**APPLICATION OF MACHINE LEARNING TECHNIQUE ASSISTED PREDICTION OF BIOACTIVITIES OF LIGANDS IN TARGETED DRUG DISCOVERY PROCESS OF LUNG CANCER FOR EGFR TARGET**

Natthakan Saengnil1, Puri Virakarin1, Thanasan Nilsu2, Bundit Boonyarit3

*1. Kamnoetvidya Science Academy*

*2. Department of Biology and Environmental Science, Kamnoetvidya Science Academy*

*3. School of Information Science and Technology, Vidyasirimedhi Institute of Science and Technology*

*Rayong, Thailand*

nuttagun\_s @ kvis.ac.th / puri\_v @ kvis.ac.th

Nowadays, more Asian people have been diagnosed with lung cancer. Among diseases in the cancer group, lung cancer was reported as the world's highest incidence in 2018. Lung cancer is one of the most common diseases in Thailand. Lung cancer is a chronic non-communicable disease caused by abnormal cells in the lung tissue that proliferate uncontrollably. One of the leading causes of lung cancer in non-smokers is the amplification or hyperactivation of Epidermal growth factor receptor (EGFR) proteins. However, the conventional laboratory of drug discovery and development processes often encounters resource and time limitations. Challenges in pharmaceutical industries include a lengthy period of 15-20 years for developing new drugs with over 2,000 million dollars expense but resulted in drug failures due to its effectiveness or safety issue during research development. Our achievement used machine learning techniques to compare Random Forest (RF), Convolutional Neural Network (CNN), and Convolutional Spatial Graph Embedding Network (C-SGEN) with Deep Neural Network (DNN) algorithms. However, the mutation of EGFR gene can lead to a buildup of drug resistance towards the existing drug targeting the wild type EGFR. This scenario would require new drugs that can effectively inhibit the mutant EGFR. In this investigation, we developed a model to predict pIC50 of ligands for kinase inhibitors with large-scale and non-redundant datasets. This work can be applied in the discovery and development processes of targeted drugs for lung cancer, focusing on wild type and mutant EGFR targets. Our findings revealed that the single task learning model with transfer learning and the multitask learning model with hyperparameter tuning presented high predictive performance. The proposed prediction model provides a powerful strategy that potentially helps researchers overcome a problem in drug discovery and development processes and mutation of EGFR, leading to increased treatment options and reducing the risk of lung cancer patients.

Web service: <https://ligegfr.vistec.ist>

**ПРИМЕНЕНИЕ ТЕХНИКИ МАШИННОГО ОБУЧЕНИЯ ПРИ ПРОГНОЗИРОВАНИИ БИОАКТИВНОСТИ ЛИГАНДОВ В ЦЕЛЕВОМ ПРОЦЕССЕ РАЗРАБОТКИ ЛЕКАРСТВЕННЫХ СРЕДСТВ ПРОТИВ РАКА ЛЕГКОГО ПРИ EGFR В КАЧЕСТВЕ ЦЕЛИ ИНГИБИРОВАНИЯ**

Саенгнил Н., Виракарин П.

*Академия Наук Камноетвидия, Районг, Таиланд*

*nuttagun\_s @ kvis.ac.th / puri\_v @ kvis.ac.th*

Белки семейства рецепторов эпидермального фактора роста (EGFR), состоящие из HER1, HER2, HER3 и HER4, являются одной из наиболее распространенных причин различных видов рака у человека, в том числе рака груди, желудка, пищевода, легких, поджелудочной железы и мочевого пузыря [1, 2]. Кроме того, мутация гена HER1 может вызвать серьезную проблему лекарственной устойчивости к существующему лекарству, нацеленному на HER1 «дикого типа» [3]. Этот сценарий требует разработки новых лекарств, способных эффективно ингибировать белки семейства EGFR.

Метод. В данном исследовании мы разработали модели для прогнозирования pIC50 лигандов для ингибиторов киназ с наборами данных, содержащими HER1, мутантный HER1, HER2, HER3 и HER4. Модели были основаны на модели LigEGFR, сверточной сети встраивания пространственного графа (C-SGEN) с алгоритмами глубокой нейронной сети (DNN) [4]. Эксперименты проводились путем тестирования средних прогностических характеристик с помощью RMSE моделей, разработанных на основе сочетания методов многозадачного обучения [5] и переноса обучения [6]. Модели были проверены на надежность с помощью метода y-скремблирования, служащего для определения устойчивости моделей.

Результаты. Это исследование показало, что однозадачный метод с методом переноса обучения и многозадачный метод обучения с методом настройки гиперпараметров продемонстрировали высокую прогностическую эффективность. Кроме того, благодаря тестированию с использованием специальных методов проверки модели показали свою надежность при прогнозировании. Предложенные нами прогностические модели представляют собой мощную стратегию, которая потенциально поможет исследователям открыть новые лекарства против белков семейства EGFR. Более того, эти модели также решают серьезную проблему виртуального скрининга лекарств с использованием методов машинного обучения с помощью ограниченных по объему экспериментальных наборов данных.