Survival prediction for lung cancer patients by integrating clinical and molecular features using machine learning

Rizwan Qureshi, City University of Hong Kong, Hong Kong

Abstract— **Lung cancer has the highest mortality among all cancer types. Accurate prognostic stratification is clinically important for designing treatment strategies for lung cancer patients. In this work, we propose a model by integrating clinical and molecular features using machine learning classifiers and molecular dynamics simulation. The proposed model achieves good accuracy with a random forest classifier. We believe that the prediction can be a promising index and may help physicians and oncologists to develop personalized therapies for lung cancer patients.**

Index terms: Cancer, Machine learning, Precision medicine

I. INTRODUCTION

Cancer is a major health problem worldwide, resulting in a loss of 9.6 million lives and 18.1 million new cases in 2018. Among all cancer types, lung cancer has the highest mortality rate, because it is diagnosed at advanced cancer stages. In the last decade there has been a great progress in the management of lung cancer patients, and epidermal growth factor receptor (EGFR) has been identified as useful biomarker. Several generation of EGFR targeted drugs has been developed to for lung cancer patients [1]. These drugs produced promising results at the initial stage, but the efficacy becomes limited due to the appearance of drug resistance. Therefore predicting the outcome of patient's survival is crucial for doctors to design therapeutic strategies [2], [3].

The completion of human genome project has allowed a move from the traditional medicine model, as one size fits all approach towards personalized therapies. The patient's genomic and genetic information explores new opportunities for treatment, care, and diagnosis [7].

Debby *et al* [1] combined clinical information and protein-drug interaction to predict the drug resistance prediction model. Ma *et al* used geometric features and energy fearures of EGFR-mutated lung cancer patients to predict the two class drug response. Qureshi and Yan investigated the stability of domains of EGFR mutants [3]. A recent review discusses challenges in drug response prediction using machine learning [4].

In this work, we present a model that combines patient's clinical information and the binding energies of protein-drug interactions. The proposed model achieves good accuracy, even with a small dataset.

II. PROPOSED MODEL

The clinical data used in this study is taken from various studies [1], [2]. The clinical information includes patient's age, sex, smoking history, progression status, mutation status, and five year survival. The mutation status is translated into 3D models of the structures. All these values are normalized between zero and one using min-max feature normalization technique.

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)} \tag{1}$$

A. Molecular dynamics simulation

Molecular dynamics (MD) simulation is a computational method to analyze the interactions between drug and the molecule at atomic scale. In this work, we performed MD simulation for EGFR mutants using Amber software. First, we perform computational modeling of the structures using Rosetta protocol. The simulations are performed using 14 parallel processor, each having 8 GB RAM. We also used NVIDIA C2075 GPU for the preparation of production files. Each program completed in 14 hours on our system with optimized code.

B. Binding free energy

The binding free energy is the measure of the drug binding affinity in a solvent environment. The MM-GBSA protocol in Amber is used to estimate the binding energy. The binding free energy is calculated based on the theory of thermodynamics.

$$\Delta G = \Delta G_{Bind,Vacuum} + \Delta G_{Solv} \tag{2}$$

ΔG shows the free energy difference between two distinct states. ΔG$_{Bind,Solv}$ and ΔG$_{Bind,Vacuum}$ corresponds to the free energy difference between the bound and unbound states of a complex in solvent and vacuum respectively. The total binding free energy is used as a feature in this work.

C. Classification

Given a patient's with clinical information, we extract binding energies for the mutant sequences. The feature values are normalized using min-max normalization. CARET package in Rstudio is used to classify the patients as alive (Yes) or dead (No). The proposed model is shown in Figure 1. Figure 2 shows the binding free energy and survival and response level. From the Figure, it is clear that the survival rate is not linearly related to the binding energies. This indicates the influence of personal features. The density distribution for features are shown in Figure 3. Six features including age, sex, smoking history, survival time, progression time

	Reference	
Prediction	Yes	No
Yes	14	9
No	6	11

TABLE I
CONFUSION MATRIX ON TESTING DATA USING CART CLASSIFIER

and Total energy are used to predict the survival of patients.

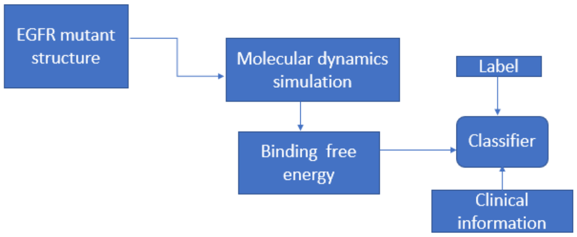

Fig. 1. The proposed model. First, we perform MD simulation for EGFR mutants and extract the binding energies between drug and the target. After that, clinical information and binding free energies are fed to the classifier to predict the survival of patients.

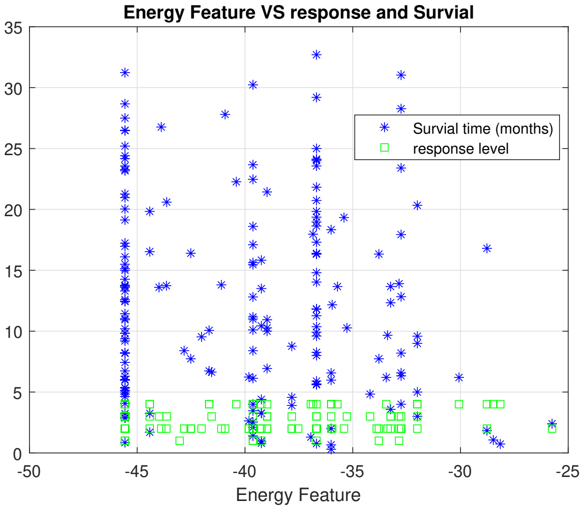

Fig. 2. Energy features vs response level and survival time in months. The drug response has four types(1,2,3,4). The lower values indicates the higher drug response.

III. RESULTS AND DISCUSSION

We used energy and clinical features to predict the five year survival for lung cancer patients. The dot results are shown is Figure 4. The Random forest classifier achieves highest accuracy. The confusion matrix is shown in Table 1. The dataset consisted of 168 patients who are treated with EGFR-targeted drugs. We divide the data into 80% and 20% as training and testing sets. The results indicates that the combination of energy and clinical features can be a good strategy to predict the survival of the patients, that may help doctors in taking optimal treatment decisions. Due to advancement in the computational speed, and the rise of deep learning, there is a growing interest in the field of personalized medicine [5], [6].

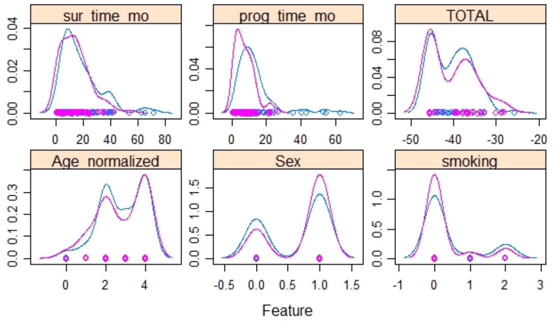

Fig. 3. Density distribution of clinical and energy features.

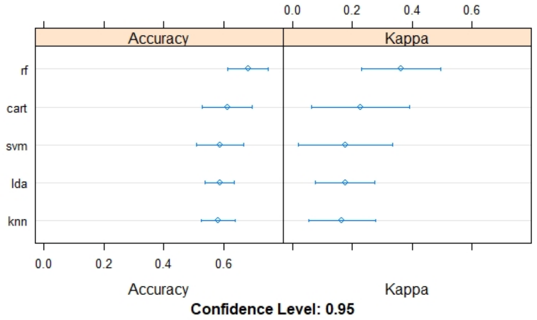

Fig. 4. Accuracy for different classifier

In the future, we will combine other features such as shape of the drug binding pocket to increase the prediction rate. We will collect more clinical data to further refine the prediction model. Other classifiers such as deep neural networks, graph neural networks will be tested to deal with the difficult learning problem.

REFERENCES

[1] Wang DD, Zhou W, Yan H, Wong M, Lee V. Personalized prediction of EGFR mutation-induced drug resistance in lung cancer. Scientific reports. 2013 Oct 4;3(1):1-8.

[2] Ma L, Wang DD, Zou B, Yan H. An eigen-binding site based method for the analysis of anti-EGFR drug resistance in lung cancer treatment. IEEE/ACM transactions on computational biology and bioinformatics. 2016 May 12;14(5):1187-94.

[3] Qureshi R, Nawaz M, Ghosh A, Yan H. Parametric models for understanding atomic trajectories in different domains of lung cancer causing protein. IEEE Access. 2019 May 22;7:67551-63.

[4] Adam G, Rampášek L, Safikhani Z, Smirnov P, Haibe-Kains B, Goldenberg A. Machine learning approaches to drug response prediction: challenges and recent progress. NPJ precision oncology. 2020 Jun 15;4(1):1-0.

[5] Chen JY, Xu H, Shi P, Culbertson A, Meslin EM. Ethics and privacy considerations for systems biology applications in predictive and personalized medicine. InBioinformatics: Concepts, Methodologies, Tools, and Applications 2013 (pp. 1378-1404). IGI Global.

[6] De Meo P, Quattrone G, Ursino D. Integration of the HL7 standard in a multiagent system to support personalized access to e-health services. IEEE Transactions on Knowledge and Data Engineering. 2010 Sep 23;23(8):1244-60.

[7] Wang F, Zhang P, Dudley J. Healthcare data mining with matrix models. InProceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016 Aug 13 (pp. 2137-2138).

[8] Qureshi, Rizwan, et al. "Computational Analysis of Structural Dynamics of EGFR and its Mutants." 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2019.

[9] Qureshi, Rizwan, Avirup Ghosh, and Hong Yan. "Correlated Motions and Dynamics in Different Domains of EGFR with L858R and T790M Mutations." IEEE/ACM Transactions on Computational Biology and Bioinformatics (2020).

[10] Qureshi, Rizwan, Muhammad Uzair, and Khurram Khurshid. "Multistage adaptive filter for ECG signal processing." 2017 International Conference on Communication, Computing and Digital Systems (C-CODE). IEEE, 2017.

[11] Khan, S., Anwar, S. M., Abbas, W., Qureshi, R. (2016). A novel adaptive algorithm for removal of power line interference from ecg signal. Science International, 28(1), 139-143.