Qiu et al. J Cheminform (2018) 10:21
https://doi.org/10.1186/s13321-018-0275-x

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

Finding the molecular scaffold of nuclear receptor inhibitors through high-throughput screening based on proteochemometric modelling

Tianyi Qiu[1,2†], Dingfeng Wu[1†], Jingxuan Qiu[1,3] and Zhiwei Cao[1*]

Abstract

Nuclear receptors (NR) are a class of proteins that are responsible for sensing steroid and thyroid hormones and certain other molecules. In that case, NR have the ability to regulate the expression of specific genes and associated with various diseases, which make it essential drug targets. Approaches which can predict the inhibition ability of compounds for different NR target should be particularly helpful for drug development. In this study, proteochemometric modelling was introduced to analysis the bioactivity between chemical compounds and NR targets. Results illustrated the ability of our PCM model for high-throughput NR-inhibitor screening after evaluated on both internal (AUC > 0.870) and external (AUC > 0.746) validation set. Moreover, in-silico predicted bioactive compounds were clustered according to structure similarity and a series of representative molecular scaffolds can be derived for five major NR targets. Through scaffolds analysis, those essential bioactive scaffolds of different NR target can be detected and compared. Generally, the methods and molecular scaffolds proposed in this article can not only help the screening of potential therapeutic NR-inhibitors but also able to guide the future NR-related drug discovery.

Keywords: Proteochemometric modelling, Nuclear receptor, Molecular scaffold, Cheminformatics

Background

As a ligand dependent transcription factors, nuclear receptors (NR) can be activated by important molecules such as steroidal hormones, endogenous hormones, glucocorticoids and thyroid hormones [1, 2]. After activation, NR can regulate the expression of specific genes and then participate in several essential physiological processes such as development, homeostasis and metabolism of the organism [1, 2]. Since NR can affect the expression of enormous genes which associated with various diseases such as diabetes and hepatic adipose infiltration, it can be considered as an appropriate therapeutic target for new drug discovery. Till now, 48 nuclear receptors have been discovered in humans [3], 23 of them are certified as drug target by U.S. Food and Drug Administration (FDA). Meanwhile, over 13% FDA approved drugs were aimed at those nuclear receptors [4]. In that case, discover novel drugs as nuclear receptor inhibitors have acquired a particular significance for NR-related metabolic diseases treatment. In drug design, scaffold is the fixed part of a molecule which is the essential part for biological activity of molecule. Therefore, scaffold based strategies were widely used for drug discovery [5–7]. It can be noticed that finding a new scaffold often lead to the discovery of a new inhibitor classes which may have the potential to become future drugs [8–10]. In that case, finding novel bioactive scaffolds is an essential process in the area of drug design.

In order to discover the molecular scaffold of a class of molecules such as NR-inhibitors, massive structure of molecules with bioactivity need to be screened and clustered to finding the consensus structure domain.

*Correspondence: zwcao@tongji.edu.cn
†Tianyi Qiu and Dingfeng Wu contributed equally to this work
[1] School of Life Sciences and Technology, Shanghai 10th People's Hospital, Tongji University, No. 1239 SiPing Road, Shanghai, China
Full list of author information is available at the end of the article

Traditionally, this screening evolving titration experiments is a time-consuming, expensive and labor-intensive process, which could be assisted by computer-aided drug design (CADD) [11]. In recent decades, different methods including virtual screening [12, 13], molecular docking [14, 15], de-novo drug design [16–18], pharmacophore modeling [19–21] and molecular dynamics [22, 23] were introduced to find bioactive molecules for further drug design. In the early 1960 s, quantitative structure activity relationship (QSAR) approach was established to discover the relationship between ligand and target [24]. In general, conventional QSAR based approaches consider structure information and bio-active value to efficiently predict the relationship between ligand and target. However, its prediction ability is limited to single target and enable to map multiple ligand-target relationship [25]. Also, the prediction ability of conventional QSARs were limited since only ligand information were used for model construction [25–27]. To avoid the shortages of QSAR, an approach relying on the description of both ligand and target to quantitatively analyze their relations was invented and termed as Proteochemometric (PCM) modeling in 2001 [28]. The main advantage of PCM modeling is to integrate information on both ligand and target to make the model applicable for multiple target screening, including GPCRs [29–31], proteases [32–34], kinases [35, 36], reverse transcriptase [37, 38]. However, according to author's knowledge, PCM for NR-inhibitor prediction was hardly reported.

In this article, two major steps including PCM modelling and scaffold finding were processed to guide the design of NR-inhibitors. Initially, based on a total number of 11 nuclear receptors and 9633 molecular compounds with EC_{50} values were derived from ONRLDB [39], a series of PCM modelling were generated to predict the inhibition ability for NR-inhibitors. After rigorous validation through both internal and external validation dataset, our PCM model was proved to have the potential ability for high-throughput NR-inhibitor screening. It should be noted that NR-targets validated in external dataset were not involved in our training set. That means for those NR proteins without enough bio-active data to establish a traditional QSAR models, our model may also have the ability to provide NR-inhibitor screening. Further, after molecular clustering based on our PCM model, novel bioactive scaffolds for NR-inhibitors can be discovered. The potential bioactive scaffolds for different NR targets were proposed for future drug discovery of NR-inhibitors.

Results and discussion

Construction of proteochemometric modeling

To build a proteochemometric modeling, three parts are necessarily needed: (1) bio-active data between multiple compounds and multiple targets; (2) descriptors which includes both ligand and target information; and (3) suitable learning methods to link descriptors and bio-active data. Here, for model construction, bio-active data with the most strict cutoff of $EC_{50} = 1$ μm was chosen as classification indicator. Then, to test the performance of different target descriptors, both sequence similarity descriptor and structure similarity descriptor were tested and four types of descriptors marked as T1–T4 were generated in this study (see "Methods" Part). Further, five different machine learning approaches including Random Forest (RF), Ridge Classifier (RC), Logistic Regression (LR), Decision Tree (DT) and Support Vector Classification (SVC) were used to establish different PCM models. Through 10-fold cross-validation, the performance of all five machine learning algorithms shows that Random Forest classifier can obtains the best prediction performance with the highest accuracy over 0.73 among all five and followed by Decision Tree (Table 1 and Additional file 1: Table S1). The AUC (area under curve) value also indicated that Random Forest classifier can achieves better prediction abilities than others for select NR-inhibitors. Therefore, Random Forest classifier was chose to establish our PCM modeling.

After that, the performances of 4 different descriptors were also tested by RF classifier (Fig. 1a). Results showed that the performance of sequence descriptors and structure descriptors are quite similar, which may cause by the fact that the structure feature of NR family is highly conserved. However, the performance of T1 and T2 is significantly better that those of T3 and T4, that means protein descriptors based on background data of whole protein family can better describe the properties of target proteins in model. And further, these protein descriptors can be extended to other proteins in the NR families. Considering that structure descriptors requires crystal structures which may not be available for several targets, in this study, sequence descriptors based on 30 background protein (T1) were used for model construction.

Further, the contribution of chemical descriptor was also analyzed. After statistic analysis, it can be found that lipo-hydro partition coefficient (MolLogP in RDKit) contains the major contribution among all ligand descriptors, which means it might be the key element for molecular with potential inhibition abilities (Additional file 2: Fig. S1). It can also found that, for both active compound and inactive compound, the distribution of MolLogP follows Normal distribution with significant difference, which were calculate through T test (P value < 0.0001). Result showed that, lipo-hydro partition coefficient is important for the activity of NR inhibitor, active compounds normally contain MolLogP around 5.775, while the MolLogP of inactive compounds were around 5.380.

Importance and P value of top 10 chemical structure descriptors can be found in Additional file 3: Table S2.

Evaluation of proteochemometric modeling

In this study, PCM modeling was systemically evaluated through both internal and external validations. By setting different cutoffs of bio-active data, results of different PCM models can be found in Fig. 1b, detailed information of model performance on all four protein descriptors can be found in Additional file 4: Table S3. Generally, all PCM models can gives outstanding performance in internal validation by achieving an AUC value above 0.870 on different cutoffs. For external validation, all PCM model can also achieves a satisfied performance with AUC value over 0.746. Above results indicate the excellent ability of our PCM model for NR-related inhibitors prediction. Also, with the increasing of cutoffs, the performance of PCM models increased synchronously. This probably caused by the fact that the unbalance between positive and negative data according to different cutoff. For example, when set $EC_{50} \leq 1$ as positive data and $EC_{50} > 1$ as negative data, the ratio (positive/negative) of training set, testing set and external validation set were all close to 1 (Additional file 5: Table S4). After the cutoff rising to 10, those ratios were quickly increased to 12.14, 12.95 and 22.76 respectively (Additional file 5: Table S4). Several reports also pointed out that the 1 μM cutoff may be more reasonable because it contains less noise [40]. In that case, the cutoff of EC_{50} value was set as 1 for further analysis.

Table 1 10-fold cross-validation results of different machine learning methods

Method	Accuracy	Precision	Recall	F1_score	AUC
RF	0.740	0.761	0.768	0.762	0.829
RC	0.624	0.643	0.713	0.674	–[a]
LR	0.453	0.490	0.000	0.000	0.452
DT	0.701	0.726	0.727	0.726	0.700
SVC	0.583	0.569	0.984	0.720	–[a]

Results in Table 1 were calculated based on descriptor T1

[a] This parameters can't be calculated in here (continuous predict values are needed to calculate AUC value)

Finding the molecular scaffolds for NR inhibitors

To further validate our PCM model, the active and inactive inhibitors were predicted through our PCM model. Then, the Rubberbanding Forcefield approach in DataWarrior [41] (release version 4.5.2) was used to mapping all compounds into a 2-dimentional area, while similar molecules were located close to each other (see "Molecular scaffold searching"). In that case, molecules with structure similarity over 0.95 will be clustered together. The molecules clustering and corresponding scaffolds for top clusters were illustrated in Fig. 2. Chemical name and smiles files of corresponding scaffolds were listed in Additional file 6: Table S5. Background color mapping of different NR proteins were derived from experimental values. Red color in background means active clusters while green ones means inactive clusters. Each spot represents one compound in our testing set, which were classified by our model. Red spot represents active compounds while green spot means inactive ones. The location of each compound determined by structure similarity, compounds with similar structures tend to clustered together. Compounds with similarity over a certain threshold will be defined as neighbors and connected with lines. The size of each compound spot is related to the number of its neighbor spots.

Generally, the prediction of our PCM model matched perfectly well with the experimental values. For three peroxisome proliferator-activated receptor (PPAR) protein

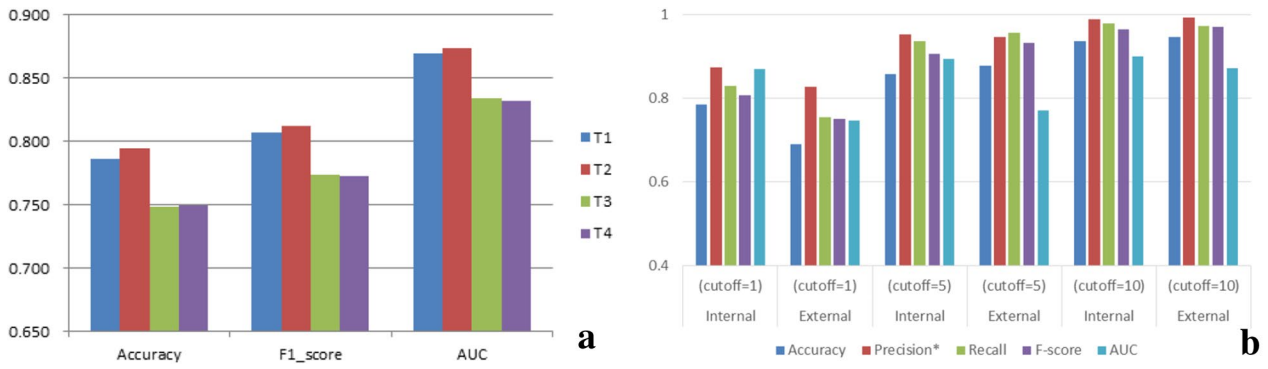

Fig. 1 Performance of PCM modeling. **a** Cross-validation performance of PCM model constructed by RF classifier based on four different protein descriptors. **b** AUC value of PCM modeling constructed by RF classifier under different cutoffs of bio-active data, this results were obtained by descriptor T1. *Precision score means the area under the precision-recall curve

targets, the top 10 clusters of each target including NR1C1 (Fig. 2a), NR1C2 (Fig. 2c) and NR1C3 (Fig. 2d) were detected and marked in each sub-graphs. For PPAR protein targets, both unique and overlapped scaffolds can be detected. For example, target NR1C1 contains 7 bioactive scaffolds (marked as S1 to S7), 2 inactive scaffolds (marked as S8 and S9) and 1 mixed scaffold (marked as S10) contains both active and inactive compounds. Among above, scaffold S1 and S6 were active in both NR1C1 and NR1C3 (Fig. 2d), while S8 and S9 were both inactive scaffold. On the other hand, different pattern can be found in target NR1C2 (Fig. 2c). In NR1C2, 7 new scaffold clusters marked as S11 to S18 were detected. Besides that, as a major inactive scaffold for NR1C1 and NR1C3, S8 was determined as active scaffold in NR1C2. Also, as an active scaffold in NR1C1 and mixed scaffold in NR1C3, scaffold S2 was defined as inactive scaffold for NR1C2. The results of two targets beside PPAR targets were quite different, totally new scaffolds were discovered and illustrated in Fig. 2e, f. All above illustrated that, even from the same protein family, the inhibitor scaffolds of different NR protein targets were still distinguishable.

Also, it should be noticed that, the bioactivity of different compounds rely on multiple factors such as side-chain composition, functional group, substituent and chirality. For instance, scaffold S10 N-benzylbenzamide contains different compounds including compound 1–3 (Fig. 2b). The molecular structure of three compounds is extremely similar except for the chirality. The stereogenic center of compound 1 (Benzenepropanoic acid, α-ethyl-4-methoxy-3-[[[[4-(trifluoromethyl)phenyl]methyl]amino]carbonyl]-, (αS)-) and compound 2 (Benzenepropanoic acid, α-ethyl-4-methoxy-3-[[[[4-(trifluoromethyl)phenyl]methyl]amino]carbonyl]-, (αR)-) are absolutely configured as S and R, respectively. Compound 3 was defined as mixture of stereoisomers which may combine with both S and R chirality.

Discussion

Computer-aided drug design (CADD) can assist and shorten the process of new drug discovery. To achieve that, one essential issue is to per-estimate the activity of different compound against different target proteins. By introducing PCM model into CADD, relationship between multiple compounds and targets can be determined. Based on high-throughput screening of compounds, bioactive molecules can be clustered and essential molecular scaffolds can be detected to guide the future development of therapeutic drugs.

In order to process high-throughput screening of bioactive inhibitors for targets from NR families, 7267 bioactive data of 11 nuclear receptors were collected to establish an in silico model. Through both internal and external validation, our PCM models were proved to be sensitive for NR-inhibitor prediction which might be benefit from our descriptors. For target descriptors, generalized sequence similarity descriptors contain information from 30 background targets from NR families. Models based on those descriptors can achieve a better prediction performance on both internal and external validation set, which means those descriptors can be extended to multiple targets from NR families. For chemical descriptors, since lipo-hydro partition coefficient contains the major contribution for classification and parameter MolLogP is distinguishable for active and inactive compounds, this may provide a clue for future therapeutic NR-inhibitors discoveries.

Another essential issue for PCM model construction is to choose the suitable machine learning method. In this study, five different machine learning methods including both regression and classification approaches were tested to establish PCM modeling. Results showed that the performance of RF and DT classifier are significantly higher than other methods, which means above algorithms might be more applicable in the case of NR-inhibitors prediction.

After high-throughput screening of NR-inhibitors, bioactive molecules could be clustered according to structure similarity and molecular scaffold enriched in each clustered can be detected and might assist the process of drug design. In this article, the appropriate models selected after evaluations were used to molecular clustering for five major NR targets. Results showed that our PCM model can successfully predict those potential NR-inhibitors which agree well with the experimental EC_{50} values. For each NR target, our algorithms can able to predict those potential therapeutic inhibitors and discover the molecular scaffolds for future drug development. Currently, this method was established on NR proteins and it can be extended to other protein targets after the accumulating of experimental data.

Methods

Data set

Training and validation dataset of nuclear receptor and its inhibitors were collected from ONRLDB [39], which including information of 11 protein targets and 9633 molecular compounds (see Additional file 7: information of 9633 compounds.sdf). After filtration, a total number of 7267 inhibitors for 11 nuclear receptors with half maximal effective concentration (EC_{50}) values were remained as our dataset. After primary statistic analyze, it can be found that the distribution of bio-active data for each protein targets were unbalanced (Fig. 3 and Additional file 8: Table S6). Major target contains more than thousands of bioactive data while several only covering tens of data. Five major protein target including NR1C1, NR1C2, NR1C3,

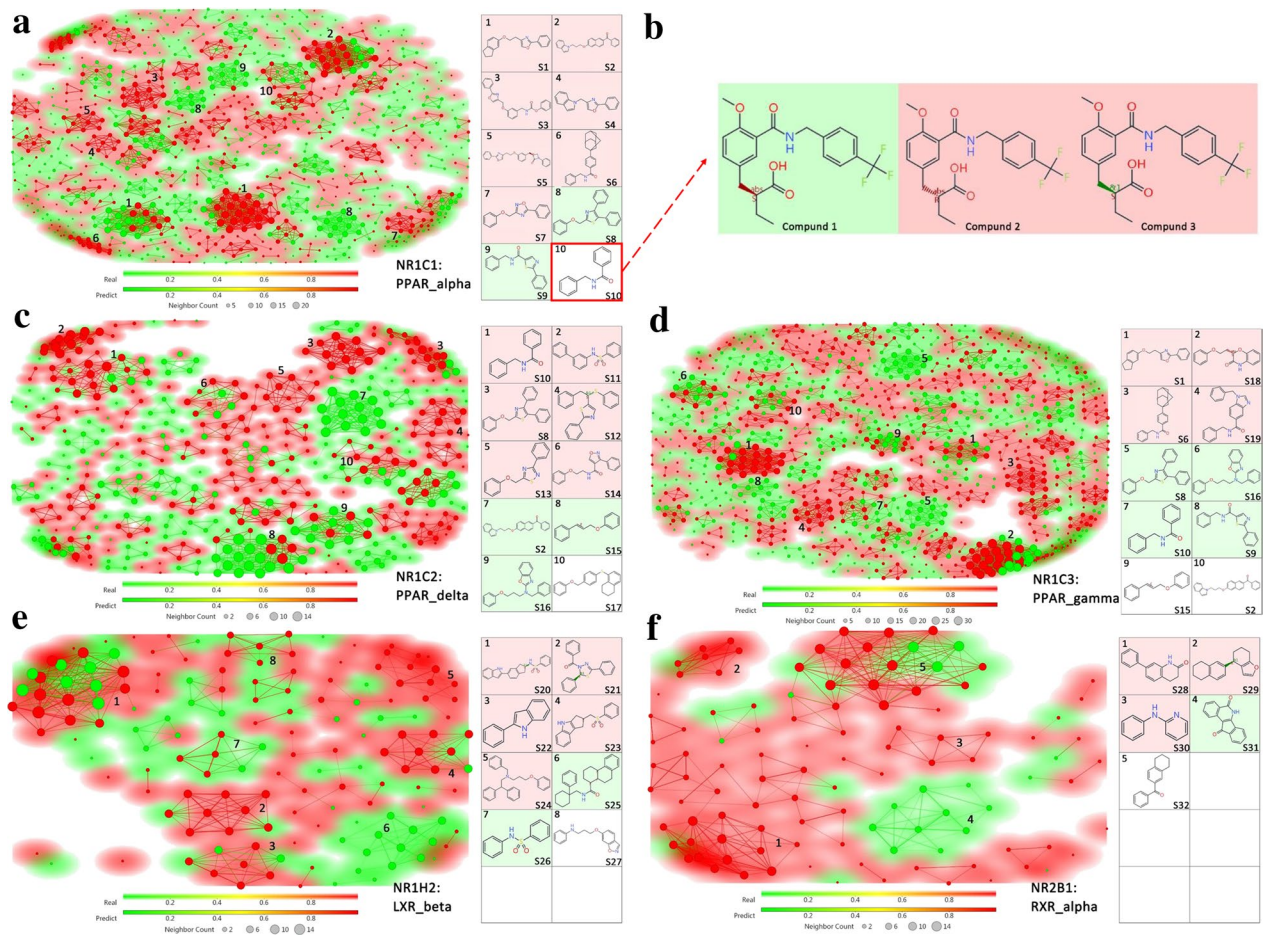

Fig. 2 Scaffold clustering of NR-inhibitors, colors in back ground and in spot represents the experimental confirmed active compounds and model predicted active compounds respectively. Red means active compounds while green means inactive compounds, white color means scaffold contains both active and inactive compounds. **a** Scaffold clustering of NR1C1-inhibitors. **b** Examples of compounds contains scaffold S10. **c** Scaffold clustering of NR1C2-inhibitors. **d** Scaffold clustering of NR1C3-inhibitors. **e** Scaffold clustering of NR1H2-inhibitors. **f** Scaffold clustering of NR2B1-inhibitors

NR1H2, NR2B1 contains over 90% of the bio-active data, which provide an abundant data for model construction. The remaining 6 targets (NR1H3, NR1H4, NR2B3, NR2B2, NR1D1, and NR1I2) with bio-active data were selected as independent validation set. After above steps, internal dataset which including 6554 bio-active data with corresponding protein targets (5 major nuclear receptor) and compounds were selected for Proteochemometric modeling. For each target, 60% of the bio-active data were chosen as training set and the rest remained as testing set. In general, 3931 bio-active data were selected as training set to generate our PCM model and the rest 2623 were used for model evaluation. Besides that, 713 bio-active data for other 6 NR proteins were collected as external validation dataset. Further, 30 crystal structures of NR from different sub-type with highest resolution were selected as background NR target (Additional file 9: Table S7).

Protein target descriptor

Here, both sequence similarity descriptors and structure similarity descriptors were used to characterize those five nuclear receptors. Firstly, a 30 protein targets from NR families can be derived from Protein Data Bank (PDB) [42] as background. For 11 protein targets in our dataset, the sequence and structure similarity compared with those 30 background protein target structures can be calculated by pairwise alignment respectively. Sequence alignment was calculated by smith-waterman alignment [43], while structure alignment was calculated by using jFATCAT [44]. Therefore, two types of generalized target descriptor including sequence similarity descriptor (T1) and structure similarity descriptor (T2) can be obtained for each protein targets. For comparison, specific descriptors based on 5 protein target from our training set instead of 30 background protein target

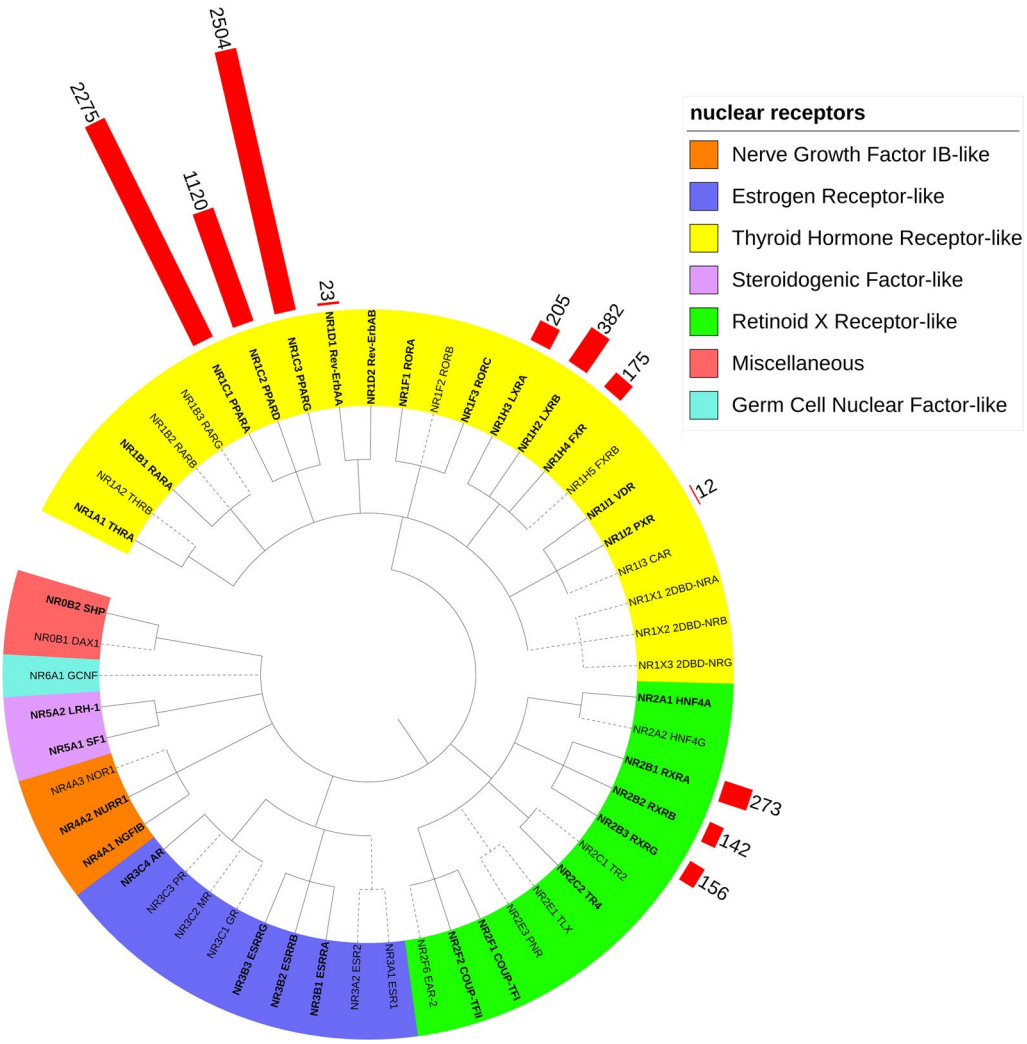

Fig. 3 Clustering tree of nuclear receptors. 7 different subtypes of NR were marked in different colors and 11 NR proteins used in this study were marked in red as well as its data distribution

were also established, recorded as T3 (specific sequence similarity descriptor based on 5 protein target) and T4 (specific structure similarity descriptor based on 5 protein target). Two generalized target descriptors can be found in Additional file 10: Table S8-1, 2 and two specific target descriptors were also listed in Additional file 11: Table S9-1, 2.

Inhibitor descriptor

Chemical structure descriptors were calculated by using RDKit (release version 2016). RDkit provides different chemical structure descriptors, which contains both chemical and physical properties such as Molecular Weight, Hydrogen Bond Donor Count, Hydrogen Bond Acceptor Count, Rotatable Bond Count and LogP etc. In addition, RDKit contains massive types of chemical descriptors derived from other tools and literatures, such as MOE-type descriptors for partial charges, MR contributions, LogP contributions, EState indices and surface area contributions integrated from molecular operating environment (MOE). In general, 187 descriptors were used to characterize the structure features of inhibitor (Additional file 12: Table S10).

Proteochemometric modeling

In this study, 4 Proteochemometric models were created from training set based on different combinations of descriptors (T1-L, T2-L, T3-L, T4-L). All models were implemented in scikit-learn (Version 0.18.1) by using Random Forest (RF) with default parameters. For

classification, different thresholds of EC_{50} were selected to distinguish positive and negative data. Here, three different thresholds ($EC_{50} < 1$ μm, $EC_{50} < 5$ μm and $EC_{50} < 10$ μm) were used for classification respectively.

Model evaluation

For each combination of descriptors, 10-fold cross-validation was carried out for the model. The performance of four models was assessed by classification accuracy. Further, both internal and external validation data were tested from different aspects to evaluate the overall performance of our models, including the area under the ROC curve (AUC) value, accuracy, precision, recall and F-score, statistical parameters were defined in the following equations:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F-score} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (4)$$

Positive samples are those with EC_{50} value below threshold. TP represents True positive, TN represents True negative, FP represents false positive and FN represent false negative.

Molecular scaffold searching

For each protein target, the similarity of corresponding molecules were analyzed based on *Rubberbanding Forcefield* approach in DataWarrior [41] (release version 4.5.2). Initially, all molecules were translated into a series of descriptors to encode various aspects of chemical structures including both 2-D and 3-D structure information. After that, calculate the entire similarity matrix between all molecules and locate most similar neighbors to be considered for every molecules. Then, stepwise relocate all molecules to ensure similar molecules were located close to each other. Finally, molecules with structure similarity over 0.95 will be clustered together [41]. For each cluster, the major Bemis-Murcko scaffold [45] (covering over 80% of the molecules in this cluster) was defined as the representative scaffold. Note that for several clusters, no major scaffold can be detected, in that case, the maximum common substructures for each two scaffolds can be calculated through RDKit and the major substructure was defined as the representative scaffold. After that, the Bemis-Murcko scaffold for each cluster can be derived and analyzed.

Additional files

Additional file 1: Table S1. 10-fold cross-validation results of different machine learning methods on four descriptors.

Additional file 2: Fig. S1. Distributions of MolLogP in both active compound and inactive compound.

Additional file 3: Table S2. Importance and P value of top 10 chemical structure descriptors.

Additional file 4: Table S3. Model performance of random forest classifier on four protein descriptors.

Additional file 5: Table S4. Data distribution of training set, testing set and external validation set.

Additional file 6: Table S5. Chemical name and smiles file of selected scaffold.

Additional file 7. Supplementary Data 1: information of 9633 compounds.

Additional file 8: Table S6. Data distribution of different NR targets.

Additional file 9: Table S7. Information of crystal structure used for descriptor generation.

Additional file 10: Table S8-1. Sequence similarity descriptors based on 30 NR proteins (T1). **Table S8-2.** Structure similarity descriptors based on 30 NR proteins (T2).

Additional file 11: Table S9-1. Sequence similarity descriptors based on 5 NR proteins (T3). **Table S9-2.** Structure similarity descriptors based on 5 NR proteins (T4).

Additional file 12: Table S10. Information of inhibitor descriptors.

Authors' contributions

TYQ and DFW developed the algorithm. TYQ and JXQ wrote the manuscript. DFW constructed the PCM model. ZWC supervised the whole project and modified the manuscript. All authors read and approved the final manuscript.

Author details

[1] School of Life Sciences and Technology, Shanghai 10th People's Hospital, Tongji University, No. 1239 SiPing Road, Shanghai, China. [2] The Institute of Biomedical Sciences, Fudan University, No. 138 Medical College Road, Shanghai, China. [3] School of Medical Instrument and Food Engineering, University of Shanghai for Science and Technology, No. 516 JunGong Road, Shanghai, China.

Acknowledgements

This work was supported in part by grants from the National Key R&D Program (2017YFC0908400, SQ2017YFC170310), the Fundamental Research Funds for the Central Universities (1350219165), the National Postdoctoral Program for Innovative Talents (BX201600033) and the China Postdoctoral Science Foundation Funded Project (2017M611451).

Competing interests

The authors declare no competing financial interests.

Availability of data and materials

All raw data used in this study are contained in the supplementary files.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Publisher's Note

Received: 27 September 2017 Accepted: 2 April 2018
Published online: 12 April 2018

References

1. Evans RM (1988) The steroid and thyroid-hormone receptor superfamily. Science 240(4854):889–895
2. Olefsky JM (2001) Nuclear receptor minireview series. J Biol Chem 276(40):36863–36864
3. Novac N, Heinzel T (2004) Nuclear receptors: overview and classification. Curr Drug Targets Inflamm Allergy 3(4):335–346
4. Overington JP, Al-Lazikani B, Hopkins AL (2006) Opinion—how many drug targets are there? Nat Rev Drug Discov 5(12):993–996
5. Dorfmueller HC, van Aalten DMF (2010) Screening-based discovery of drug-like O-GlcNAcase inhibitor scaffolds. FEBS Lett 584(4):694–700
6. Camarero JA (2011) Novel peptide-based scaffolds for drug discovery. Curr Pharm Des 17(38):4224–4225
7. Chang CF, Lin WH, Ke YY, Lin YS, Wang WC, Chen CH, Kuo PC, Hsu JTA, Uang BJ, Hsieh HP (2016) Discovery of novel inhibitors of Aurora kinases with indazole scaffold: in silico fragment-based and knowledge-based drug design. Eur J Med Chem 124:186–199
8. Ge Y, Jin Y, Wang C, Zhang J, Tang Z, Peng J, Liu K, Li Y, Zhou Y, Ma X (2016) Discovery of novel Bruton's Tyrosine Kinase (BTK) inhibitors bearing a N,9-diphenyl-9H-purin-2-amine Scaffold. ACS Med Chem Lett 7(12):1050–1055
9. Lu P, Liu X, Yuan X, He M, Wang Y, Zhang Q, Ouyang PK (2016) Discovery of a novel NEDD8 activating enzyme inhibitor with piperidin-4-amine Scaffold by structure-based virtual screening. ACS Chem Biol 11(7):1901–1907
10. Shiokawa Z, Kashiwabara E, Yoshidome D, Fukase K, Inuki S, Fujimoto Y (2016) Discovery of a novel Scaffold as an indoleamine 2,3-dioxygenase 1 (IDO1) inhibitor based on the pyrrolopiperazinone alkaloid. Longamide B ChemMedChem 11(24):2682–2689
11. Rohrer SP, Birzin ET, Mosley RT, Berk SC, Hutchins SM, Shen DM, Xiong YS, Hayes EC, Parmar RM, Foor F et al (1998) Rapid identification of subtype-selective agonists of the somatostatin receptor through combinatorial chemistry. Science 282(5389):737–740
12. Geromichalos GD, Alifieris CE, Geromichalou EG, Trafalis DT (2016) Overview on the current status of virtual high-throughput screening and combinatorial chemistry approaches in multi-target anticancer drug discovery; Part I. J Buon 21(4):764–779
13. Cruz-Monteagudo M, Schurer S, Tejera E, Perez-Castillo Y, Medina-Franco JL, Sanchez-Rodriguez A, Borges F (2017) Systemic QSAR and phenotypic virtual screening: chasing butterflies in drug discovery. Drug Discov Today 22(4):994–1007
14. Ragno R, Frasca S, Manetti F, Brizzi A, Massa S (2005) HIV-reverse transcriptase inhibition: inclusion of ligand-induced fit by cross-docking studies. J Med Chem 48(1):200–212
15. Ragno R, Mai A, Sbardella G, Artico M, Massa S, Musiu C, Mura M, Marturana F, Cadeddu A, La Colla P (2004) Computer-aided design, synthesis, and anti-HIV-1 activity in vitro of 2-alkylamino-6-[1-(2,6-difluorophenyl) alkyl]-3,4-dihydro-5-alkylpyrimidin-4(3H)-ones as novel potent non-nucleoside reverse transcriptase inhibitors, also active against the Y181C variant. J Med Chem 47(4):928–934
16. Takeda S, Kaneko H, Funatsu K (2016) Chemical-space-based de novo design method to generate drug like molecules. J Chem Inf Model 56(10):1885–1893
17. Menegatti S, Zakrewsky M, Kumar S, De Oliveira JS, Muraski JA, Mitragotri S (2016) De novo design of skin-penetrating peptides for enhanced transdermal delivery of peptide drugs. Adv Healthc Mater 5(5):602–609
18. Schneider G, Funatsu K, Okuno Y, Winkler D (2017) De novo drug design—Ye olde Scoring Problem Revisited. Mol Inform 36:1–2. https://doi.org/10.1002/minf.201681031
19. Xie HD, Qiu KX, Xie XG (2014) 3D QSAR studies, pharmacophore modeling and virtual screening on a series of steroidal aromatase inhibitors. Int J Mol Sci 15(11):20927–20947
20. Liu XF, Ouyang SS, Yu BA, Liu YB, Huang K, Gong JY, Zheng SY, Li ZH, Li HL, Jiang HL (2010) PharmMapper server: a web server for potential drug target identification using pharmacophore mapping approach. Nucleic Acids Res 38:W609–W614
21. Wang X, Shen YH, Wang SW, Li SL, Zhang WL, Liu XF, Lai LH, Pei JF, Li HL (2017) PharmMapper 2017 update: a web server for potential drug target identification with a comprehensive target pharmacophore database. Nucleic Acids Res 45(W1):W356–W360
22. Zhao HT, Caflisch A (2015) Molecular dynamics in drug design. Eur J Med Chem 91:4–14
23. Ndagi U, Mhlongo NN, Soliman ME (2017) The impact of Thr91 mutation on c-Src resistance to UM-164: molecular dynamics study revealed a new opportunity for drug design. Mol BioSyst 13(6):1157–1171
24. Hansch C, Steward AR (1964) The use of substituent constants in the analysis of the structure-activity relationship in penicillin derivatives. J Med Chem 7:691–694
25. van Westen GJP, Wegner JK, IJzerman AP, van Vlijmen HWT, Bender A (2011) Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. Medchemcomm 2(1):16–30
26. Qiu T, Qiu J, Feng J, Wu D, Yang Y, Tang K, Cao Z, Zhu R (2017) The recent progress in proteochemometric modelling: focusing on target descriptors, cross-term descriptors and application scope. Brief Bioinform 18(1):125–136
27. Wu DF, Huang Q, Zhang YD, Zhang QC, Liu Q, Gao J, Cao ZW, Zhu RX (2012) Screening of selective histone deacetylase inhibitors by proteochemometric modeling. BMC Bioinform 13:212. https://doi.org/10.1186/1471-2105-13-212
28. Lapinsh M, Prusis P, Gutcaits A, Lundstedt T, Wikberg JES (2001) Development of proteo-chemometrics: a novel technology for the analysis of drug–receptor interactions. Bba-Gen Subjects 1525(1–2):180–190
29. Cortes-Ciriano I, van Westen GJP, Lenselink EB, Murrell DS, Bender A, Malliavin T (2014) Proteochemometric modeling in a Bayesian framework. J Cheminform 6:35. https://doi.org/10.1186/1758-2946-6-35
30. Gao J, Huang Q, Wu DF, Zhang QC, Zhang YD, Chen T, Liu Q, Zhu RX, Cao ZW, He Y (2013) Study on human GPCR-inhibitor interactions by proteochemometric modeling. Gene 518(1):124–131
31. Lapinsh M, Veiksina S, Uhlen S, Petrovska R, Mutule I, Mutulis F, Yahorava S, Prusis P, Wikberg JES (2005) Proteochemometric mapping of the interaction of organic compounds with melanocortin receptor subtypes. Mol Pharmacol 67(1):50–59
32. Huang Q, Jin HX, Liu Q, Wu Q, Kang H, Cao ZW, Zhu RX (2012) Proteochemometric modeling of the bioactivity spectra of HIV-1 protease inhibitors by introducing protein–ligand interaction fingerprint. PLoS ONE 7(7):e41698. https://doi.org/10.1371/journal.pone.0041698
33. Prusis P, Lapins M, Yahorava S, Petrovska R, Niyomrattanakit P, Katzenmeier G, Wikberg JES (2008) Proteochemometrics analysis of substrate interactions with dengue virus NS3 proteases. Bioorgan Med Chem 16(20):9369–9377
34. Prusis P, Junaid M, Petrovska R, Yahorava S, Yahorau A, Katzenmeier G, Lapins M, Wikberg JES (2013) Design and evaluation of substrate-based octapeptide and non substrate-based tetrapeptide inhibitors of dengue virus NS2B-NS3 proteases. Biochem Biophys Res Commun 434(4):767–772
35. Lapins M, Wikberg JES (2010) Kinome-wide interaction modelling using alignment-based and alignment-independent approaches for kinase description and linear and non-linear data analysis techniques. BMC Bioinform 11:339. https://doi.org/10.1186/1471-2105-11-339
36. Subrarnanian V, Prusis P, Pietila LO, Xhaard H, Wohlfahrt G (2013) Visually interpretable models of kinase selectivity related features derived from field-based proteochemometrics. J Chem Inf Model 53(11):3021–3030
37. Junaid M, Lapins M, Eklund M, Spjuth O, Wikberg JES (2010) Proteochemometric modeling of the susceptibility of mutated variants of the HIV-1 virus to reverse transcriptase inhibitors. PLoS ONE 5(12):e14353. https://doi.org/10.1371/journal.pone.0014353
38. van Westen GJP, Wegner JK, Geluykens P, Kwanten L, Vereycken I, Peeters A, IJzerman AP, van Vlijmen HWT, Bender A, (2011) Which compound to select in lead optimization? Prospectively validated proteochemometric models guide preclinical development. PLoS ONE 6(11):e27518. https://doi.org/10.1371/journal.pone.0027518
39. Nanduri R, Bhutani I, Somavarapu AK, Mahajan S, Parkesh R, Gupta P (2015) ONRLDB-manually curated database of experimentally validated ligands for orphan nuclear receptors: insights into new drug discovery. Database-Oxford

40. Lusci A, Fooshee D, Browning M, Swamidass J, Baldi P (2015) Accurate and efficient target prediction using a potency-sensitive influence-relevance voter. J Cheminform 7:63. https://doi.org/10.1186/s13321-015-0110-6
41. Sander T, Freyss J, von Korff M, Rufener C (2015) DataWarrior: an open-source program for chemistry aware data visualization and analysis. J Chem Inf Model 55(2):460–473
42. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S et al (2002) The Protein Data Bank. Acta Crystallogr Sect D Biol Crystallogr 58(Pt 6 No 1):899–907
43. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147(1):195–197
44. Prlic A, Bliven S, Rose PW, Bluhm WF, Bizon C, Godzik A, Bourne PE (2010) Pre-calculated protein structure alignments at the RCSB PDB website. Bioinformatics 26(23):2983–2985
45. Bemis GW, Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. J Med Chem 39(15):2887–2893