ASIAN JOURNAL OF PHARMACEUTICAL AND CLINICAL RESEARCH

INNOVARE ACADEMIC SCIENCES
Knowledge to Innovation

Vol 8, Issue 6, 2015

ISSN - 0974-2441

Research Article

EMPLOYING RECURSIVE PARTITION AND REGRESSION TREE METHOD TO INCREASE THE QUALITY OF STRUCTURE-BASED VIRTUAL SCREENING IN THE ESTROGEN RECEPTOR ALPHA LIGANDS IDENTIFICATION

ENADE PERDANA ISTYASTONO*

Department of Pharmacy, Division of Drug Design and Discovery, Faculty of Pharmacy, Sanata Dharma University, Depok, Sleman, Yogyakarta 55282, Indonesia. Email: enade@usd.ac.id

Received: 04, August 2015, Revised and Accepted: 29 September 2015

ABSTRACT

Objective: Increase the predictive quality of the structure-based virtual screening (SBVS) protocol to identify potent ligands for estrogen receptor alpha (ERα).

Methods: Employing recursive partition and regression tree (RPART) method to identify potent ligands for ERα among their decoys by using molecular docking scores and the protein-ligand interaction fingerprint bitstrings as the predictors. These predictors were obtained from previously published SBVS campaign to identify potent ligands for ERα. The quality of the protocol by using RPART method was assessed by examining the enrichment factors and the accuracy in 95% level of confidence compared to the reference protocol.

Results: The decision tree resulted from analysis using RPART method increased the enrichment factor and the accuracy values of the SBVS protocol from 18.5 to 247.9 and from 0.975 to 0.989, respectively. Notably, the accuracy value of the protocol using the decision tree was statistically significant in 95% level of confidence while the reference protocol was not.

Conclusion: RPART method could lead to a significant increase of the SBVS quality to identify potent ligands for ERα.

Keywords: Recursive partition and regression tree, Molecular docking, Interaction fingerprint, Estrogen receptor alpha.

INTRODUCTION

The understanding of molecular determinants of protein-ligand binding has served as a key success in structure-based drug design and discovery [1-6]. The information resulted from site-directed mutagenesis (SDM) studies provides an important basis on which residues an active ligand might bind in the relevant interaction pocket [1,2,4,7,8]. Employing this knowledge has been proven to increase the quality of structure-based virtual screening (SBVS) and to assist in the elucidation on how active ligands bind in their receptor targets [1,2,5,6,9-11]. Unfortunately, this important information on SDM results is not available for every relevant target for drug discovery purposes, e.g., for breast cancer drug discovery targeting estrogen receptor alpha (ERα) [11-14]. Therefore, method development to increase the quality of SBVS campaigns as well as to virtually identify molecular determinants to guide SDM studies is of considerable interest [1,4].

Tamoxifen, by binding to ERα [15], has served as one of the drugs of choice in the chemotherapy for breast cancer treatment [16]. The compound is metabolized to 4-hydroxy-tamoxifen and N-*des*-methyl-4-hydroxo-tamoxifen, which bind to ERα with 30-1000 times stronger compared to tamoxifen [17]. Visual inspection of the crystal structure of 4-hydroxy-tamoxifen binds to ERα discovered that the ERα binding pocket has circa 70 residues (Fig. 1) [11-13,18]. The visual inspection has also discovered two hydrogen bond (H-bond) networks formed by the co-crystalized ligand 4-hydroxy-tamoxifen with the ERα binding pocket, i.e., (i) the phenol moiety of the co-crystal ligand with GLU353, ARG394 and a conserved water molecule, and (ii) the protonated amine of the co-crystal ligand with THR347 and ASP351 [11,13,18]. An ionic interaction between the protonated amine of the co-crystal ligand and ASP351 has also been observed [11,13]. These interactions might play an important role in the ERα-ligand binding, which can, therefore, be used to increase the SBVS quality significantly [5,6,10,19].

The freely and publicly available PyPLIF, a software to identify protein-ligand interaction fingerprints (PLIF) offers opportunities to develop a method for the identification of the interactions that play an important role in the protein-ligand binding [11,19]. Since there is no molecular determinants data resulted from SDM studies on ERα-ligand binding that could assist the development of more predictive SBVS protocol to identify ligands for ERα, PyPLIF could be pivotal in the identification of the molecular determinants [11,13,19]. The research presented in this article made use of in-house data from a previously published research project on the development and validation of a retrospective SBVS protocol using a database of useful decoys enhanced version (DUD-e) to identify ligands for ERα [13,20]. The ChemPLP scores and the PLIF bitstrings resulted from the SBVS protocol [13] were used as the descriptors in the construction of a decision tree using recursive partition and regression tree (RPART) method [21-23]. The best decision tree resulted from the analysis using RPART could increase the virtual screening quality significantly.

MATERIALS AND METHODS

Materials

ChemPLP scores [24] and PLIF bitstrings [11,25,26] resulted from retrospective SBVS campaigns on ERα ligands and decoys [20] were obtained from our in-house database [13]. The data sets consisted of the ChemPLP scores and PLIF bitstrings of the ligands and decoys docking poses with the best ChemPLP score for each compound. The packages "rpart" [21,23] and "caret" [22,23] were employed in the statistical analysis using R computational statistics software version 3.2.1 (R-3.2.1) [23].

Methods

By employing the "RPART" package in R-3.2.1 [21,23], the best decision tree for every data set was constructed and selected. The decision tree provided the lowest cross-validated prediction error (CV-err)

was selected [21]. The tree was subsequently used to predict using the predictors in the data set and confusion matrix, i.e. consisted of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), was created [21,27]. The enrichment factor (EF=(TP/(TP+FN))/(FP/(TN+FP))) [28,29] and accuracy (ACC=TP+TN/[TP+TN+FP+FN]) [22,23] values were then calculated and compared to the values of the reference protocol [30]. At 95% level of confidence, the confidence interval of the ACC value and the p value to examine whether the accuracy was higher than the "no information rate" (the largest class percentage in the data) were calculated using "confusionMatrix" module in the "caret" package of R-3.2.1 [22,23] to examine the significance of the ACC value.

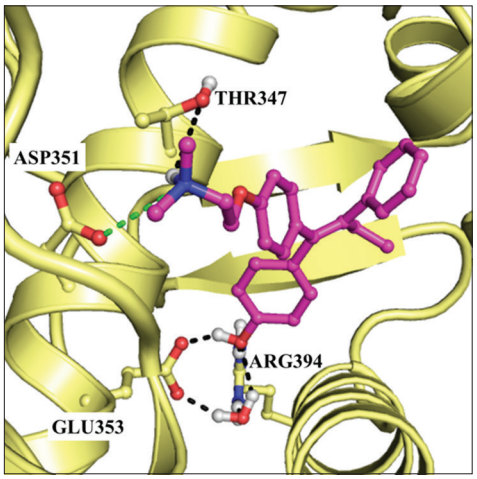

Fig. 1: The co-crystal ligand 4-hydroxytamoxifen (carbon atoms are in magenta) in the estrogen receptor alpha (ERα) (carbon atoms are in light yellow) binding pocket [18]. ERα is presented in the cartoon mode while the crystal structure pose is presented in the sticks mode. Only polar hydrogens (presented in white), residues (presented in sticks mode, carbon atoms are in light yellow) with hydrogen bond interaction (presented in black dashes) and ionic interaction (presented in green dashes) to the ligand, and a conserved water molecule [12,18] are presented for the sake of clarity. Nitrogen and oxygen atoms are presented in blue and red, respectively. The figure was prepared by employing the same point of view and similar rendering with the figure in Setiawati *et al*. [13]

RESULTS

The research presented in this article aimed to examine if employing ChemPLP scores and PLIF bitstrings from previously published retrospective SBVS campaigns as predictors in RPART analysis could increase the quality of the SBVS. The RPART analysis resulted in decision trees presented in Table 1. The decision tree with the lowest CV-err value was selected as the best decision tree to be employed further in determine whether a compound was a potent ERα ligands (Fig. 2). Comparison of the statistical significances of the results between the reference protocol [13] and the decision tree showed that the decision tree significantly increased the virtual screening quality to identify potent ERα ligands (Table 2).

DISCUSSION

By employing the selected decision tree (Table 1 and Fig. 2), the quality of the SBVS to identify potent ERα ligands has increased significantly (Table 2). The reference protocol [13] used only the ChemPLP scores and has resulted in better EF value (18.5) compared to the EF value of the original SBVS (EF=15.4) accompanying the publication of DUD-e [20]. Interestingly, a decision tree using the PLIF bitstrings identified using PyPLIF [11,26] as additional descriptors accompanying ChemPLP score has been constructed here using RPART method [21,23] and could outperform the quality of the SBVS protocols [13,20], significantly (Table 2). On the other hand, a combination of ChemPLP scores [24]

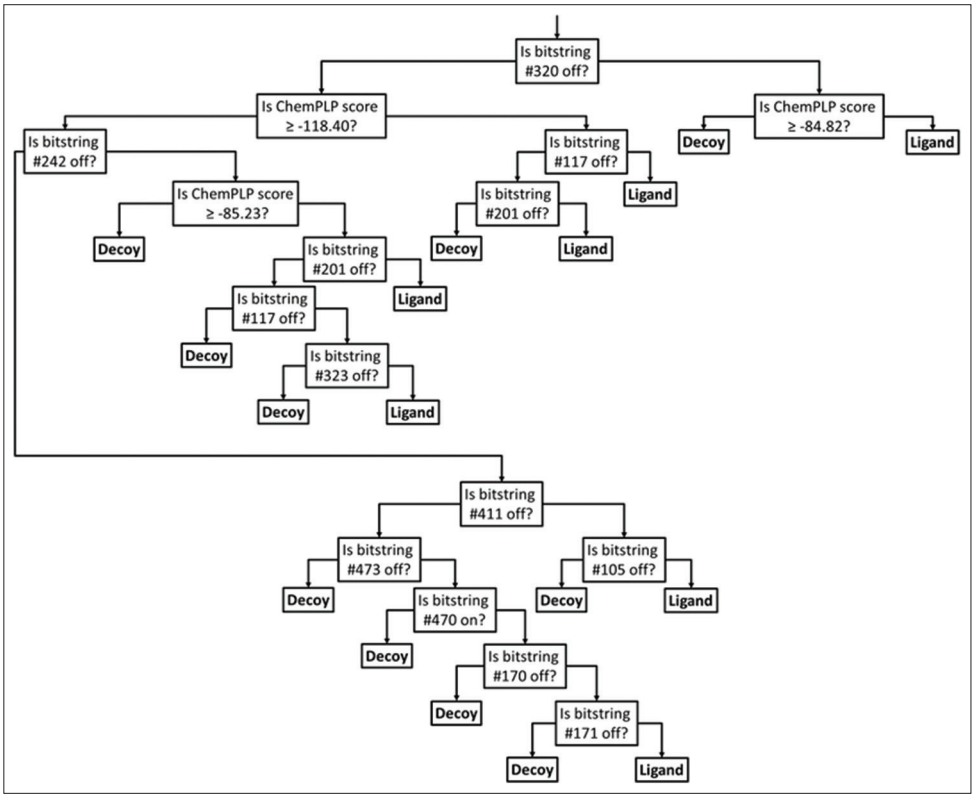

Fig. 2: The decision tree adopted from the best one resulted from the recursive partition and regression tree method (Table 1). If the answer of the question in the box is "Yes," then the path goes to the left arrow, otherwise it goes to the right arrow [21]

Table 1: Decision trees resulted from employing RPART method on the SBVS results to identify potent ERα ligands

Number	CP[a]	CV-err[b]	CV-std[c]
1	0.1097	1.0000	0.0506
2	0.0705	0.8903	0.0048
3	0.0313	0.8355	0.0464
4	0.0183	0.7781	0.0447
5	0.0141	0.7781	0.0447
6	0.0104	0.7415	0.0437
7[d]	0.0100	0.7023	0.0425

[a]Complexity parameter of the decision tree; [b]Cross-validated prediction error; [c]Cross-validated standard deviation; [d]The decision tree with the lowest CV-err and the lowest CV-std involves the following descriptors: ChemPLP score and PLIF bitstrings number 105, 117, 170, 171, 201, 242, 320, 323, 411, 470, and 473 [Figure 2]. RPART: Recursive partition and regression tree, SBVS: Structure-based virtual screening, ERα: Estrogen receptor alpha, PLIF: Protein-ligand interaction fingerprints

Table 2: Statistical significances of the best decision tree compared to the reference protocol

SBVS protocol	Confusion matrix				EF	ACC
	TP	FN	TN	FP		
Reference[a]	71	312	20478	207	18.5	0.975
Employing the best decision tree[b]	202	181	20641	44	247.9	0.989**

[a]Refer to [13]; [b]Table 1 and Figure 2; **p (ACC > "no information rate") <0.01. SBVS: Structure-based virtual screening, TP: True positive, FN: False negative, TN: True negative, FP: False positive, EF: Enrichment factor

Table 3: The important PLIF bitstrings in the selected decision tree

Bitstring number[a]	Corresponding residue	Interaction type[b]
320	GLY420	Hydrogen bond (protein as acceptor)
242	ARG394	Hydrogen bond (protein as donor)
117	GLU353	Hydrogen bond (protein as acceptor)
411	GLY521	Hydrogen bond (protein as acceptor)
473	CYS530	Hydrogen bond (protein as donor)
105	ASP351	Electrostatic interaction (protein as anion)
201	LEU387	Hydrogen bond (protein as acceptor)
470	CYS530	Non polar interaction
170	TRP383	Aromatic face-to-face
171	TRP383	Aromatic edge-to-face
323	MET421	Non polar interaction

[a]Sorted by the importance in the decision tree [Figure 2] based on the sequence of appearance; [b]For more explanation see [11,19,25]. PLIF: Protein-ligand interaction fingerprints

and the Tanimoto similarity to a reference PLIF bitstrings [25] showed successful retrospective and prospective SBVS campaigns to discover novel active fragments for histamine H_1 receptor [6]. Recently, retrospective and prospective SBVS campaigns were reported could discover novel active fragments for histamine H_4 receptor with only using the Tanimoto similarity to a reference PLIF bitstrings but with two distinct reference ligands and two different templates in the building the receptor-ligand complexes for performing SBVS campaigns [29]. Since different references and templates could complementary resulted in a successful SBVS campaign [29], a reference-independent method using PLIF bitstrings could be of considerable interest instead of that instead of using Tanimoto similarity to a reference PLIF bitstrings as one of the scoring functions. Fortunately, RPART method [21] has provided such approaches and, in this research, resulted in a significantly improved quality of the SBVS protocol (Tables 1 and 2).

The prospective hit rate of the SBVS using combined objective scoring functions outperformed the one using a single objective scoring function [6,29]. As suggested by Istyastono *et al.* [29] and shown by de Graaf *et al.* [6], the optimization in using combined scoring functions could lead to a significantly better SBVS protocol quality. Notably, both SBVS protocols [6,29] made use of previous information of the molecular determinants in protein-ligand binding [7,31] to select only poses that have interaction with the pivotal ASP residue in the ligand binding [6,29], which increased the SBVS qualities significantly. These indicated that prior knowledge of the molecular determinants of the protein-ligand binding is pivotal to have robust SBVS protocols. However, not every relevant drug target has the privilege to have such information. The using of PLIF bitstrings as descriptors in this research offered the opportunity to identify virtually the molecular determinants in ERα-ligand binding (Table 3). In Fig. 2, these residues participate in different branches, which indicates that the effect of these molecular determinants is ligand dependent [1,2]. Notably, PLIF bitstrings #242, #117, and #105 were previously recognized in the visual inspection [13,18] as the plausible molecular determinants since they participated in the H-bond networks of the 4-hydroxytamoxifen binding to ERα (Fig. 1). However, only these residues, CYS530, TRP383, and MET421 that could be employed further in novel ligand designs since their PLIF bitstrings could correspond to the interaction of the side chain, while other PLIF bitstrings could only correspond to the main chain [11,25,32].

The method used in this research could be categorized as a binary quantitative structure-activity relationship [27] by using ChemPLP scores [24] and PLIF bitstrings [11,19,25] from previous retrospective SBVS campaigns [13] as the descriptors instead of using the physicochemical properties of the compounds [27,33-36]. Since the descriptors resulted from SBVS campaigns, the visual inspections on the corresponding docking poses and the information from the decision tree resulted from RPART method could provide more intuitive and powerful tool for the design of novel potent ligands [6,9,29,37]. Notably, the protocol resulted in this research could be employed to identify potent ligands for ERα. In turn, this offers the opportunity to identify potent phytoestrogens [38-40] and could provide early alarms of their activity and toxicity since they are available in daily foods [34,40-42].

CONCLUSIONS

Employing RPART method has led to a significant increase in the SBVS quality to identify potent ligands for ERα. Besides increasing the SBVS quality, analysis using RPART method on post-SBVS campaigns, which result in docking scores and PLIF bitstrings, could also assist the identification of the molecular determinants in protein-ligand bindings. These strategies could, therefore, be further employed and examined in the construction of SBVS protocols to identify potent ligands for other pharmaceutical relevant targets.

ACKNOWLEDGMENTS

The author thanks Florentinus D.O. Riswanto and Sri H. Yuliani for the preparation of the in-house results of the previously published SBVS campaigns on ERα. This research was financially supported by Indonesian Directorate General of Higher Education (Competitive Research Block Grant 1320/K5/KM/2014).

REFERENCES

1. Istyastono EP, Nijmeijer S, Lim HD, van de Stolpe A, Roumen L, Kooistra AJ, *et al.* Molecular determinants of ligand binding modes in the histamine H_4 receptor: Linking ligand-based three-dimensional quantitative structure − Activity relationship (3D-QSAR) models to *in silico* guided receptor mutagenesis studies. J Med Chem 2011;54(23):8136-47.
2. Lim HD, de Graaf C, Jiang W, Sadek P, McGovern PM, Istyastono EP, *et al.* Molecular determinants of ligand binding to H_4R species variants. Mol Pharmacol 2010;77(5):734-43.

3. Wijtmans M, de Graaf C, de Kloe G, Istyastono EP, Smit J, Lim H, *et al.* Triazole ligands reveal distinct molecular features that induce histamine H_4 receptor affinity and subtly govern H_4/H_3 subtype selectivity. J Med Chem 2011;54(6):1693-703.
4. Istyastono EP, de Graaf C, de Esch IJ, Leurs R. Molecular determinants of selective agonist and antagonist binding to the histamine H_4 receptor. Curr Top Med Chem 2011;11(6):661-79.
5. Sirci F, Istyastono EP, Vischer HF, Kooistra AJ, Nijmeijer S, Kuijer M, *et al.* Virtual fragment screening: Discovery of histamine H_3 receptor ligands using ligand-based and protein-based molecular fingerprints. J Chem Inf Model 2012;52(12):3308-24.
6. de Graaf C, Kooistra AJ, Vischer HF, Katritch V, Kuijer M, Shiroishi M, *et al.* Crystal structure-based virtual screening for fragment-like ligands of the human histamine H_1 receptor. J Med Chem 2011;54(23):8195-206.
7. Shin N, Coates E, Murgolo NJ, Morse KL, Bayne M, Strader CD, *et al.* Molecular modeling and site-specific mutagenesis of the histamine-binding site of the histamine H_4 receptor. Mol Pharmacol 2002;62(1):38-47.
8. Lim HD, Jongejan A, Bakker RA, Haaksma E, de Esch IJ, Leurs R. Phenylalanine 169 in the second extracellular loop of the human histamine H_4 receptor is responsible for the difference in agonist binding between human and mouse H_4 receptors. J Pharmacol Exp Ther 2008;327(1):88-96.
9. Kufareva I, Katritch V. Participants of GPCR Dock, Stevens RC, Abagyan R. Advances in GPCR modeling evaluated by the GPCR Dock 2013 assessment: Meeting new challenges. Structure 2014;22(8):1120-39.
10. Yuniarti N, Ikawati Z, Istyastono EP. The importance of ARG513 as a hydrogen bond anchor to discover COX-2 inhibitors in a virtual screening campaign. Bioinformation 2011;6(4):164-6.
11. Radifar M, Yuniarti N, Istyastono EP. PyPLIF: Python-based Protein-Ligand Interaction fingerprinting. Bioinformation 2013;9(6):325-8.
12. Anita Y, Radifar M, Kardono LB, Hanafi M, Istyastono EP. Structure-based design of eugenol analogs as potential estrogen receptor antagonists. Bioinformation 2012;8(19):901-6.
13. Setiawati A, Riswanto FD, Yuliani SH, Istyastono EP. Retrospective validation of a structure-based virtual screening protocol to identify ligands for estrogen receptor alpha and its application to identify the alpha-mangostin binding pose. Indones J Chem 2014;14:103-8.
14. Bayala B, Bassole IH, Scifo R, Gnoula C, Morel L, Lobaccaro JM, *et al.* Anticancer activity of essential oils and their chemical components - A review. Am J Cancer Res 2014;4(6):591-607.
15. Harvey JM, Clark GM, Osborne CK, Allred DC. Estrogen receptor status by immunohistochemistry is superior to the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer. J Clin Oncol 1999;17(5):1474-81.
16. BIG - Collaborative Group, Mouridsen H, Giobbie-Hurder A, Goldhirsch A, Thürlimann B, Paridaens R, *et al.* Letrozole therapy alone or in sequence with tamoxifen in women with breast cancer. N Engl J Med 2009;361(8):766-76.
17. Desta Z, Ward BA, Soukhova NV, Flockhart DA. Comprehensive evaluation of tamoxifen sequential biotransformation by the human cytochrome P450 system *in vitro*: Prominent roles for CYP3A and CYP2D6. J Pharmacol Exp Ther 2004;310(3):1062-75.
18. Shiau AK, Barstad D, Loria PM, Cheng L, Kushner PJ, Agard DA, *et al.* The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. Cell 1998;95(7):927-37.
19. Salentin S, Haupt VJ, Daminelli S, Schroeder M. Polypharmacology rescored: Protein-ligand interaction profiles for remote binding site similarity assessment. Prog Biophys Mol Biol 2014;116(2-3):174-86.
20. 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):6582-94.
21. Therneau T, Atkinson B, Ripley B. Rpart: Recursive partitioning and regression trees. R Package Version 4.1-9. Available from: http://www.CRAN.R–project.org/package=rpart; 2015.
22. Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, *et al.* Caret: Classification and Regression Training. R Package Version 6.0-52. Available from: http://www.CRAN.R–project.org/package=caret; 2015.
23. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. Available from: http://www.R–project.org/; 2015.
24. Korb O, Stützle T, Exner TE. Empirical scoring functions for advanced protein-ligand docking with PLANTS. J Chem Inf Model 2009;49(1):84-96.
25. Marcou G, Rognan D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. J Chem Inf Model 2007;47(1):195-207.
26. Radifar M, Yuniarti N, Istyastono EP. PyPLIF-assisted redocking indomethacin-(*R*)-alpha-ethyl-ethanolamide into cyclooxygenase-1. Indones J Chem 2013;13:283-6.
27. Luo M, Wang XS, Roth BL, Golbraikh A, Tropsha A. Application of quantitative structure-activity relationship models of 5-HT_{1A} receptor binding to virtual screening identifies novel and potent 5-HT_{1A} ligands. J Chem Inf Model 2014;54(2):634-47.
28. de Graaf C, Rognan D. Selective structure-based virtual screening for full and partial agonists of the beta2 adrenergic receptor. J Med Chem 2008;51(16):4978-85.
29. Istyastono EP, Kooistra AJ, Vischer H, Kuijer M, Roumen L, Nijmeijer S, *et al.* Structure-based virtual screening for fragment-like ligands of the g protein-coupled histamine H_4 receptor. Medchemcomm 2015;6:1003-17.
30. Setiawati A, Riswanto FO, Yuliani SH, Istyastono EP. Anticancer activity of mangosteen pericarp dry extract against MCF-7 breast cancer cell line though estrogen receptor-α. Indones J Pharm 2014;25:119-24.
31. Vroling B, Sanders M, Baakman C, Borrmann A, Verhoeven S, Klomp J, *et al.* GPCRDB: Information system for G protein-coupled receptors. Nucleic Acids Res 2011;39:D309-19.
32. Loving K, Alberts I, Sherman W. Computational approaches for fragment-based and de novo design. Curr Top Med Chem 2010;10(1):14-32.
33. Golbraikh A, Muratov E, Fourches D, Tropsha A. Data set modelability by QSAR. J Chem Inf Model 2014;54(1):1-4.
34. Appiah-Opong R, Commandeur JN, Istyastono E, Bogaards JJ, Vermeulen NP. Inhibition of human glutathione S-transferases by curcumin and analogues. Xenobiotica 2009;39(4):302-11.
35. Strasser A. Molecular modeling and QSAR-based design of histamine receptor ligands. Expert Opin Drug Discov 2009;4(10):1061-75.
36. Lim HD, Istyastono EP, van de Stolpe A, Romeo G, Gobbi S, Schepers M, *et al.* Clobenpropit analogs as dual activity ligands for the histamine H_3 and H_4 receptors: Synthesis, pharmacological evaluation, and cross-target QSAR studies. Bioorg Med Chem 2009;17(11):3987-94.
37. Andrews SP, Brown GA, Christopher JA. Structure-based and fragment-based GPCR drug discovery. ChemMedChem 2014;9(2):256-75.
38. Matsuda H, Shimoda H, Morikawa T, Yoshikawa M. Phytoestrogens from the roots of *Polygonum cuspidatum* (Polygonaceae): Structure-requirement of hydroxyanthraquinones for estrogenic activity. Bioorg Med Chem Lett 2001;11(14):1839-42.
39. Hopert AC, Beyer A, Frank K, Strunck E, Wünsche W, Vollmer G. Characterization of estrogenicity of phytoestrogens in an endometrial-derived experimental model. Environ Health Perspect 1998;106(9):581-6.
40. Helferich WG, Andrade JE, Hoagland MS. Phytoestrogens and breast cancer: A complex story. Inflammopharmacology 2008;16(5):219-26.
41. Kamatou GP, Vermaak I, Viljoen AM. Eugenol – From the remote Maluku Islands to the international market place: A review of a remarkable and versatile molecule. Molecules 2012;17(6):6953-81.
42. Murphy PA, Barua K, Hauck CC. Solvent extraction selection in the determination of isoflavones in soy foods. J Chromatogr B Analyt Technol Biomed Life Sci 2002;777(1-2):129-38.