Drug Design, Development and Therapy

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Prediction of selective estrogen receptor beta agonist using open data and machine learning approach

This article was published in the following Dove Press journal:
Drug Design, Development and Therapy
18 July 2016
Number of times this article has been viewed

Ai-qin Niu[1]
Liang-jun Xie[2]
Hui Wang[1]
Bing Zhu[1]
Sheng-qi Wang[3]

[1]Department of Gynecology, the First People's Hospital of Shangqiu, Shangqiu, Henan, People's Republic of China; [2]Department of Image Diagnoses, the Third Hospital of Jinan, Jinan, Shandong, People's Republic of China; [3]Department of Mammary Disease, Guangdong Provincial Hospital of Chinese Medicine, the Second Clinical College of Guangzhou University of Chinese Medicine, Guangzhou, People's Republic of China

Background: Estrogen receptors (ERs) are nuclear transcription factors that are involved in the regulation of many complex physiological processes in humans. ERs have been validated as important drug targets for the treatment of various diseases, including breast cancer, ovarian cancer, osteoporosis, and cardiovascular disease. ERs have two subtypes, ER-α and ER-β. Emerging data suggest that the development of subtype-selective ligands that specifically target ER-β could be a more optimal approach to elicit beneficial estrogen-like activities and reduce side effects.

Methods: Herein, we focused on ER-β and developed its in silico quantitative structure-activity relationship models using machine learning (ML) methods.

Results: The chemical structures and ER-β bioactivity data were extracted from public chemogenomics databases. Four types of popular fingerprint generation methods including MACCS fingerprint, PubChem fingerprint, 2D atom pairs, and Chemistry Development Kit extended fingerprint were used as descriptors. Four ML methods including Naïve Bayesian classifier, k-nearest neighbor, random forest, and support vector machine were used to train the models. The range of classification accuracies was 77.10% to 88.34%, and the range of area under the ROC (receiver operating characteristic) curve values was 0.8151 to 0.9475, evaluated by the 5-fold cross-validation. Comparison analysis suggests that both the random forest and the support vector machine are superior for the classification of selective ER-β agonists. Chemistry Development Kit extended fingerprints and MACCS fingerprint performed better in structural representation between active and inactive agonists.

Conclusion: These results demonstrate that combining the fingerprint and ML approaches leads to robust ER-β agonist prediction models, which are potentially applicable to the identification of selective ER-β agonists.

Keywords: estrogen receptor subtype β, selective estrogen receptor modulators, quantitative structure-activity relationship models, machine learning approach

Introduction

Estrogen receptors (ERs) are nuclear transcription factors and hormone-regulated modulators of intracellular signaling and gene expression.[1–4] There are two subtypes of ERs, ER-α and ER-β. ER-α is encoded by the ESR1 gene on chromosome 6, and ER-β is encoded by the ESR2 gene on chromosome 14.[5] Both ER-α and ER-β are widely distributed in many kinds of cells and tissues, and modulate biological functions in several organ systems, such as endocrine, reproductive, skeletal, cardiovascular, and central nervous systems. ER-α is predominantly expressed in mammary gland, ovary, uterus, male reproductive organs (testes and epididymis), prostate, liver, heart, bone, adipose tissue, vascular system, and brain. ER-β is mainly expressed in

Correspondence: Sheng-qi Wang
Department of Mammary Disease, Guangdong Provincial Hospital of Chinese Medicine, the Second Clinical College of Guangzhou University of Chinese Medicine, Dade Road No 111, Guangzhou 510120, People's Republic of China
Email wsq2011@126.com

submit your manuscript | www.dovepress.com
Dovepress
http://dx.doi.org/10.2147/DDDT.S110603

mammary gland, ovary (granulosa cells), bladder, prostate (epithelium), adipose tissue, immune system, colon, heart, vascular system, lung, and brain.[6,7] The ER-α subtype has a more prominent role in the mammary gland, uterus, the preservation of skeletal homeostasis, and the regulation of metabolism. The ER-β subtype has a more profound effect on the immune and central nervous systems. What is more, ER-β generally counteracts the ER-α promoted cell hyperproliferation in tissues such as breast and uterus.[4,8]

Abnormal ER signaling leads to development of a variety of diseases including osteoporosis. Estradiol replacement therapy is used in the clinic for the treatment of osteoporosis. However, estradiol replacement therapy often leads to an increased risk of breast and endometrial cancers, and thromboembolism due to the ER-α promoted cell hyper-proliferation.[4] Selective estrogen receptor modulators (SERMs) are a class of drugs that act on the ER. A characteristic that distinguishes these substances from pure ER agonists and antagonists (that is, full agonists and silent antagonists) is that their action is different in various tissues, thereby granting the possibility to selectively inhibit or stimulate estrogen-like action in various tissues.[9,10] Following tamoxifen, the first SERM, a number of other anti-estrogens have been developed. Good SERMs would display antagonist activity in the mammary gland and uterus, and agonist activity in cardiovascular, skeletal, and central nervous systems.[9,11,12] Emerging data suggest that ER-β subtype-selective ligands could be used to elicit beneficial estrogen-like activities and reduce side effects.[4,13–15] These results inspired the medical researchers to discover selective ER-β agonists. Roberts et al found sulfonamides as selective ER-β agonists.[16] Paterni et al identified a series of new salicylketoxime derivatives that display unprecedentedly high levels of ER-β selectivity, and one compound was further proved to be active in an in vivo xenograft model of human glioma.[17]

Computational approaches in medicinal chemistry provide important tools for lead discovery and lead optimizations. Machine learning methods are widely applied in computer aided drug design, particularly in the ligand based virtual screening. Zang et al developed binary classification models using a large collection of environmental chemicals from ER assays by quantitative structure-activity relationship (QSAR) and machine learning methods.[18] Ng et al developed a classification model using decision forest to predict environmental chemicals binding to ER.[19] However, previous QSAR studies mainly focused on toxicity or endocrine disruption activity predictions for environmental chemicals. Furthermore, there had been rare reports focusing on the subtype-selective ER agonist prediction.

Owing to the significance of the selective ER-β agonists, as discussed above, we proposed a protocol to predict selective ER-β agonists using a machine learning approach (Figure 1). Due to the difficulty in developing a regression model for a large structural diverse dataset, binary

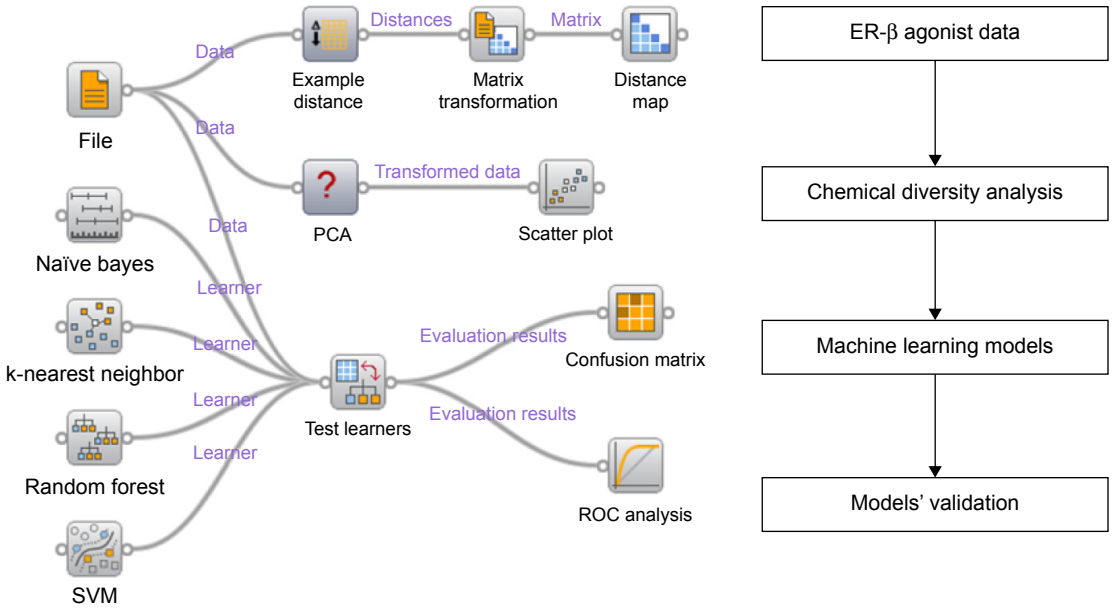

Figure 1 The data analysis and machine learning schema.
Notes: Step 1: collect ER-β agonist data from public database. Step 2: chemical diversity analysis. Step 3: construct machine learning models. Step 4: validate the constructed models.
Abbreviations: ER, estrogen receptor; SVM, support vector machine; ROC, receiver operating characteristic; PCA, principal component analysis.

classification approaches were used here. In this work, we collected a dataset of selective ER-β agonists from an open database (ChEMBL, www.ebi.ac.uk/chembl) and performed the dataset analysis using principal component analysis (PCA) and distance analysis. Then we constructed the prediction models using various molecular fingerprints and machine learning approaches. The accuracies and robustness of the prediction models were further validated, and the performance of the machine learning methods and the molecular fingerprints was compared. These models could be useful in the discovery of selective ER-β agonists.

Materials and methods

Dataset

The ER-β bioactive agonists were downloaded from ChEMBL database (ChEMBL 20 release). Duplicates and salts were removed using Open Babel.[20] Compounds with unclear EC_{50} data, for example <1,000 nM, were removed. The active ER-β agonist was defined as having an EC_{50} less than 10 μM. The inactive agonist was defined as having an EC_{50} more than 10 μM. Finally a dataset was constructed which contained 356 active agonists and 107 inactive agonists. The balancing of the dataset is important for developing a robust model. Machine learning approaches are likely to perform poorly in situations with data imbalance between the classes.[21,22] In order to balance the dataset, we generated a decoy dataset (249 compounds) using the DUD-E online automated tool.[23] Finally, a dataset with 356 active compounds and 356 inactive compounds was obtained.

Molecular fingerprints

Molecular fingerprints are representations of chemical structures originally developed for substructure and similarity searching, but later widely used for descriptors in QSAR studies.[24] Four popular fingerprint generation methods in chemoinformatics including Chemistry Development Kit extended fingerprint (ExtFP, 1024 bits), MACCS fingerprint (MACCSFP, 166 bits), PubChem fingerprint (PubChemFP, 881 bits), and 2D atom pairs (AP2D, 780 bits) were used in this study. All the fingerprints were generated using the PaDEL-Descriptor software.[25]

Naïve Bayesian (NB) classification

The NB classification method is a simple classification method based on the Bayes' theorem as described below:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

The prior probability can be estimated from the training set, while the marginal probability can be ignored. The details of NB classifier building have been described elsewhere.[26,27] NB classification can process large amounts of data, learn fast, and noise data tolerance. The NB classifiers were developed in Orange with default settings. Laplace method was used for probability estimation.

k-nearest neighbor (KNN)

KNN classifier can predict a test sample based on the closest training examples. The nearness is measured by similarity or distance based on vectors in a multidimensional feature space. In the classification process, "k" was a user-defined value, and an unlabeled vector was classified by assigning the label that was most frequent in the k-nearest training samples. The KNN classifiers were developed in Orange using Euclidean distance and the value of "k" was set to three.

Random forest (RF)

The RF was first proposed by Breiman.[28] The RF method is based upon an ensemble of decision trees, from which the prediction of a continuous variable is provided as the average of the predictions of all trees. The advantages of RF in QSAR include high accuracy of prediction, built-in descriptor selection, and a method for evaluating the importance of descriptors in the QSAR model. The details of training procedures are described elsewhere.[29] The RF classifiers were developed in Orange and the number of trees in forest was set to ten, splitting was stopped in RF with nodes of five or fewer instances.

Support vector machine (SVM)

SVM is a general data modeling methodology, originally developed by Vapnik, aimed at minimizing the structural risk under the frame of Vapnik Chervonenkis theory.[30] The principle of SVM is to find a hyperplane in a high dimensional space to separate the positives and negatives.[31] In this work, the radial basis function kernel function was used and the cost was set to 1.00. SVM models were developed using Orange.

Model validation

Cross validation method was employed to test the model performance and robustness. In 5-fold cross validation, the dataset was divided into five subsets, four subsets were chosen as training sets which left one subset as test set in each run. After five runs, each subset was used as test set and the entire dataset was predicted. The quality of the model was

evaluated by the quantity of true positives (TP), true negatives (TN), false positives (FPos), and false negatives (FN). Then the sensitivity (SE), the specificity (SP), the classification accuracy (CA) and the Matthews correlation coefficient (MCC) were calculated using the following equations. Furthermore, the receiver operating characteristic (ROC) curve was plotted and the area under the ROC curve (AUC) was calculated. The values of AUC range from 0–1.0, and 1.0 indicates a perfect model, 0.5 indicates a random model, and >0.8 indicates a good model.

$$\begin{aligned} SE &= \frac{TP}{TP+FN} \\ SP &= \frac{TN}{TN+FPos} \\ CA &= \frac{TP+TN}{TP+TN+FPos+FN} \\ MCC &= \frac{TP\times TN - FPos\times FN}{\sqrt{(TP+FN)(TP+FPos)(TN+FN)(TN+FPos)}} \end{aligned} \quad (2)$$

Results and discussion

Chemical diversity analysis

The diversity is important when building a QSAR model.[32] The PCA was performed here to explore the chemical space of the dataset, which contained 356 active agonists and 356 inactive agonists. For each molecule, four types of fingerprints (ExtFP 1024 bits, MACCSFP 166 bits, PubChemFP 881 bits, and AP2D 780 bits) were calculated as descriptors. Each compound is represented by a multi-dimensional vector, the dimension of which is equal to the bit-length of a fingerprint. A reducing dimension calculation was processed in the PCA. The top two principal components were preserved and plotted as illustrated in Figure 2. Each node represents a molecule of the entire dataset. The actives and inactives were rendered in black and gray color, respectively. The actives and the inactives cover the same chemical space, suggesting the diversity of this collected dataset and the reasonability of the decoy generation methods. The distance of the compounds in this dataset was calculated

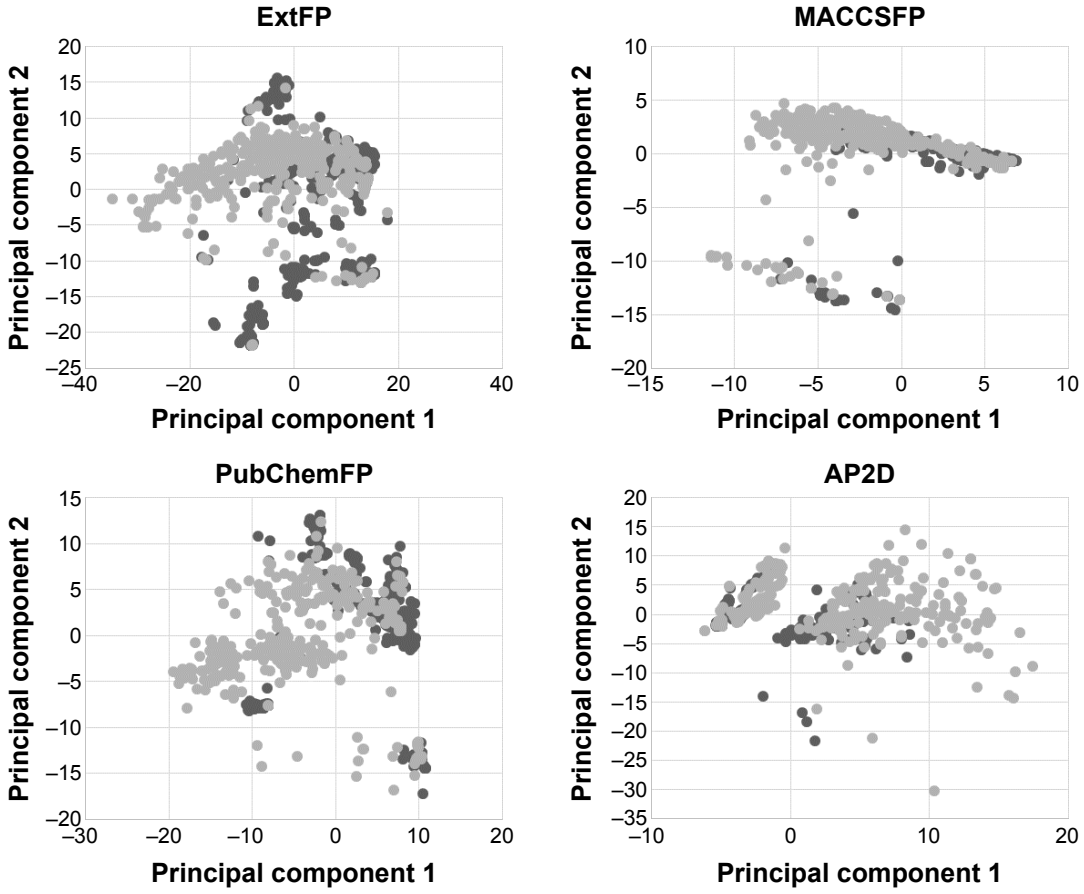

Figure 2 Principal component analysis (PCA) of the dataset.
Notes: The PCA was based on four types of fingerprints. Each dot represents a unique compound of the dataset. Black dots represent active compounds, whereas gray dots represent inactive compounds.
Abbreviations: Ext, extended; AP2D, 2D atom pairs; FP, fingerprints.

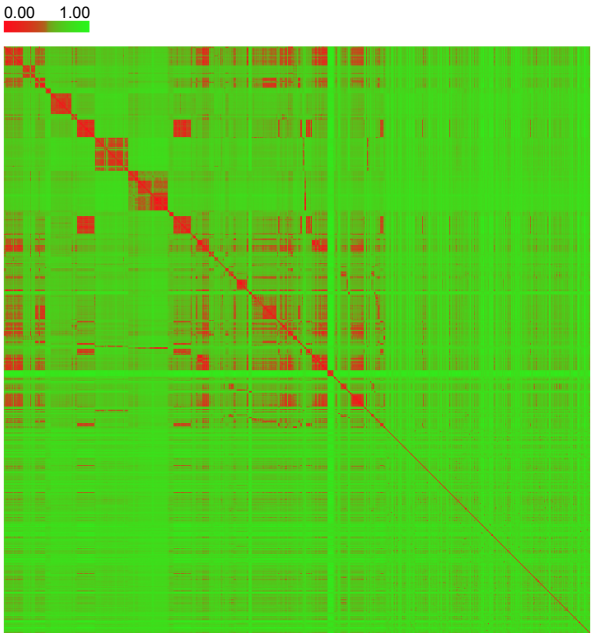

Figure 3 The heat map of distance matrix for the compounds in the collected dataset.
Note: Green represents a large distance and structural dissimilarity.

Table 1 Model performances of 5-fold cross validation

Model	CA	SE	SP	AUC	MCC
NB-ExtFP	0.8244	0.8792	0.7697	0.8633	0.6528
KNN-ExtFP	0.8693	0.9270	0.8118	0.9171	0.7437
RF-ExtFP	0.8750	0.8680	0.8820	0.9450	0.7501
SVM-ExtFP	0.8834	0.9270	0.8399	0.9407	0.7698
NB-MACCSFP	0.7921	0.8146	0.7697	0.8532	0.5849
KNN-MACCSFP	0.8707	0.9045	0.8371	0.9302	0.7433
RF-MACCSFP	0.8693	0.8961	0.8427	0.9475	0.7398
SVM-MACCSFP	0.8665	0.9410	0.7921	0.9153	0.7414
NB-PubChemFP	0.7950	0.8371	0.7528	0.8544	0.5920
KNN-PubChemFP	0.8539	0.8764	0.8315	0.9044	0.7086
RF-PubChemFP	0.8652	0.8961	0.8343	0.9408	0.7317
SVM-PubChemFP	0.8539	0.9354	0.7725	0.9103	0.7175
NB-AP2D	0.7710	0.8483	0.6938	0.8151	0.5487
KNN-AP2D	0.8357	0.8680	0.8034	0.8883	0.6728
RF-AP2D	0.8314	0.9354	0.7275	0.9056	0.6777
SVM-AP2D	0.8132	0.8933	0.7331	0.8453	0.6346

Abbreviations: NB, Naïve Bayesian; KNN, k-nearest neighbor; RF, random forest; SVM, support vector machine; Ext, extended; AP2D, 2D atom pairs; FP, fingerprints; SE, sensitivity; SP, specificity; AUC, area under the receiver operating characteristic curve; MCC, Matthews correlation coefficient; CA, classification accuracy.

using Euclidean distance based on the ExtFP. A distance matrix (712×712) was generated and plotted with a heat map. The distance values were normalized to interval 0–1. One represents the largest distance (green) and suggests the structural dissimilarity. As shown in Figure 3, most areas in the heat map were green, indicating the chemical diversity of this dataset.

Performance of cross validation

In order to evaluate the performance of models, 5-fold cross validation was employed here. We developed 16 models based on four types of fingerprints and four classifiers. The CA, SE, SP, AUC, and MCC values are listed in Table 1. The ranges of CA, SE, SP, AUC, and MCC were 0.7710–0.8834, 0.8146–0.9410, 0.6938–0.8820, 0.8151–0.9475, and 0.5487–0.7698, respectively. The ROC curves of the 16 models are illustrated in Figure 4. The AUC values of all models were greater than 0.8, indicating the good performance of the constructed models. The excellent models (MCC >0.75) were SVM combined with ExtFP, and RF combined with ExtFP. SVM-ExtFP achieved performances of CA 0.8834 and MCC 0.7698. RF-ExtFP achieved a performance of CA 0.8750 and MCC 0.7501. Ten-fold cross validation method was also employed and the model performances were evaluated and are listed in Table S1. Compared with 5-fold cross validation results, 10-fold cross validation results tended to be a bit more optimistic and showed a similar trend. In the following study, the 5-fold cross validation results were used.

For comparison, we developed models using true inactive agonists that do not include decoys. The 5-fold cross validation model performance is listed in Table S2. The mean value of the MCC for the 16 models is 0.3518, this indicates a poor performance of the imbalanced dataset when compared with the balanced dataset, which achieved a mean value of MCC 0.6881.

In order to compare the performances of different machine learning methods, we ranked the performances of the models with the same fingerprint using the values of MCC. The rank results are shown in Figure 5. NB ranked fourth with each type of fingerprint. KNN ranked first with the MACCSFP. RF ranked first with PubChemFP and AP2D. SVM ranked first with ExtFP. These results suggest the model performance varies with different combinations of machine learning approaches and molecular fingerprints. Taken together, NB performs worst compared with KNN, RF, and SVM. RF and SVM are superior to other methods for the classification of ER-β agonists. In Zang et al's binary classification models of a large collection of environmental chemicals from ER assays, they obtained the best model using SVM.[18] This consistency suggests SVM is a suitable machine learning method for this target.

We further compared the performances of the fingerprints. Model performances with various fingerprints and the same machine learning methods were ranked, as shown in Figure 6. As is obvious from the figure, ExtFP performed best and AP2D performed worst. ExtFP ranked first for any machine learning method mentioned here. In contrast, AP2D

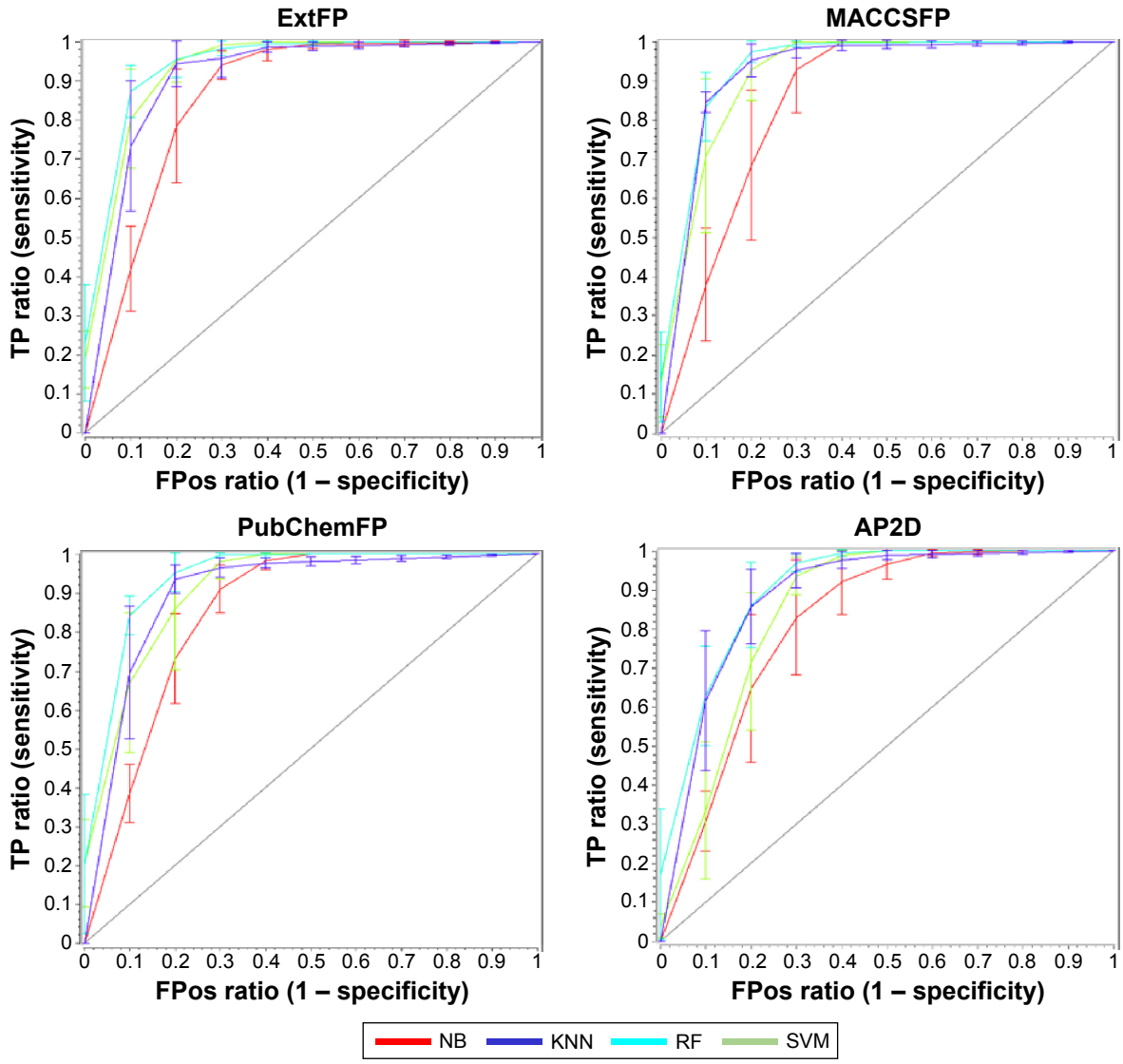

Figure 4 The ROC curves of the 5-fold cross validation models based on four types of fingerprints (FP) and four machine learning approaches.
Note: The error bar in the curve is based on five runs of the 5-fold cross validation process.
Abbreviations: ROC, receiver operating characteristic; NB, Naïve Bayesian; KNN, k-nearest neighbor; RF, random forest; SVM, support vector machine; Ext, extended; AP2D, 2D atom pairs; TP, true positives, FPos, false positives.

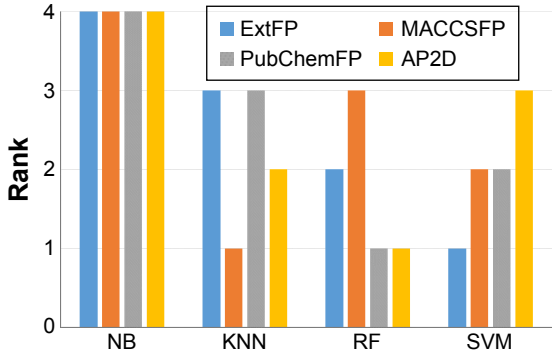

Figure 5 Performance ranking of machine learning methods with various fingerprints (FP).
Note: Take KNN for example, KNN ranked first with MACCSFP, ranked second with AP2D, and ranked third with ExtFP or PubChemFP.
Abbreviations: NB, Naïve Bayesian; KNN, k-nearest neighbor; RF, random forest; SVM, support vector machine; Ext, extended; AP2D, 2D atom pairs.

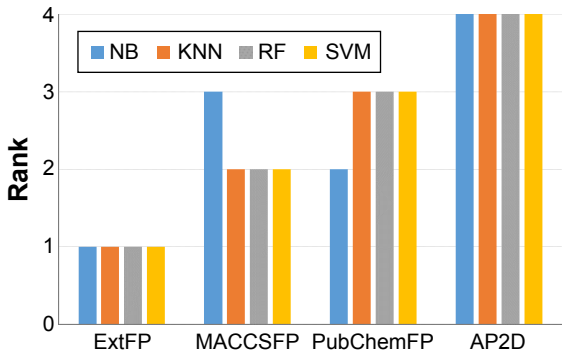

Figure 6 Performance ranking of fingerprints (FP) in various machine learning methods.
Note: Take MACCSFP for example, MACCSFP ranked third in NB, and ranked second in KNN, RF, and SVM.
Abbreviations: NB, Naïve Bayesian; KNN, k-nearest neighbor; RF, random forest; SVM, support vector machine; Ext, extended; AP2D, 2D atom pairs.

ranked last. ExtFP had 1024-bit length and AP2D had 780-bit length. Usually, the length of the fingerprints may affect the performance. However, MACCSFP with only 166-bit length, ranked second in KNN, RF, and SVM models, and ranked third in NB models. These results suggest that the model's performance is not dependent on the length of the fingerprints but the structural representation methods. Furthermore, MACCSFP is a good fingerprint generation method for capturing the structural patterns of ER-β agonists.

Performance of test set

To further evaluate the robustness of the model and to prove the observations in the cross validations, the test set was randomly split from the original dataset. The ratio of training set against test set was 2:1. The model performances for the test set are summarized in Table 2. For machine learning methods, RF ranked first with MACCSFP and PubChemFP. NB ranked last with any fingerprints. For fingerprints, ExtFP ranked first in NB and SVM. MACCSFP ranked first in KNN and second in RF and SVM. MACCSFP showed good performance and AP2D performed worst. This is in agreement with the previous observation in the cross validation.

We further collected an external test dataset from two literature sources,[17,33] which included eleven ER-β selective agonists. We added eleven decoys to obtain a dataset with 22 compounds in total. We trained models using compounds from ChEMBL and predicted the external test dataset. The models' performances are summarized in Table 3. All those models showed a good performance in classifying agonists and non-agonists.

Table 2 Model performances of test set

Model	CA	SE	SP	AUC	MCC
NB-ExtFP	0.8314	0.8876	0.7752	0.8689	0.6670
KNN-ExtFP	0.8612	0.9165	0.8058	0.9061	0.7268
RF-ExtFP	0.8769	0.8678	0.8860	0.9450	0.7538
SVM-ExtFP	0.8835	0.9248	0.8421	0.9403	0.7696
NB-MACCSFP	0.8062	0.8322	0.7802	0.8643	0.6132
KNN-MACCSFP	0.8731	0.8975	0.8488	0.9161	0.7472
RF-MACCSFP	0.8773	0.8926	0.8620	0.9483	0.7549
SVM-MACCSFP	0.8674	0.9455	0.7893	0.9174	0.7438
NB-PubChemFP	0.8136	0.8612	0.7661	0.8661	0.6301
KNN-PubChemFP	0.8657	0.8934	0.8380	0.9108	0.7325
RF-PubChemFP	0.8806	0.9248	0.8364	0.9491	0.7642
SVM-PubChemFP	0.8616	0.9347	0.7884	0.9158	0.7310
NB-AP2D	0.7719	0.8529	0.6909	0.8198	0.5511
KNN-AP2D	0.8260	0.8322	0.8198	0.8858	0.6521
RF-AP2D	0.8273	0.9091	0.7455	0.8917	0.6635
SVM-AP2D	0.8157	0.8802	0.7512	0.8514	0.6367

Abbreviations: NB, Naïve Bayesian; KNN, k-nearest neighbor; RF, random forest; SVM, support vector machine; Ext, extended; AP2D, 2D atom pairs; FP, fingerprints; SE, sensitivity; SP, specificity; AUC, area under the receiver operating characteristic curve; MCC, Matthews correlation coefficient; CA, classification accuracy.

Table 3 Model performances of external test set

Model	CA	SE	SP	AUC	MCC
NB-ExtFP	1.0000	1.0000	1.0000	1.0000	1.0000
KNN-ExtFP	0.8182	0.8182	0.8182	0.8967	0.6364
RF-ExtFP	0.8182	0.6364	1.0000	1.0000	0.6831
SVM-ExtFP	0.8636	0.7273	1.0000	1.0000	0.7559
NB-MACCSFP	0.9091	0.8182	1.0000	1.0000	0.8321
KNN-MACCSFP	0.9091	0.8182	1.0000	1.0000	0.8321
RF-MACCSFP	0.8636	0.7273	1.0000	1.0000	0.7550
SVM-MACCSFP	0.9091	0.8182	1.0000	1.0000	0.8321
NB-PubChemFP	0.9545	0.9091	1.0000	1.0000	0.9129
KNN-PubChemFP	0.9091	0.8182	1.0000	1.0000	0.8321
RF-PubChemFP	0.9091	0.8182	1.0000	1.0000	0.8321
SVM-PubChemFP	0.9091	0.8182	1.0000	1.0000	0.8321
NB-AP2D	0.8182	0.7273	0.9091	0.9504	0.6472
KNN-AP2D	0.8182	0.6364	1.0000	1.0000	0.6831
RF-AP2D	0.9545	1.0000	0.9091	0.9917	0.9129
SVM-AP2D	0.9545	0.9091	1.0000	1.0000	0.9129

Abbreviations: NB, Naïve Bayesian; KNN, k-nearest neighbor; RF, random forest; SVM, support vector machine; Ext, extended; AP2D, 2D atom pairs; FP, fingerprints; SE, sensitivity; SP, specificity; AUC, area under the receiver operating characteristic curve; MCC, Matthews correlation coefficient; CA, classification accuracy.

Conclusion

Emerging data suggest that ER-β subtype-selective ligands could be used to elicit beneficial estrogen-like activities and reduce side effects. There have been rare reports focusing on the subtype-selective ER agonist prediction. Owing to the significance of the selective ER-β agonists, in this work, we collected a dataset of selective ER-β agonists and performed the dataset analysis using PCA and distance analysis. Subsequently, we constructed the classification models of selective ER-β agonists using multiple machine learning methods and various molecular fingerprints. The models were validated through cross validation methods and test set validations. The range of classification accuracies was 77.10% to 88.34%, and the range of AUC values was 0.8151 to 0.9475, evaluated by the 5-fold cross validation. Comparison analysis suggests that both the RF and the SVM are superior to other machine learning methods for the classification of selective ER-β agonists. Chemistry Development Kit ExtFP and MACCSFP performed better in structural representation between active and inactive agonists. These models are robust and accurate, and could be applied in the virtual screening of large chemical libraries to identify selective ER-β agonists.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Jia M, Dahlman-Wright K, Gustafsson JA. Estrogen receptor alpha and beta in health and disease. *Best Pract Res Clin Endocrinol Metab*. 2015; 29(4):557–568.

2. Katzenellenbogen BS, Choi I, Delage-Mourroux R, et al. Molecular mechanisms of estrogen action: selective ligands and receptor pharmacology. *J Steroid Biochem Mol Biol*. 2000;74(5):279–285.
3. Evers NM, van den Berg JH, Wang S, et al. Cell proliferation and modulation of interaction of estrogen receptors with coregulators induced by ERα and ERβ agonists. *J Steroid Biochem Mol Biol*. 2014; 143:376–385.
4. Paterni I, Granchi C, Katzenellenbogen JA, Minutolo F. Estrogen receptors alpha (ERα) and beta (ERβ): subtype-selective ligands and clinical potential. *Steroids*. 2014;90:13–29.
5. Nilsson S, Koehler KF, Gustafsson JA. Development of subtype-selective oestrogen receptor-based therapeutics. *Nat Rev Drug Discov*. 2011;10(10):778–792.
6. Drummond AE, Fuller PJ. The importance of ERbeta signalling in the ovary. *J Endocrinol*. 2010;205(1):15–23.
7. Taylor AH, Al-Azzawi F. Immunolocalisation of oestrogen receptor beta in human tissues. *J Mol Endocrinol*. 2000;24(1):145–155.
8. Heldring N, Pike A, Andersson S, et al. Estrogen receptors: how do they signal and what are their targets. *Physiol Rev*. 2007;87(3):905–931.
9. Jordan VC, Gapstur S, Morrow M. Selective estrogen receptor modulation and reduction in risk of breast cancer, osteoporosis, and coronary heart disease. *J Natl Cancer Inst*. 2001;93(19):1449–1457.
10. Riggs BL, Hartmann LC. Selective estrogen-receptor modulators – mechanisms of action and application to clinical practice. *N Engl J Med*. 2003;348(7):618–629.
11. Dhingra K. Selective estrogen receptor modulation: the search for an ideal hormonal therapy for breast cancer. *Cancer Invest*. 2001;19(6): 649–659.
12. Maximov PY, Lee TM, Jordan VC. The discovery and development of selective estrogen receptor modulators (SERMs) for clinical practice. *Curr Clin Pharmacol*. 2013;8(2):135–155.
13. Minutolo F, Macchia M, Katzenellenbogen BS, Katzenellenbogen JA. Estrogen receptor beta ligands: recent advances and biomedical applications. *Med Res Rev*. 2011;31(3):364–442.
14. Hinsche O, Girgert R, Emons G, Grundker C. Estrogen receptor β selective agonists reduce invasiveness of triple-negative breast cancer cells. *Int J Oncol*. 2015;46(2):878–884.
15. Marzioni M, Torrice A, Saccomanno S, et al. An oestrogen receptor β-selective agonist exerts anti-neoplastic effects in experimental intrahepatic cholangiocarcinoma. *Dig Liver Dis*. 2012;44(2):134–142.
16. Roberts LR, Armor D, Barker C, et al. Sulfonamides as selective oestrogen receptor β agonists. *Bioorg Med Chem Lett*. 2011;21(19): 5680–5683.
17. Paterni I, Bertini S, Granchi C, et al. Highly selective salicylketoxime-based estrogen receptor β agonists display antiproliferative activities in a glioma model. *J Med Chem*. 2015;58(3):1184–1194.
18. Zang Q, Rotroff DM, Judson RS. Binary classification of a large collection of environmental chemicals from estrogen receptor assays by quantitative structure-activity relationship and machine learning methods. *J Chem Inf Model*. 2013;53(12):3244–3261.
19. Ng HW, Doughty SW, Luo H, et al. Development and Validation of Decision Forest Model for Estrogen Receptor Binding Prediction of Chemicals Using Large Data Sets. *Chem Res Toxicol*. 2015;28(12):2343–2351.
20. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. *J Cheminform*. 2011;3:33.
21. Klein K, Hennig S, Paul SK. A Bayesian Modelling Approach with Balancing Informative Prior for Analysing Imbalanced Data. *PloS One*. 2016;11(4):e0152700.
22. Datta S, Das S. Near-Bayesian Support Vector Machines for imbalanced data classification with equal or unequal misclassification costs. *Neural Netw*. 2015;70:39–52.
23. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem*. 2012;55(14):6582–6594.
24. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model*. 2010;50(5):742–754.
25. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem*. 2011;32(7): 1466–1474.
26. Chen L, Li Y, Zhao Q, Peng H, Hou T. ADME evaluation in drug discovery. 10. Predictions of P-glycoprotein inhibitors using recursive partitioning and naive Bayesian classification techniques. *Mol Pharm*. 2011;8(3):889–900.
27. Sun H. A naive bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. *J Med Chem*. 2005;48(12): 4031–4039.
28. Breiman L. Random forests. *Machine Learning*. 2001;45(1):5–32.
29. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci*. 2003;43(6): 1947–1958.
30. Cheng F, Yu Y, Shen J, et al. Classification of cytochrome P450 inhibitors and noninhibitors using combined classifiers. *J Chem Inf Model*. 2011; 51(5):996–1011.
31. Heikamp K, Bajorath J. Support vector machines for drug discovery. *Expert Opin Drug Discov*. 2014;9(1):93–104.
32. Xu C, Cheng F, Chen L, et al. In silico prediction of chemical Ames mutagenicity. *J Chem Inf Model*. 2012;52(11):2840–2847.
33. Chen L, Wu D, Bian HP, et al. Selective ligands of estrogen receptor β discovered using pharmacophore mapping and structure-based virtual screening. *Acta Pharmacol Sin*. 2014;35(10):1333–1341.

Supplementary materials

Table S1 Ten-fold cross validation model performance

Model	CA	SE	SP	AUC	MCC
NB-ExtFP	0.8272	0.8876	0.7669	0.8631	0.6593
KNN-ExtFP	0.8708	0.9298	0.8118	0.9245	0.7468
RF-ExtFP	0.8820	0.8848	0.8792	0.9539	0.7641
SVM-ExtFP	0.8833	0.9298	0.8371	0.9423	0.7702
NB-MACCSFP	0.7850	0.8006	0.7697	0.8533	0.5705
KNN-MACCSFP	0.8806	0.9045	0.8567	0.9287	0.7621
RF-MACCSFP	0.8903	0.9185	0.8624	0.9541	0.7821
SVM-MACCSFP	0.8679	0.9438	0.7921	0.9212	0.7446
NB-PubChemFP	0.7992	0.8455	0.7528	0.8531	0.6009
KNN-PubChemFP	0.8736	0.9129	0.8343	0.9173	0.7495
RF-PubChemFP	0.8736	0.9213	0.8258	0.9480	0.7506
SVM-PubChemFP	0.8524	0.9354	0.7697	0.9133	0.7149
NB-AP2D	0.7681	0.8399	0.6966	0.8112	0.5421
KNN-AP2D	0.8427	0.8624	0.8230	0.8964	0.6859
RF-AP2D	0.8287	0.9213	0.7360	0.9031	0.6689
SVM-AP2D	0.8131	0.8876	0.7388	0.8473	0.6335

Abbreviations: NB, Naïve Bayesian; KNN, k-nearest neighbor; RF, random forest; SVM, support vector machine; Ext, extended; AP2D, 2D atom pairs; FP, fingerprints; SE, sensitivity; SP, specificity; AUC, area under the receiver operating characteristic curve; MCC, Matthews correlation coefficient; CA, classification accuracy.

Table S2 Five-fold cross validation model performance using experimental inactive agonists

Model	CA	SE	SP	AUC	MCC
NB-ExtFP	0.7083	0.7247	0.6542	0.7385	0.3322
KNN-ExtFP	0.8425	0.9354	0.5327	0.7831	0.5219
RF-ExtFP	0.8035	0.9129	0.4393	0.7852	0.3966
SVM-ExtFP	0.8121	0.9522	0.3458	0.8275	0.3914
NB-MACCSFP	0.7169	0.7781	0.5140	0.7309	0.2715
KNN-MACCSFP	0.8186	0.9157	0.4953	0.7999	0.4517
RF-MACCSFP	0.8271	0.9298	0.4860	0.8141	0.4707
SVM-MACCSFP	0.8164	0.9522	0.3645	0.8083	0.4095
NB-PubChemFP	0.6305	0.6489	0.5701	0.6940	0.1883
KNN-PubChemFP	0.8380	0.9073	0.6075	0.7922	0.5312
RF-PubChemFP	0.7947	0.9326	0.3364	0.7969	0.3377
SVM-PubChemFP	0.7905	0.9382	0.2991	0.8057	0.3116
NB-AP2D	0.5376	0.5225	0.5888	0.5837	0.0938
KNN-AP2D	0.8098	0.8989	0.5140	0.7504	0.4380
RF-AP2D	0.7905	0.9944	0.1121	0.6727	0.2622
SVM-AP2D	0.7840	0.9860	0.1121	0.6324	0.2199

Abbreviations: NB, Naïve Bayesian; KNN, k-nearest neighbor; RF, random forest; SVM, support vector machine; Ext, extended; AP2D, 2D atom pairs; FP, fingerprints; SE, sensitivity; SP, specificity; AUC, area under the receiver operating characteristic curve; MCC, Matthews correlation coefficient; CA, classification accuracy.