwede, The SWISS_MODEL Workspace: a web-based environment for protein structure homology modelling, Bionformatics 22 (2006) 195–201.
[38] R. Adamczak, A. Porollo, J. Meller, Combining prediction of secondary structure and solvent accessibility in proteins, Proteins 59 (2005) 467–475.
[39] H. Ashkenazy, E. Erez, E. Martz, T. Pupko, N. Ben-Tal, ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids, Nucl. Acids Res. 38 (2010) 529–533.
[40] M.M. Gromiha, S. Selvaraj, Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction, J. Mol. Biol. 310 (2001) 27–32.
[41] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.
[42] I.K. McDonald, J.M. Thornton, Satisfying hydrogen bonding potential in proteins, J. Mol. Biol. 238 (1994) 777–793.
[43] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, in: ACM SIGKDD Explorations Newsletter, 2009, pp. 10–18, 11.
[44] Yugandhar, Gromiha, Feature selection and classification of protein-protein complexes based on their binding affinities using machine learning approaches, Proteins 82 (2014) 2088–2096.
[45] J. Schlessinger, Cell signaling by receptor tyrosine kinases, Cell 103 (2000) 211–225.
[46] C.H. Yun, T.J. Boggon, Y. Li, M.S. Woo, H. Greulich, M. Meyerson, M.J. Eck, Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity, Cancer Cell. 11 (2007) 217–227.
[47] S. Khan, M. Vihinen, Spectrum of disease-causing mutations in protein secondary structures, BMC Struct. Biol. 7 (2007) 56.
[48] K. Saraboji, M.M. Gromiha, M.N. Ponnuswamy, Relative importance of secondary structure and solvent accessibility to the stability of protein mutants: a case study with amino acid properties and energetics on T4 and human lysozymes, Comput. Biol. Chem. 29 (2005) 25–35.
[49] C.T. Saunders, B. David, Evaluation of structural and evolutionary contributions to deleterious mutation prediction, J. Mol. Biol. 4 (2002) 891–901.
[50] A.V. Jorge, D.R. Ripoll, H.A. Scheraga, Physical reasons for the unusual α-helix stabilization afforded by charged or neutral polar residues in alanine-rich peptides, Proc. Natl. Acad. Sci. U.S.A. 97 (2000) 13075–13079.
[51] L. Gonzalez, D.N. Woolfson, T. Alber, Buried polar residues and structural specificity in the GCN4 leucine zipper, Nat. Struct. Biol. 3 (1996) 1011–1018.
[52] S. Luo, D. Lam, Oncogenic driver mutations in lung cancer, Tran. Resp. Med. 1 (2013) 6.
[53] J.C. Lee, I. Vivanco, R. Beroukhim, J.H. Huang, W.L. Feng, R.M. DeBiasi, I.K. Mellinghoff, Epidermal growth factor receptor activation in glioblastoma through novel missense mutations in the extracellular domain, PLoS Med. 3 (2006) e485.
[54] C. Peraldo-Neia, G. Migliardi, M. Mello-Grand, F. Montemurro, R. Segir, Y. Pignochino, M. Aglietta, Epidermal Growth Factor Receptor (EGFR) mutation analysis, gene expression profiling and EGFR protein expression in primary prostate cancer, BMC Cancer 11 (2011) 31.
[55] F.L. Simonetti, C. Tornador, N. Nabau-Moretó, M.A. Molina-Vila, C. Marino-Buslje, Kin-Driver: a database of driver mutations in protein kinases, Database (2014) bau104.
[56] K.D. Carey, A.J. Garton, M.S. Romero, J. Kahler, S. Thomson, S. Ross, M.X. Sliwkowski, Kinetic analysis of epidermal growth factor receptor somatic mutant proteins shows increased sensitivity to the epidermal growth factor receptor tyrosine kinase inhibitor, erlotinib, Cancer Res. 66 (2006) 8163–8171.

[57] I. Yee-San Tam, E.L.H. Leung, V.P.C. Tin, D.T.T. Chua, A.D.L. Sihoe, L.C. Cheng, M.P. Wong, Double EGFR mutants containing rare EGFR mutant types show reduced in vitro response to gefitinib compared with common activating missense mutations, Mol. Cancer Ther. 8 (2009) 2142–2151.
[58] C. Ma, S. Wei, Y. Song, T790M and acquired resistance of EGFR TKI: a literature review of clinical reports, J. Thorac. Dis. 3 (2011) 10.
[59] L. Lin, T.G. Bivona, Mechanisms of resistance to epidermal growth factor receptor inhibitors and novel therapeutic strategies to overcome resistance in NSCLC patients, Chemother. Res. Pract. (2012), http://dx.doi.org/10.1155/2012/817297
[60] V. Potapov, M. Cohen, G. Schreiber, Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details, Protein Eng. Des. Sel. 22 (2009) 553–560.
[61] R. Nagarajan, S.P. Chothani, C. Ramakrishnan, M. Sekijima, M.M. Gromiha, Structure based approach for understanding organism specific recognition of protein-RNA complexes, Biol. Direct. 10 (2015) 8.
[62] C.H. Yun, K.E. Mengwasser, A.V. Toms, M.S. Woo, H. Greulich, K.K. Wong, M. Meyerson, M.J. Eck, The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP, Proc. Natl. Acad. Sci. U.S.A. 105 (2008) 2070–2075.