[P34] Development of QSAR classification models for identifying EGFR inhibitors using machine learning approaches

Subhash M Agarwal

Bioinformatics Division, Institute of Cytology and Preventive Oncology, Noida-201301, India

Introduction:
Epidermal Growth Factor Receptor (EGFR) is a well studied clinically established Lung cancer drug target. In past, several QSAR models have been developed that consider few molecules of similar nature identified using a single bioassay system. As previously developed models have limited coverage, it has become necessary to develop model that considers heterogeneous dataset of molecules covering broad chemical space so that the pace of EGFR inhibitor drug discovery is accelerated.

Objective:
1.To develop a literature curated database of small synthetic inhibitors of EGFR.
2.To develop global QSAR model using a large dataset of structurally diverse molecules.

Material and Methods:
A collection of ~3500 anti-EGFR inhibitors with diverse structural scaffolds that used different experimental assay protocols was obtained from the EGFRIndb database developed from our group. Machine learning methods including IBK, Naive Bayes, SVM and Random forest were then used to construct QSAR models.

Results:
We have developed EGFRIndb, a literature curated database of EGFR inhibitors consisting of 4581 compounds having in vitro inhibitory activities against EGFR or its different isoforms. For each compound, database provides information on structure, experimentally determined inhibitory activity of compound against kinase as well as various cell lines, properties (physical, elemental and topological) and drug likeness.
We then developed and evaluated classification model using various algorithms in Weka package and SVM light. Models were developed using fivefold cross validation and 881 PubChem FP (fingerprints). It was observed that Random Forest algorithm performed best among various classifiers and achieved 80.2% sensitivity, 77.2% specificity and 77.6% accuracy along with 0.44 MCC. We also developed a user-friendly web-server implementing the algorithm for prediction of newer anti EGFR molecules.

Conclusions:
We have designed, developed and validated a single QSAR model that accounts for heterogeneous data i.e. diverse structural scaffold and different experimental assay. This highly accurate prediction models would be useful to design and discover novel EGFR inhibitors.

References:
1.QSAR based model for discriminating EGFR inhibitors and non-inhibitors using Random forest. Biol Direct. 10 (2015) 10.
2.Agarwal SM et. al. *Anticancer Agents Med Chem*. 14 (2014) 928-35.