International Journal of Mathematics and
Computer Science, **17**(2022), no. 1, 461–468

Prediction of Estrogen Receptor alpha Antagonists Using Deep Neural Network

Sara Salah Mohamed

Department of Mathematics
Faculty of Science
New Valley University
Assiut, Egypt

email: dr.sarasalah799@gmail.com

(Received August 31, 2021, Accepted September 28, 2021)

Abstract

Breast cancer is one of the most common diseases whose seriousness has raised concern and anxiety worldwide with over 7 million cases and around 685 thousand deaths globally (according to the World Health Organization(WHO)), making it one of the world's most prevailing diseases. Scientists and doctors tried to cure breast cancer disease using many types of therapeutic treatment, the most common of them being Endo-therapy. Unfortunately however, Endo-therapy did ot have much better results than the other types of treatments. Some breast cancer cells have estrogen receptors which can attract estrogen proteins and make a cancer cell grow. In this study, we introduce deep neural networks on QSAR model to find inhibitors for those estrogen receptors and thus stop the breast cancer cells from growing. We present a comprehensive comparison between our model (the DNNR model) and the Random forest model, in which the DNNR outperformed the RF algorithm (an increase of 21.2% in the accuracy of the algorithm).

1 Introduction

The female sex hormones (estrogens) have an important role in the menstrual cycle and the development of the female characteristics. More extensive research, though, made scientists find out that estrogens can affect the

Key words and phrases: Estrogen Receptor Alpha, Deep Learning, Breast Cancer, Drug Design, Chemoinformatics.
AMS (MOS) Subject Classifications: 92C40, 68T07, 92-08.
ISSN 1814-0432, 2022, http://ijmcs.future-in-tech.net

oxidative stress and that is by preventing the ROS (reactive oxygen species) from generating. Some breast cancer cells need estrogen to grow. These cancer cells have special proteins inside called estrogen receptors (ER). When estrogen hormones attach to estrogen receptors, the cancer cells with these receptors grow. About 65-75 percent of breast cancers are positive for estrogen receptors. ER regulates the signaling of estrogens and their biological effects. ER is divided into two subtypes: ERα and ERβ. In normal cells, ERα is expressed in only a small portion of the cell (10-20 percent in mammary glands), while ERβ is expressed in a larger portion of the cell (80-85 percent in mammary glands)[1]. However, in breast cancer cells ERα is expressed more and ERβ is expressed less, and this is one of the reasons ERα is used as an indicator of breast cancer cells. Estrogen receptor-positive breast cancers can be treated with hormone therapy drugs. ERα is used as a therapeutic target for those drugs in breast cancer cells. Therapeutic treatment is done by blocking the estrogen receptors' transcription from the body. Antiestrogens are agents that can obstruct the production of estrogen in breast cancer cells. Examples of Antiestrogens are Tamoxifin and Fulvestrant. The main problem, though, is the development of resistance in breast cancer cells to antiestrogen agents. Approaches using computations can be of a lot of help at this subject and that is because trying to find new treatments is expensive and time consuming and needs a lot of labor. As a result, we try to use the virtual screening technology to help us find one [2]. Virtual screening has a lot of types inclusing Ligand based drug design (LBDD) and structure-based drug design (SBDD). LBDD is related more to machine learning and the quantitative structureactivity relationship (QSAR) of the drug [3], while SBDD is related more to the molecular structure of the drug [4]. In the last decade, deep learning became a very important topic for solving a lot of problems and challenges in many fields including the chemoinformatics, bioinformatics, system biology. The main contribution of this paper is utilizing deep learning neural network on QSAR model to predict inhibitor drugs (chemical compounds) for ERα. The dataset was taken from ChEMBL database and much preprocessing were done to curate the data and encoding it to be prepared for learning and testing with the deep neural network. The rest of this paper is organized as follows: Section 2 introduces a literature review. Section 3 presents materials and methods. Section 4 introduces experimental results. Finally, conclusions are presented in section 5.

2 Literature Review

Drug design is one of the most challenging processes in the pharmaceutical industry. It is complicated, takes a very long time, and is resource consuming. In this process, the researchers try to find small molecules with desired physicochemical and biological properties from the huge chemical space that is 10^{60} feasible molecules. This large space leads researchers to use computer-aided drug discovery (CADD) to speed up the process of finding the desired molecule. In recent years, the rapid development of computational resources and the massive and enormous amount of biological and chemical database supporting the drug discovery (such as DrugBank, KEGG, ChEMBL) lead to the use of the artificial intelligence algorithms to mine the data and classify it according to common features and pathways. Machine learning and deep learning are subclasses of artificial intelligence. We will briefly outline many studies that involved the use of deep leaning and Machine learning in the drug design field. Aliper et al. [5] trained deep learning neural network on the transcriptomic data to find the pharmacological properties of several drugs and compounds through several biological systems. Arach and Bouden [6] used an ensemble classifier to perform analysis on breast cancer dataset; they found that incorporation between Naive Bayes (NB) and RF gets superior results compared to other classifiers. Nishant et al. [7] was trying to use the deep learning techniques to early diagnose COVID-19 and those techniques were able to speed up the drug development but required more accurate clinical data. Tsou et al. [8] made a comparative study between deep learning neural network and the random forest algorithm to find an inhibitor drug for the triple negative breast cancer in which it proved that both of them can achieve great results and can be adapted to be used in identifying the active and non-active compounds in different diseases. Suvannang et al. [9] utilized random forest algorithm to predict inhibitors for ERα; the best accuracy obtained for the training set was 94% and for the testing set was 73%. The proposed deep learning neural network method attempts to increase the accuracy of predicting ERα active inhibitors and optimize the learning.

3 Material and Methods

3.1 QSAR Model

QSAR plans to create models that can represent the relations between the chemical structures of the molecules and their biological activities. The chemical structure of a certain molecule can be represented by its descriptors: its chemical and physical-chemical properties which can be later used to create a mathematical equation: $y = f(d)$, Where y is a biological property of compound, d is a vector of descriptors, and f is a function that extracts important features. For this function, we can use many AI algorithms such as machine learning algorithms and deep learning algorithms.

3.2 Deep learning

DL is one of the modern and new areas in the drug design field. It has resolved many problems and challenges that were faced by other ML algorithms (for example, healthcare, and image denoising)[10]. Deep learning is an algorithm that was created by resembling the neurons in the human brain and that gives it the ability to learn from the history of the data it collected, in which each neuron can be treated as a feature that can help us classify the complex factors. It has the ability of linear and non-linear algorithms. Deep learning uses a cascade of several layers to filter the data, with each layer using the results from the previous one and that gives it the ability to become more accurate by learning from its previous data. Resembling the neurons in the human brain, each neuron is activated by stimuli from a neighboring neuron. The deep learning algorithm uses the multi layered strategy to help in a lot of different problems.

3.3 Dataset Preprocessing

In this study, the dataset used was extracted from ChEMBL database(chembl_23) which targets human Estrogen receptor alpha (ERα) and it consists of 5809 compounds with 10 936 bioactivity data points. The bioactivity unit used is IC50. We extracted records with ASSAY_TYPE is B. Then many preprocessing steps were followed inspired by [9] where the records with smiles equal NaN and duplicate smiles notation were deleted. The final data set consist of 1231 compounds. The IC50 values were converted to pIC50 allowing the data to be more uniformly distributed, where PIC50=$-\log_{10}(IC50)$.

3.4 Encoding Data and Building the DNNR Model

The Fingerprint descriptors were used to encode the structures of the compounds. As there are many types of fingerprint descriptors, o investigate the performance of our model, we used 12 classes of fingerprint descriptors as shown in Table 2 with the number of descriptors for each class computed using the PaDEL fingerprint descriptors software [11]. After the encoding process, the final dataset was randomly divided into two sets with a ratio of 7:3.

3.5 Model Assessment

In this study, we used the square value of the Pearsons correlation coefficient (R2) and the root mean squared error (RMSE) for both the training and the testing datasets to quantify and express the difference in efficiencies between our model and the RF algorithm. The goal of our study was to construct deep learning model that is capable of predicting the biological activity pIC50 as a function of molecular fingerprints. We used DNNRegressor (DNNR) from the Tensorflow Python API, in which we build a feedforward multi-layer neural network that is trained with a set of molecular fingerprints data in order to perform regression task on similar unseen molecular compounds.

4 Results and Discussions

In this experiment, we made a comparative study between the RF algorithm and our model DNNR, in which DNNR consists of three layers where each layer has [1024, 512, 256] neurons, respectively. We used Adam optimizer with parameters (learning rate=0.001, beta1=0.9, beta2=0.999, epsilon=1e-08). The number of iterations for network training was 20000. Finally, to avoid an overfitting problem we used a dropout regularization technique and set dropout = 0.2. For RF algorithm, we used the same configurations as in [9]. All the experiments were run on a PC with 1.99 GHz Intel core i7 and 8GB RAM. Table 1 represents the results of the two models in both datasets in which the RT2 and the RMSE_T are the values of the training dataset, while the QEx2 and RMSE_E are for the testing dataset. Generally, an acceptable model according to the statistical threshold of Golbraikh and Tropsha [12] is that RT2>60 and QEx2>80. As we can see in Table 1, our model superiorly outperformed the RF algorithm in both datasets. The DNNR algorithm outperformed the RF algorithm with an average increase of around 21.2% in the testing dataset. As we can notice in Table 1, the results indicate that the two best models from both datasets are the KlekotaRoth count(RT2 =0.99, RMSE_T =0.25) and the KlekotaRoth(RT2=0.97, RMSE_T =0.28), in which they give an average increase of 13.5% in the RT2 and an average decrease of 46% in the RMSE_T from the results of the RF algorithm.

Table 1: *

Table 1:THE PERFORMANCE OF RF ALGORITHM AND DNNR FOR 12 FINGERPRINT CLASSES

	RF				**DNNR**			
Fingerprint class	**Training set**		**Testing set**		**Training set**		**Testing set**	
	RT2	**RMSE_T**	**QEx2**	**RMSE_E**	**RT2**	**RMSE_T**	**QEx2**	**RMSE_E**
AtomPairs 2D count	0.93	0.38	0.73	0.53	0.96	0.34	0.80	0.61
AtomPairs 2D	0.85	0.54	0.68	0.62	0.92	0.44	0.83	0.60
CDK fingerprinter	0.87	0.51	0.71	0.56	0.96	0.41	0.85	0.59
CDK extended	0.84	0.55	0.67	0.65	0.93	0.47	0.88	0.58
CDK graph only	0.81	0.60	0.70	0.58	0.93	0.42	0.85	0.60
E-state	0.80	0.63	0.64	0.71	0.89	0.50	0.79	0.67
KlekotaRoth count	0.91	0.41	0.72	0.54	0.99	0.25	0.91	0.49
KlekotaRoth	0.82	0.60	0.70	0.59	0.97	0.28	0.90	0.53
MACCS	0.86	0.52	0.71	0.58	0.96	0.35	0.86	0.59
PubChem	0.84	0.57	0.71	0.56	0.97	0.36	0.82	0.59
Substructure count	0.94	0.34	0.73	0.52	0.97	0.32	0.80	0.64
Substructure	0.87	0.51	0.68	0.63	0.91	0.48	0.86	0.61

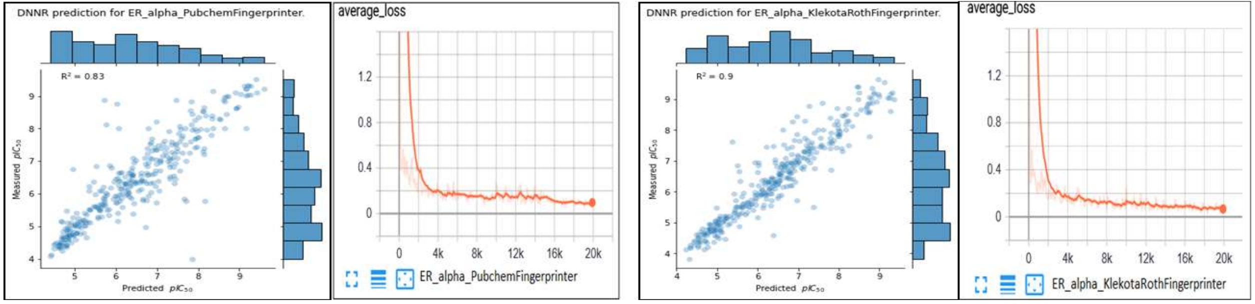
Table 2: *

Table 2: The difference between R square values for training and testing datasets

Fingerprint Class	No of Descriptors	RF RT2 -QEx2	DNNR RT2 -QEx2
AtomPairs 2D count	780	0.20	**0.16**
AtomPairs 2D	780	0.17	**0.09**
CDK fingerprinter	1024	0.16	**0.11**
CDK extended	1024	0.18	**0.05**
CDK graph only	1024	0.11	**0.08**
E-State	79	0.16	**0.10**
KlekotaRoth count	4860	0.19	**0.08**
KlekotaRoth	4860	0.12	**0.07**
MACCS	166	0.15	**0.10**
PubChem	881	0.12	0.15
Substructure count	307	0.21	**0.17**
Substructure	307	0.19	**0.05**

As we can see in Table 2, the fluctuations in the accuracy between the training dataset and the testing dataset in the DNNR is less than that of the RF which means that the DNNR algorithm is more trained to handle new data with better performance and accuracy, performing better than the RF algorithm in all the 12 fingerprint classes except for the Pubchem class. Figure 1 shows the scatter plots of experimental versus predicted pIC50 values for the KlekotaRoth Fingerprint class and PubChem Fingerprint classes. This also represents the average losses versus the number of steps for them, in which it shows that the learning is stable for both of them.

Figure 1: The experimental versus predicted pIC50 values for the KlekotaRoth Fingerprint class and PubChem Fingerprint classes



5 Conclusion

Breast cancer disease is one of the diseases that raised public concern in the last century. ER α and ER β are responsible for this cancer because of their implications in the clinical results. Some types of therapies target both of those ERs, but that often leads to increased risks of breast and endometrial cancers. ERα is the target for the positive breast cancer cells. The laboratory processes to find the inhibitors for the breast cancer cells are expensive not to mention taking a lot of time. As a result, we developed other methods to predict ER endocrine agitation and to simplify their classification achieving surprising results by using our network DNNR on the QSAR model.

References

[1] Xiaocong Pang, Weiqi Fu, Jinhua Wang, De Kang, Lvjie Xu, Ying Zhao, Ai-Lin Liu, Guan-Hua Du, Identification of estrogen receptor α antagonists from natural products via in vitro and in silico approaches, Oxidative medicine and cellular longevity, (2018).

[2] Antonio Lavecchia, Carmen Di Giovanni, Virtual screening strategies in drug discovery: a critical review, Current medicinal chemistry, **20,** no. 23, (2013), 2839–2860.

[3] Bruno J. Neves, Rodolpho C. Braga, Cleber C. Melo-Filho, José Teófilo Moreira-Filho, Eugene N. Muratov, Carolina Horta Andrade, QSAR-based virtual screening: advances and applications in drug discovery, Frontiers in pharmacology, **9,** (2018), 1275.

[4] Antonio Lavecchia, Deep learning in drug discovery: opportunities, challenges and future prospects, Drug discovery today, **24,** no. 10, (2019), 2017–2032.

[5] Alexander Aliper, Sergey Plis, Artem Artemov, Alvaro Ulloa, Polina Mamoshina, Alex Zhavoronkov, Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data, Molecular pharmaceutics, **13,** no. 7, (2016), 2524–2530.

[6] Soumya Arach, Halima Bouden, Performance Analysis on Three Breast Cancer Datasets using Ensemble Classifiers Techniques, Computer Science, **14,** no. 4, (2019), 935–952.

[7] Nishant Jha, Deepak Prashar, Mamoon Rashid, Mohammad Shafiq, Razaullah Khan, Catalin I. Pruncu, Shams Tabrez Siddiqui, M. Saravana Kumar, Deep Learning Approach for Discovery of In Silico Drugs for Combating COVID-19, Journal of Healthcare Engineering, (2021).

[8] Lun K. Tsou, Shiu-Hwa Yeh, Shau-Hua Ueng, Chun-Ping Chang, Jen-Shin Song, Mine-Hsine Wu, Hsiao-Fu Chang et al., Comparative study between deep learning and QSAR classifications for TNBC inhibitors and novel GPCR agonist discovery, Scientific reports, **10,** no. 1, (2020), 1–11.

[9] Naravut Suvannang, Likit Preeyanon, Aijaz Ahmad Malik, Nalini Schaduangrat, Watshara Shoombuatong, Apilak Worachartcheewan, Tanawut Tantimongcolwat, Chanin Nantasenamat, Probing the origin of estrogen receptor alpha inhibition via large-scale QSAR study, RSC advances, **8,** no. 21, (2018), 11344–11356.

[10] Pramila P. Shinde, Seema Shah, A review of machine learning and deep learning applications, Fourth international conference on computing communication control and automation, IEEE, (2018), 1–6

[11] Chun Wei Yap, PaDELdescriptor: An open source software to calculate molecular descriptors and fingerprints, Journal of computational chemistry, **32** , no. 7, (2011), 1466–1474.

[12] Alexander Tropsha, Alexander Golbraikh, Predictive QSAR modeling workflow, model applicability domains, and virtual screening, Current pharmaceutical design, **13,** no. 34, (2007), 3494–3504.