Parametric Models for Understanding Atomic Trajectories in Different Domains of Lung Cancer Causing Protein

QUreshi, Rizwan; NAWaz, Mehmood; GHOsh, Avirup; YAN, Hong

Published in:
IEEE Access

Published: 01/01/2019

Document Version:
Final Published version, also known as Publisher’s PDF, Publisher’s Final version or Version of Record

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.1109/ACCESS.2019.2918294

Publication details:

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher's definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher's copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.
Parametric Models for Understanding Atomic Trajectories in Different Domains of Lung Cancer Causing Protein

RIZWAN QURESHI©, MEHMOOD NAWAZ©, AVIRUP GHOSH©, (Student Member, IEEE), AND HONG YAN©, (Fellow, IEEE)
Department of Electrical Engineering, City University of Hong Kong, Hong Kong

Corresponding author: Rizwan Qureshi (engr.rizwanqureshi786@gmail.com)

This work was supported in part by the Hong Kong Research Grants Council under Project 11200818, and in part by the City University of Hong Kong under Project 9610034.

ABSTRACT Non-small cell lung cancer (NSCLC) is a major cause of death worldwide. About 80% to 85% of lung cancer cases are NSCLC. It is well known that mutation of the epidermal growth factor (EGFR) may lead to the NSCLC. The first generation drugs are effective initially, but almost all patients develop drug resistance after about a year due to a secondary mutation. The computational methods are an efficient tool for investigating drug resistance, design, and discovery. Moreover, molecular dynamics (MD) simulation enables us to study and analyze the behavior of proteins and molecules at the atomic level. MD simulations offer extraordinary insight about biomolecules and are a valuable tool for computer aided drug discovery. Earlier studies on EGFR only focused on the kinase domain. Because EGFR is a multi-domain protein, mutations in the kinase domain may affect the function in other domains. Therefore, it is important to investigate the complete structure of the EGFR and its mutants. In this paper, we first generate the complete structure of the EGFR and perform MD simulation for the wildtype EGFR, EGFR with L858R mutation and EGFR with L858R and T790M mutation. We divide the complete structure of the EGFR and its mutants into 8 domains according to the reference crystal structure. We then consider atom trajectories as time series signals and estimate the power spectral densities using the auto-regressive integrated (ARI) model, which shows interesting insight. Dynamic time warping is used to analyze the similarity between each domain of the structures. Interesting patterns are observed which may be useful for investigating drug resistance and design. Furthermore, Pearson correlation coefficient, peaks, and widths of the power spectral density are calculated for each domain. The simulation results provide useful insight about conformation dynamics of the EGFR, such as atom motion and protein stability. The domains are less correlated in L858R type and even weaker when the second mutation occurs. The warping patterns are changed due to mutation and the movement of atoms is distorted. Hence, it is difficult for a drug to bind to the protein. These findings will be useful in understanding the characteristics of the EGFR and for computer aided drug design process for the NSCLC patients.

INDEX TERMS Autoregressive integrated model, drug resistance, dynamic time warping, epidermal growth factor receptor, molecular dynamics, non-small cell lung cancer.

I. INTRODUCTION
Cancer is one of the most deadly diseases worldwide, caused by the uncontrolled cell growth [1], [2]. According to the World Health Organization (WHO), cancer is the second largest cause of death with an estimated rate of 9.6 million deaths in 2018 and globally one out of every six deaths is due to cancer [3]. The stage of a lung cancer describes its spread and determined by the parameter TNM. Here T denotes the size of the tumor, N shows its spread to lymph nodes and M refers whether it has metastasized into tissues. T1N0M0 denotes the first stage and T1N1M1b denotes the last stage. Early cancer detection stages have higher survival rate, approximately 50% and the last stage have 2% survival rate [4].

The treatment strategies for lung cancer depend upon the detection stage and patient’s characteristics. For example, if the cancer is confined to one lung and patient is in good...
health, surgery can be applied to remove the cancerous cells followed by chemotherapy to destroy any cell that have remained in the body. Radiotherapy is recommended, when the cancer cells have not spread too far but surgery is not possible, due to patient’s health condition. Chemotherapy is recommended, if the cancer cells have spread too far for surgery and radiotherapy to be effective. In some cases, biological or targeted therapy is recommended as an alternate to chemotherapy [5]. In targeted therapy, EGFR is identified as a pathogenic factor in NSCLC and several抗癌 drugs have been targeted against the EGFR [6]. EGFR activates downstream pathways that allow the tumor cells to survive and grow.

The US food and drug administration agency (FDA) approved two EGFR targeting small molecule inhibitors gefitinib and erlotinib for NSCLC patients who had responded to conventional chemotherapy [7]. The EGFR mutations also show a striking correlation with patient characteristics such as gender and region. For example, EGFR kinase inhibitor gefitinib (Iressa) causes a decrease in the size of tumor, more frequently in Japanese patients [1]. Most of the anticancer therapeutics tend to bind the EGFR to stop the downstream signaling. About 60% of NSCLC patients show over-expression of EGFR, making it an appealing target for many treatment strategies [8].

Despite the initial dramatic response to EGFR mutant cases treated with gefitinib or erlotinib, the development of drug resistance after about a year limits the ability of these drugs to significantly prolong the survival rate. A single secondary mutation T790M is present in a subset of EGFR mutant tumors after initial response to gefitinib or erlotinib [9]. The T790M mutants retain low-nanomolar affinity to gefitinib. Moreover, T790M mutation activates the WT EGFR and increases the ATP affinity of the oncogenic L858R mutant. The increased ATP is the primary mechanism due to which T790M mutation causes drug resistance [9]. Studies in [1] and [7] also show that the mutation of EGFR is one of the major cause of drug resistance. Some studies also suggest the presence of T790M mutation even before the patient is exposed to drugs [10], [11]. A third generation drug Osimertinib is developed to treat NSCLC patients if the cancer cells are positive to T790M mutation [12]. The resistance to Osimertinib usually develops after about 10 months. EGFR C797S mutation is the most commonly observed mutation to Osimertinib [13].

The mutation is the alteration in the residue sequence of a protein. The alteration includes substitution, insertion, deletion, modification and duplication of residues in the protein sequence [1]. L858R and T790M have been identified as the most frequent mutation types in NSCLC patients [14]. In this paper, we analyze the motion of wildtype and these two mutated type EGFR. Earlier studies have been done using only the kinase domain of EGFR. To the best of our knowledge, this is the first study on the complete structure of EGFR and its mutants.

Both experimental and computational methods have been applied to investigate the motion of EGFR mutants induced drug resistance and discovery. The past decade has witnessed tremendous growth in computational capabilities that expedite the drug design process [15]. Molecular dynamics (MD) simulation aided with parallel computing and GPUs and coupled with advanced deep learning algorithms can solve many challenging problems of protein dynamics and structure prediction. It offers great opportunities to study and analyze the motion of the atoms/residues at the atomic level. The simulation can uncover hidden motions and important features of bio-molecules at atomic scales, which can be useful for computer aided drug design for NSCLC patients.

With this motivation, we aim to investigate the atomic trajectories of EGFR and its mutant structures. We use MD simulation to generate the trajectory of 200 ps motion of EGFR. We then apply auto-regressive integrated (ARI) model, Pearson correlation coefficient and dynamic time warping (DTW) to extract useful information from these trajectories. Our contribution can be itemized as follows.

- Generating complete EGFR structure.
- Generating 200 frames of dynamics simulation of EGFR and its mutants.
- Correlation analysis of each domain.
- Analysis of power spectral density of each domain.
- Analysis using dynamic time warping of each domain.
- Drug resistance analysis.

The purpose of the paper is twofold. The first is to investigate the atomic trajectories of EGFR and drug resistance with the help of computational methods. The second is to attract computer scientists and signal processing experts to explore this potential avenue and contribute to cancer data analysis. The rest of the paper is organized as follows. In Section 2, prior work is discussed, and in Section 3, we describe database, mutation and homology modeling. In Section 4, we present our methodology. Section 5 presents analysis. Section 6 highlights results and discussion. Finally, Section 7 concludes the paper with future possibilities.

II. PRIOR WORK

A lot of research work has been done using experimental methods to investigate the drug resistance [2]. Developing a market approved drug costs approximately $2.6 billion [15]. Computational methods are also applied due to low cost, flexibility, easy implementation and large datasets [16], [17]. Recent work in [18] predicts the protein function based on an iterative algorithm. Ma et al. [19] employed support vector machine based classifiers to correlate the energy binding features with geometrical shape of the protein. Wang et al. [20] used some personal clinical information such as; age, gender, sex, smoking history with binding energy of EGFR to build a classifier for the prediction of response levels of drugs, they achieved an accuracy of 95.83%. Mustafa and Grubmüller studied the conformational regulation of the EGFR kinase core by the juxtamembrane and C-terminal tail [21].
Lange et al. [22] used full correlation analysis based on mutual analysis to detect all the correlated motion in conformational dynamics. Simulation results show that they were able to achieve better results than principal component analysis. Ma et al. [23] employed alpha shape dynamics to model the EGFR mutant induced drug resistance. In a recent study, Ghosh and Yan [24] showed that mutation breaks the hydrogen bonds in EGFR and the overall systems is less stable.

Ramanathan et al. [25] used tensor analysis to characterize the contact map of protein dynamics. They were able to interpret the motion of complex protein in real time. Balasubramaniam et al. [26] employed artificial intelligence for molecular communication. They provided two design aspects of neuronal networks which include (i) design of an interface between nano-device and neurons that can initiate signaling, and (ii) the design of transmission scheduling to ensure that the signal from multiple devices will reach the receiver with minimum loss. Papaleo et al. [27] used dynamic cross correlation matrix to find the correlated motions between domains of a peptidase. Ralaho et al. [28] performed experiments on different organs of humans and mouse, and they were able to show that both out frame and in frame transcript can be used to regulate protein activity or localization. Song and Yan [29] proposed a method composed of autoregressive models and hidden Markov models to detect the hidden periodicities and solenoid recognition in a protein sequence.

Inspired from all these contributions, we use parametric spectral analysis based on autoregressive integrated to analyze the motions of different domains of EGFR and its mutants. Pearson correlation coefficient is used to check the stability between different domains of EGFR and its mutant. Simulation results show that mutant structures are less correlated than the WT EGFR. Moreover, dynamic time warping (DTW) is applied to each domain. Simulation results show interesting patterns and useful insight of protein dynamics. All these methods produce consistence result and show meaningful insight that can be useful for computer aided drug discovery for NSCLC patients.

A. STRUCTURE OF EGFR

EGFR consists of a large extracellular region, a single spanning transmembrane region (TM), an intracellular region, tyrosine kinase domain and a C-terminal regulatory region. The extracellular region is further divided into four subdomains. Domains 1 and 3 are homologous binding domain and domain 2 and 4 are cysteine rich. The complete structure of EGFR is shown in Figure 1. The extracellular and intracellular region of the EGFR are briefly studied in [30]–[32]. The different activation states of an intracellular kinase domain is explained in [33]. Moreover, the complete structure is also reviewed in [34].

B. HOMOLOGY MODELING

We use homology modeling to obtain the complete structure of EGFR and its mutant structures. Homology modeling is the construction of a target protein from its template by aligning the residue sequences. Homology modeling is widely used in predicting the protein structure and often it can provide reliable and accurate structures [35]. Homology modeling is performed in Rosetta to predict the complete structure of EGFR [36], as shown in Fig. 2b. PSIRED [37] and Swiss-model are used to obtain the target secondary structure file. The predicted structures are refined using the energy minimization program sander in Amber [39].

C. MUTATION

Single point mutation is the replacement of an amino acid with another and double-point mutation is the change in amino acid sequence at two points residue sequence. In this paper, we investigate two most frequent mutations in lung cancer patients. L858R is the alteration of leucine (L) into arginine (R) at the 858th position. T790M is the alteration of Theocine into Methionine at index 790 of the residue sequence, as shown in Fig. 2 (c) and (d), respectively. First, we align the templates to the target structures and perform homology modeling to predict the complete structure. The First mutation (L858R) is applied at 858th position of the complete structure and the second mutation (T7790M) is applied at 790 residue index of the EGFR. Mutagenesis protocol of Pymol [36] is adopted to generate the mutant structures.
R. Qureshi et al.: Parametric Models for Understanding Atomic Trajectories

![Figure 2](image-url)

**FIGURE 2.** In (a), the templates of different domains are shown. 4UV7 represents extracellular domain (1-4), 2N5S shows trans-membrane domain and juxtamembrane domain (5-6) and 2ITY represents kinase domain (7). In (b), the alignment of templates and the complete EGFR structure is shown. In (c), there is L858R single point mutation. In (d), there are L858R and T790M double point mutation.

which uses a rotamer library for predicting different structures. Since we only dealt with simple point mutation in our analysis, such structures can have good accuracy. The rotamer having the highest percentage for being correct is selected to generate the mutant structures.

The models predicted by simulation may not be accurate and assessment of a model is a very important task. We verified our predicted models based on the full atom energy scoring function and the one with the minimum energy score was selected for further analysis.

### III. PROPOSED FRAMEWORK

#### A. MOLECULAR DYNAMICS SIMULATION

Molecular dynamics (MD) is a simulation method that can be used to analyze the motion and behavior of proteins, atoms, molecules and complexes [39], [40]. It enables us to analyze the motions of molecules at the atomic level. MD provides computational methods for computer aided drug design. With the advancement in big data, artificial intelligence and parallel computing methods, there is a significant improvement in simulation speed, accuracy and now it is possible to have long simulations. These simulation coupled with the proliferation of experiment structural data have made MD more appealing over experimental methods [41], as they are flexible, efficient and cost effective.

MD Simulations are based on the Newton’s classical laws of physics. Using the Newton’s second law of motion, the trajectories of positions, accelerations and velocities of each atom can be obtained.

In this paper, we have used Amber [39] to perform the MD simulations. Amber is one of the most widely used packages for biomolecule simulation, such as protein. There are other molecular computer simulation program as well, such as, GROMACS [42] and CHARMM [43]. We chose Amber because our group developed software based on this platform before and we joined an active user community [44]. We used ff14SB force field and a TIP3P water box for each structure in xleap for solvation. Xleap is a program in Amber to generate the coordinates and topology file for MD Simulations. The total energy in the force field is the sum of bonded terms, bonds stretching, angle bending, torsion terms of covalent bonds and non-bonded terms, represented as electrostatic and Vander Waals forces as shown in Eq.1.

\[
E_{\text{total}} = E_{\text{stretch}} + E_{\text{bend}} + E_{\text{tor}} + E_{\text{electrostatic}} + E_{\text{vdw}}
\]  

The next step is energy minimization to guarantee a stable structure. We performed a minimization of 1000 steps to remove the bad contacts of the solute. Then we conducted a series of steps as shown in Figure 3 (heating, density equilibration and constant pressure equilibration) to equilibrate the system. At the desired temperature, pressure and density, the MD production was performed for two ns. The production files were saved for every 10 ps, resulting in 200 frames for each mutant structure. After the MD production files are generated, the trajectory for each mutant complex is obtained by using CPPTRAJ package [45].

MD simulations are performed on 12 processors of Intel Core-i7@3.4 GHz processor having 8 GB RAM and NVidia Tesla c2075 GPU.

#### B. ROOT MEAN SQUARE DEVIATION

Root mean square deviation (RMSD) is used to measure the structural difference between coordinate sets [46]. RMSD is defined as

\[
\text{RMSD} = \frac{1}{M} \sqrt{\frac{1}{N} \sum_{i=0}^{N} m_i \cdot [(X_i - Y_i)^2]}
\]  

where N is the number of atoms, \(m_i\) is the mass of \(atom_i\), \(X_i\) is the set of target coordinates, \(Y_i\) is the set of reference coordinates and M is the total mass. The Histogram of RMSD is shown in Fig.4. The blue curve is the density estimate of the histogram. The density estimate is the count of each bin divided by the total number of observations multiplied by the size of each bin.

#### C. ROOT MEAN SQUARE FLUCTUATION

Root mean square fluctuations (RMSF) measures the residue wise fluctuations [25]. RMSF highlights the portion of the structure which are fluctuating from there mean position. RMSF plot for wild type EGFR is shown in Fig. 5. RMSF is

FIGURE 3. The proposed framework. First we extract the templates from PDB [3], use homology modeling to align the complete structure followed by two types of mutations to prepare three protein data types. We perform 200 ps MD simulation to generate the trajectory of EGFR and its mutant sequences. These 200 ps simulations are used as time series signals for further analysis. We apply the ARI model, Pearson Correlation and dynamic time warping for the analysis of MD trajectories.

FIGURE 4. Root mean square deviation (RMSD). The X-axis represents RMSD values in Å and Y-axis represents the density of each value. RMSD values within 3 Å are reliable for analysis [47], [48]. The blue curve is the density estimate of the histogram.

calculated using equation 3.

$$RMSF(t) = \sqrt{\frac{1}{m-1} \sum_j (|r_j^t - r_j^\theta|^2)}$$  (3)

where m is the total number of frames, $r_j^t$ is the positional vector of $i^{th}$ atom in $j^{th}$ frame and $r_j^\theta$ is the mean position of $i^{th}$ atom. $|r|$ denotes the euclidean distance of vector r.

D. EXTRACTION OF CA ATOMS
Let Z represent the entire protein and $z_1, z_2, \ldots, z_8$ represent the 8 domains, as given by 4.

$$Z = z_1, z_2, z_3 \ldots z_8$$  (4)

FIGURE 5. Root mean square fluctuations per residue. The maximum fluctuation is observed between residues 500 - 700.

The protein structure consists of residues and in each residue there are different types of atoms.

$$\begin{align*}
z_1 &= Z(\text{resseq}(1 : 165), \text{CA}) \\
z_2 &= Z(\text{resseq}(166 : 310), \text{CA}) \\
z_3 &= Z(\text{resseq}(311 : 480), \text{CA}) \\
z_4 &= Z(\text{resseq}(481 : 620), \text{CA}) \\
z_5 &= Z(\text{resseq}(621 : 643), \text{CA}) \\
z_6 &= Z(\text{resseq}(644 : 685), \text{CA}) \\
z_7 &= Z(\text{resseq}(686 : 953), \text{CA}) \\
z_8 &= Z(\text{resseq}(954 : 1186), \text{CA})
\end{align*}$$  (5)

We extracted the spatial coordinates of only CA-atoms from the protein using equation 5, because as long as c-alpha atom
is visible, the whole backbone of the protein is visible [49]. The coordinates are translated and rotated. We, first find the centroid by averaging the coordinates and re-center them to remove the translation. The optimal rotation is found using the singular value decomposition which minimizes the least square function [50]. The scripts are written in MATLAB 2018b, using molecular dynamics toolbox [51] and RStudio using Bio3D package [52].

IV. ANALYSIS

There are various programs for the analysis of MD trajectories, such as MDanalysis [53] and CPPTRAJ [45]. They offer a number of functions for the analysis of MD trajectories such as RMSD, RMSF and radius of gyration. However, only a limited analysis can be performed using these programs. In this paper, we use time series models to analyze the MD trajectories.

First, we perform spectral analysis inspired by the spectral speech analysis. We tested both parametric and non-parametric models power spectral density estimation. In the non-parametric models, periodogram [54] is used to estimate the power spectral density [55]. However, results were difficult to analyze as periodogram has low resolution and introduce biases in the spectrum estimation. We also used multitaper [56] method and welch method [57], but the results were hard to interpret as shown in Figure 9.

Parametric models have high frequency resolution and avoid spectral leakage [58]. In this paper, we use autoregressive integrated model based power spectral density estimation on each domain of EGFR and its mutant structures. Parameters of the model are estimated using the forward linear prediction method. Simulation results show useful insight.

To see the relationship between different domains of EGFR and its mutants, we calculated the correlation between them. Pearson correlation coefficient is used due to linearity among the variables. Simulation result shows mutant structures are less correlated, as shown in Figure 7.

Dynamic time warping is applied as our final analysis method to highlight the differences in warping patterns of each domain of EGFR and its mutant structures. DTW has a quadratic time and space complexity and it is mostly used for small time series analysis, i.e., a few thousand points [59]. There are other modified and faster versions of DTW [60], [61] as well, which can be used for the comparison of long time series. Each domain of EGFR contains a few thousands of point, so DTW is computed easily for our dataset. All these methods are implemented in MATLAB 2018B.

A. AUTOREGRESSIVE INTEGRATED MODEL

In an autoregressive integrated model, the value from a time series is regressed upon previous values. The AR model is similar to the Fourier transform [62], which is used to estimate the spectrum of a time series signal. The AR model has higher resolution and stable power spectra than the DFT [63], [64].

Let Eq. 6 represent the coordinates of MD trajectories. We first divide the entire structure into eight domains and use CA atoms of each domain. Each domain is considered as a time signal and ARI model is applied to it.

\[
Z = [z(\text{CA}1) \ldots z(\text{CA}8)]^T
\]

(7)

Equation 8 represents a general p order AR model. We have used a differencing function to remove the trends and stabilize the mean of equation 8. Differencing is a popular method to make the time series stationary [19], [65].

\[
z(n) = a_0 z(n-1) + a_1 z(n-2) + \ldots + a_p z(n-p) + \delta
\]

(8)

where \(z(n)\) is the differenced sequence.

In equation 8, \(p\) and \(a\) represents the order and coefficients of the ARI model and \(\delta\) is the estimation error which is assumed to be Gaussian. \(p\) is chosen as \(n/2\) [66], [67], where \(n\) is the length of the sequence. The ARI model in the compact form can be written as 10.

\[
z(n) = Za + \delta
\]

(10)
FIGURE 7. Pearson correlation coefficient among domains of EGFR (a) Wildtype (b) L858R Mutated type (c) L858R-T790M mutated and resistive type. The blue circle shows positive correlation and red circle shows negative correlation. The color intensity shows the strength of correlation.

and $\delta$ is the column vector of estimation error. There are several ways to estimate the prediction coefficients such as Yule walker method [47], maximum likelihood estimation [68] and forward linear prediction [69]. In this paper, we use linear prediction model to estimate the coefficients, which minimizes the prediction error in the least square sense.

$$A_n = Z^{-1} \ast z + \delta_n$$  \hspace{1cm} (11)

Matrix $Z$ should be full rank and symmetric to guarantee the inversion. $\delta_n$ is the estimation error vector, $A_n$ is the set of prediction coefficients. The information lies in the AR coefficients. We use these coefficients to calculate Power Spectral Density (PSD) using Eq. 12.

$$P_{ARSD} = \frac{\sigma^2}{1 + \sum_{i=0}^{N} a_k \ast \exp(-jwk)^2}$$ \hspace{1cm} (12)

Here $\sigma^2$ represents the variance and $w$ is the normalized angular frequency. We also calculated the center, upper and lower frequency of the spectral density using 13.

$$\begin{align*}
Uf &= t > \max(t)/2 \\
Lf &= t < \max(t)/2 \\
Cf &= \sqrt{(Uf \ast Lf)}
\end{align*}$$ \hspace{1cm} (13)

Here $Uf$ and $Lf$ represents the frequency above and below the half power respectively and $Cf$ represents the center frequency. The shift in the center frequency is used to analyze the motion of atoms [70], [71].

B. DYNAMIC TIME WARPING

Dynamic time warping (DTW) is a popular method for measuring the similarity between time series signals, which may vary in speed and direction. DTW has been previously applied to compare different speech patterns, data mining and
information retrieval [72], [73]. DTW has also been used to automatically cope with time deformations and different speed associated with time series data in many real world application [74]–[76].

DTW finds an optimal match between two time series by allowing a non-linear mapping and minimizing the distance between the sequence [77]. A recent paper is very pertinent in this regard [78]. We use DTW to analyze the motion patterns of each domain.

Let us assume \( D_1 = (x_1, x_2, x_3 \ldots x_n) \) and \( D_2 = (y_1, y_2, y_3 \ldots y_m) \) represents two domains of the EGFR, containing n and m samples of CA atoms respectively. In the first step, a cost matrix is created using the following equation 14.

\[
D(i, j) = \text{Dist}(i, j) + \min \begin{cases} 
D(i - 1, j) \\
D(i, j - 1) \\
D(i - 1, j - 1)
\end{cases}
\]

Here \( D(i, j) \) represents the distance between \( i \) and \( j \) atoms of domains 1 and 2 respectively, the second term is the minimum distance from the nearest neighbors. In this way, we construct a \( M \times N \) cost matrix.

In the second step, we construct the warp path by using the cost matrix. The steps for the warp path is given below.

- Construction of warp path \( W (W_1, W_1, W_2, \ldots, W_n) \) using the cost matrix
- From \((1, 1)\) to \((m, n)\)
- \(i \) and \(j \) must be monotonically increasing
- \( w_{k} = (i, j) = w_{k+1} = (i', j') \) if \( i' \leq i + 1, j' \leq j + 1 \)
- Move upward, right or diagonal
- Backtracking and greedy search

Warping path represents the movement of atoms in each domain. The warping paths are compared in terms of dynamic distance. Warping path distance is calculated using

FIGURE 8. Power spectral densities of 8 domains of EGFR and its mutants structures. For a more precise representation, tabular values are also shown in Table 4.

TABLE 1. Analysis of Peaks, widths and center frequencies. There is a shift in the center frequency of each domain. Also the number of peaks are decreased in mutated and resistive types, specially in the kinase domain and domain 4. The maximum width is obtained in mutated type.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Domain 1</th>
<th>Domain 2</th>
<th>Domain 3</th>
<th>Domain 4</th>
<th>Domain 5</th>
<th>Domain 6</th>
<th>Domain 7</th>
<th>Domain 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Peaks</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Maximum Peak</td>
<td>86.0071</td>
<td>64.3762</td>
<td>69.3412</td>
<td>62.0354</td>
<td>74.4387</td>
<td>74.2985</td>
<td>63.2218</td>
<td>71.6852</td>
</tr>
<tr>
<td>Minimum Peak</td>
<td>58.80073</td>
<td>33.6223</td>
<td>31.5432</td>
<td>21.5268</td>
<td>53.1169</td>
<td>29.9600</td>
<td>33.3732</td>
<td></td>
</tr>
</tbody>
</table>

Mutated Type

| Center Frequency | 7.0711   | 8.7178   | 7.5498   | 8.8154   | 8.6023   | 10.3441  | 7.5498   | 7.000    |
| No. of Peaks    | 5        | 6        | 7        | 6        | 7        | 6        | 7        | 6        |
| Maximum Peak    | 81.7580  | 70.4224  | 79.8705  | 71.3277  | 69.8388  | 62.4231  | 82.8298  | 80.7281  |
| Minimum Peak    | 37.8458  | 36.0325  | 32.1044  | 29.4513  | 23.0416  | 28.3897  | 25.1045  | 30.8331  |

Mutated and Resistive Type

| No. of Peaks    | 5        | 6        | 7        | 6        | 6        | 4        | 7        |
| Maximum Peak    | 74.1132  | 65.8745  | 84.7964  | 76.2841  | 58.2577  | 72.3576  | 63.6026  | 82.7039  |
| Minimum Peak    | 37.2264  | 35.5124  | 32.9306  | 34.1209  | 19.6692  | 25.8266  | 31.9309  | 29.2417  |

\[ \text{Dist}(w) = \sum_{k=1}^{L} \text{Dist}(w_{ki}, w_{kj}) \]  

C. PEARSON CORRELATION COEFFICIENT

Pearson correlation coefficient is a useful method to find the linear correlation between two time series [79]. Let us assume domain 1 of protein X contains n samples \((x_{11}, x_{12}, x_{13}, \ldots, x_{1n})\) and domain 2 of protein Y also contains n samples \((y_{11}, y_{12}, y_{13}, \ldots, y_{1n})\) are represented as time series signal, then the Pearson Correlation between the two domains of protein is defined as 16.

\[ r = \frac{\sum_{i=1}^{N} x_{i} y_{i} - \left( \sum_{i=1}^{N} x_{i} \right) \left( \sum_{i=1}^{N} y_{i} \right)}{\sqrt{\left( \sum_{i=1}^{N} x_{i}^{2} \right) \left( \sum_{i=1}^{N} y_{i}^{2} \right) - \left( \sum_{i=1}^{N} x_{i} \right)^{2} \left( \sum_{i=1}^{N} y_{i} \right)^{2}}} \]  

The range of correlation is \([-1, 1]\). When \(r = 1\), X and Y are completely correlated, \(r = -1\), X and Y are completely opposite of each other and \(r = 0\) reflects no correlation, X and Y are independent of each other. The higher the value of \(r\), the greater is the correlation between X and Y. The correlation results are shown in Figure 7. In our scenario, positive correlation relates to the movement of atoms in the same direction and overall structure of the protein is stable and easier for the drug to bind. The negative correlation reflects the movement of atoms in opposite directions, reflecting that the structure of the protein is destabilized and difficult for the drug to bind. This information can be useful for dealing with the drug resistance problem.

V. RESULTS & DISCUSSION

Domains 1-to 4 have a strong positive correlation in all the three types of structures, which shows that the mutation does not have much effect on these regions of EGFR. However domain 4 is less correlated in mutant structures compared to WT as shown in Figures 7-b and c.
Domains 5-to 8 have a strong negative correlation in mutant structures, mutation is applied in this part of the EGFR and this is most affected region. The mutation effects the domains of protein and they become less correlated especially in domain 7, Figure 7-b and in domain 8, Figure 7-c. Negative correlation shows that atoms move in the opposite directions and the domains are less stable, Hence it is difficult for a drug to bind the protein and stop the downstream signaling. This can be one of the reasons for drug resistance.

The power spectral densities for EGFR and its mutant structures using parametric methods are shown in Figure 8. The non-parametric methods are also shown in Figure 9. In Figure 8, the peaks are shifted and distorted in mutated type and even more in mutated and resistive type, especially in the kinase domain, domain 4 and domain 8. There is not much difference until domain 3, but still a frequency shift can be observed. In domain 4, there is a clear shift in the third peak. Kinase domain is affected the most and peaks get distorted in the mutated type and even more distorted in the mutated and resistive type. Detailed information about power spectral density curves is given in table 1. The sub-domain 4 of extracellular region is also affected. The mutant structures are less stable than the WT.

Warping patterns for each domain of EGFR and its mutant sequence are shown in Figure 10. The Patterns are almost similar for domains 1 - 4, 6 and 7. There is a visible difference in domain 8 of protein when the first mutation occurs. The distance between domain 1, 4 and 8 decreases in mutated type and they come close to each other in mutated and resistive type. Domains 5 and 6 also have different patterns. Domain 8 (tail) suddenly starts to rise when the first mutation happens.

We have also listed the warping path distance in Table 2. All the values are in nm. The distance increases in kinase domain for the mutated and resistive type respectively. The warping patterns can be useful to understand the motion of each domain of EGFR and its mutant structures and for computer aided drug design for NSCLC patients.

VI. CONCLUSION AND FUTURE WORK

In this paper, we investigated the behavior of the wildtype EGFR and its mutants. We generated 200 frames of MD simulation for EGFR and its mutant sequences using parallel computing and AMBER tools. We used power spectral densities, dynamic time warping and the Pearson correlation coefficient to analyze the motion of EGFR and its mutant sequences. Strong correlation is observed in the wildtype EGFR, less in mutated type and even weaker in the mutated and resistive type. Hence, the domains are more stable in wildtype EGFR, less stable in mutated EGFR and even weaker in mutated and resistive type. This can be assumed as, under no mutation the atoms are tightly coupled and it is easy for the drug to bind the protein. The power spectral density also shows similar results. We also showed that mutation in the kinase domain also affects the other domain as well and overall the structure is less stable. Earlier studies on the kinase domain also verify that mutant EGFR structures are less stable than the wildtype [24], [36], [80]. Hence, the mutation destroys the atoms/residues and it is difficult for the drug to bind the mutated and resistive EGFR. This may be one of the reasons for drug resistance.

This study can help us in understanding drug resistance mechanism and designing new treatment strategies for NSCLC patients. Insights into EGFR trajectories can provide the clue to design future therapies for NSCLC patients. We also hope that this article will encourage signal processing specialist to contribute to this challenging area. The rapid
advancement in the field of machine learning [81] and its combination with molecular dynamics can be integrated to create accurate useful applications [15].

The Deep neural networks can learn the atomic trajectories and find suitable descriptors, which can be used as an input. A recently developed tool at Princeton University DeePMD [82] is pertinent in this regard. It is now possible to visualize and run long MD simulations [83] using GPUs that can help doctors to visualize protein-drug interactions. Computational methods have a great potential in designing new drugs for NSCLC patients and bring a significant contribution to the society.

ACKNOWLEDGMENT

The authors would like to thank Dr. Sheheryar Khan and Mr. Ali Raza Shahid for their valuable support and suggestions.

DATA AVAILABILITY

The MD simulation data and relevant codes that supports this study is available from the corresponding author upon reasonable request.

REFERENCES


RIZWAN QURESHI received the bachelor’s degree in electronic engineering from the Mehran University of Engineering & Technology, Jamshoro, in 2010, and the M.Sc. degree in electrical engineering from the Institute of Space Technology, Islamabad, Pakistan, in 2015. He is currently pursuing the Ph.D. degree in electrical engineering with the City University of Hong Kong. Before joining City University, he was a Lecturer with the Electrical Engineering Department, COMSATS University Islamabad, Wah Campus, Pakistan. His research interests include bioinformatics, computer vision, signal and image processing, energy management, and spectral imaging.

MEHMOOD NAWAZ received the B.S. degree from Bahauddin Zakariya University, Pakistan, in 2012, and the M.Sc. degree from Shanghai Jiao Tong University, China, in 2017. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, City University of Hong Kong. His research interests include image processing, graph matching, pattern recognition, and facial expression analysis.

AVIRUP GHOSH received the B.Tech. degree in electronics and communication engineering from the West Bengal University of Technology, in 2013, and the M.Tech. degree in mechatronics from the Vellore Institute of Technology, Vellore, in 2015. He is currently pursuing the Ph.D. degree with the Electrical Engineering Department, City University of Hong Kong. He is working on hydrogen bond analysis in lung cancer and drug resistance mechanism. His research interests include genomic data analysis, structural biology, and gene mutation for lung cancer.

HONG YAN received the Ph.D. degree from Yale University. He was a Professor of imaging science with The University of Sydney. He is currently a Chair Professor of computer engineering with the City University of Hong Kong. His research interests include image processing, pattern recognition, and bioinformatics. He has more than 600 journal and conference publications in these areas. He is an IAPR Fellow. He received the 2016 Norbert Wiener Award from the IEEE SMC Society for contributions to image and biomolecular pattern recognition techniques.