JPP Journal of Pharmacy and Pharmacology

JPP 2010, 62: 1717–1728
© 2010 The Authors
JPP © 2010 Royal Pharmaceutical Society of Great Britain
Received February 13, 2010
Accepted May 24, 2010
DOI
10.1111/j.2042-7158.2010.01154.x
ISSN 0022-3573

Research Paper

Molecular docking and QSAR studies of aromatase inhibitor androstenedione derivatives

Partha Pratim Roy and Kunal Roy

Drug Theoretics and Cheminformatics Laboratory, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Kolkata, India

Abstract

Objectives Aromatase (CYP19) inhibitors have emerged as promising candidates for the treatment of estrogen-dependent breast cancer. In this study, a series of androstenedione derivatives with CYP19 inhibitory activity was subjected to a molecular docking study followed by quantitative structure–activity relationship (QSAR) analyses in search of ideal physicochemical characteristics of potential aromatase inhibitors.

Methods The QSAR studies were carried out using both two-dimensional (topological, and structural) and three-dimesional (spatial) descriptors. We also used thermodynamic parameters along with 2D and 3D descriptors. Genetic function approximation (GFA) and genetic partial least squares (G/PLS) were used as chemometric tools for QSAR modelling.

Key findings The docking study indicated that the important interacting amino acids in the active site were Met374, Arg115, Ile133, Ala306, Thr310, Asp309, Val370, Leu477 and Ser478. The 17-keto oxygen of the ligands is responsible for the formation of a hydrogen bond with Met374 and the remaining parts of the molecules are stabilized by the hydrophobic interactions with the non-polar amino acids. The C2 and C19 positions in the ligands are important for maintaining the appropriate orientation of the molecules in the active site. The results of docking experiments and QSAR studies supported each other.

Conclusions The developed QSAR models indicated the importance of some Jurs parameters, structural parameters, topological branching index and E-state indices of different fragments. All the developed QSAR models were statistically significant according to the internal and external validation parameters.

Keywords CYP19; docking; GFA; G/PLS; QSAR

Introduction

Non-communicable chronic diseases, such as cancer, are fast replacing communicable diseases in India and other developing countries in terms of occurrence and impact. The burden of cancer is still an increasing concern worldwide in spite of the advancement of diagnosis and treatment.[1] Cancer is responsible for about 12.5% of deaths worldwide and it is estimated that the number of cancer patients with different types of cancer (such as breast, prostate, lung, uterine and cervical tumours) will be 15 million by 2020.[2,3] Breast cancer is one of the most common varieties of female cancer worldwide and the disease is generally supposed to be a major cause of morbidity and mortality in both pre- and postmenopausal women.[4] Approximately one-third of the breast cancer patients and two-thirds of postmenopausal breast cancer is estrogen dependent or estrogen receptor positive.[5,6] The proportion of tumours sensitive to estrogens increases with age and thus postmenopausal women are more susceptible to developing breast cancers than premenopausal women.[7] The role of estrogens in the development of breast cancer by activating the transcription factor for cancer cell proliferation has been established in earlier years.[5]

Estrogen is also an issue of concern in male physiology or reproduction, and the androgen/estrogen ratio is crucial for inducing different conditions such as apoptosis or ovulation of oocytes.[8] In males, androgen acts as a prohormone for the production of estrogens in the target cell.[9] Aromatase (estrogen synthetase) controls the androgen/estrogen ratio in vertebrates intercellularly.[8] Aromatase (CYP19: EC 1.14.14.1) is a member of the reticulum-bound cytochrome P450 super family. The two main components of the catalytic complex are the CYP19 enzyme with a crucial iron-binding porphyrin ring

Correspondence: Dr Kunal Roy, Drug Theoretics and Cheminformatics Laboratory, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India.
E-mail: kunalroy_in@yahoo.com

as a prosthetic group and NADPH–cytochrome P450 reductase as an electron donor. The main reaction catalysed by aromatase is the conversion of androstenedione and testosterone into estrone and estradiol.[10–12] Aromatase is the main enzyme responsible for the production of circulating estrogens in the peripheral tissue, such as liver, muscle, adipose and most importantly breast tumour tissue, to stimulate tumour cells, as estrogen is no longer made in the ovaries after menopause.[13] The peripheral production of estrogen necessitates the development of aromatase inhibitors for the treatment of breast cancer.[14] Aromatase expression is controlled by eight tissue specific promoters. In normal breast cells, promoter I.4 is expressed for transcription whereas aromatase expression is shifted from promoter I.4 to promoters I.3 and II in breast cancer cells.[15] Apart from this, recent study indicates that aromatase expression regulation is controlled by cyclooxygenase (COX)-I and COX-II inhibitors in SK-BR-3 breast cells.[16] Two main strategies have been applied most frequently for the treatment of estrogen-sensitive breast cancers. The first one is to block estrogen synthesis with inhibitors of aromatase and the other is the application of anti-estrogens to ameliorate the growth effects of estrogens on tumors.[17] Thus suppression of estrogen biosynthesis by aromatase inhibition represents an effective approach for the treatment of hormone-sensitive breast cancer.[18] Depending upon the chemical structure, aromatase inhibitors are grouped into steroidal (Type I) and non-steroidal (Type II) categories.[19–21] Differences in the modes of action of the two types of aromatase inhibitors are due to structural difference. The steroidal inhibitors produce irreversible inhibition by competing with the natural substrate for the active site of the enzyme (they act as false substrates and are processed to intermediates that bind irreversibly to the active site), whereas non-steroidal inhibitors produce reversible inhibition forming a coordinate bond with the heme iron (in the case of non-steroidal inhibitors, the coordinate bond is strong but reversible; the activity is regained after removal of the inhibitor from the active site).[14] The most important feature of the aromatase inhibitors is the coordination of the ligand with the iron atom of the heme moiety. In the case of steroidal inhibitors, the C19 methyl hydrogens coordinate with the heme moiety.[22] Over the past two decades, several steroidal (exemestan, formestane) and non-steroidal (anastrozole, letrozole) aromatase inhibitors have been developed and widely used as first-line drugs in breast cancers.[23–25] Several analogues of natural androstenedione have been found to have potent aromatase activity.[26–28] Long-term clinical use of aromatase inhibitors produces different adverse effects. Therefore, development of potent selective steroidal aromatase inhibitors is one of the major challenges in this century.[29–30]

The recently solved crystal structure of placental aromatase enzyme (pdb code 3EQM) allows us to study the critical interactions at the active site of the enzyme with the inhibitors.[31] Different docking studies were done on the theoretical 3D model of aromatase (e.g. pdb code 1TQA).[32–36] A few quantitative structure–activity relationship (QSAR) studies have also been reported on selected classes of aromatase inhibitors.[37,38]

The binding characteristics and interactions of steroidal aromatase inhibitors in the active site, as well the properties important for binding (electronic, hydrophobic and steric features), are required to be explored in designing more selective aromatase inhibitors. In connection with our previous study with non-steroidal aromatase inhibitors,[39] we have performed here a molecular docking study followed by QSAR analysis taking spatial, thermodynamic and structural descriptors and selected topological parameters using a series of androstenedione analogues to explore the important properties of potent aromatase inhibitors.[26,40–44] The novelty of this work is that we have used the recently solved crystal structure of placental aromatase enzyme (pdb code 3EQM)[31] for molecular docking in this study and the results obtained from docking have been cross-checked with QSAR studies.

Materials and Methods

The dataset

The inhibitory activity of a series of androstenedione analogues towards human CYP19 reported in the literature[26,40–44] has been used as the model data set for this study (Figure 1, Table 1). The inhibitory potencies of the compounds (IC50 (nM)) have been converted to the logarithmic scale (pIC50 (mM)) and then used for subsequent QSAR analyses as the response variable.

Molecular docking

The crystal structure of human placental aromatase cytochrome P450 in complex with androstenedione (EC: 1.14.14.1, 3EQM.pdb)[31] has been obtained from the RCSB protein data bank (http://www.pdb.org). The enzyme is co-crystallized with androstenedione, protoporphirin IX containing Fe^{3+} and phosphate ion. We performed the docking studies by using the LigandFit of Receptor-ligand interactions protocol section of Discovery Studio 2.1.[45] Initially there was a pretreatment process for both the ligands and the enzyme (aromatase). For ligand preparation, all the duplicate structures were removed and the options for ionization change, tautomer generation, isomer generation, Lipinski filter and 3D generator have been set true. For enzyme preparation, the whole enzyme was selected and hydrogen atoms were added to it. The pH of the protein was set in the range of 6.5–8.5. Then we defined the aromatase enzyme as the receptor and the active site was selected based on the ligand binding domain of bound ligand androstenedione. Then the pre-existing ligand (androstenedione) was removed and freshly prepared ligand (compound from the dataset in Figure 1) prepared by us was placed. Then from the receptor–ligand interaction section LigandFit was chosen. We used the pre-processed receptor and ligand as inputs. 'Dreiding' was selected as the energy grid. The conformational search of the ligand poses was performed by the Monte Carlo trial method. Torsional step size for polar hydrogen was set at 10. The docking was performed with consideration of electrostatic energy. Maximum internal energy was set at 10 000 Cal. Pose saving and interaction filters were set as default. Fifty poses were docked for each compound. During the procedure of docking, no attempt was made to minimize the ligand–enzyme complex (rigid docking). After completion of docking, the docked enzyme (protein–ligand complex) was analysed to investigate the type

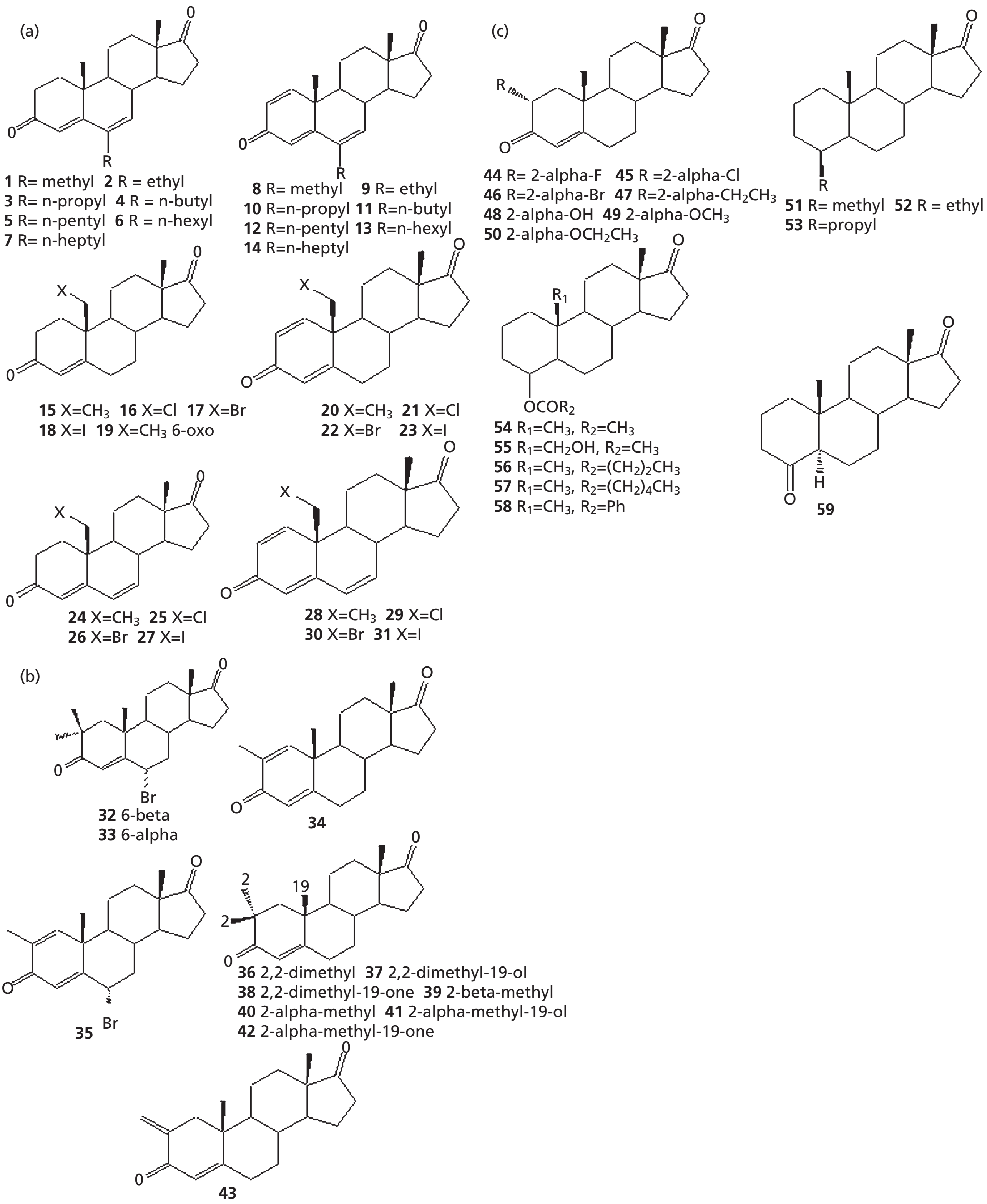


Figure 1 Structures of androstenedione analogues.

of interaction. Ten docking poses saved for each compound were ranked according to their dock score function. The pose (conformation) having the highest consensus dock score was selected and analysed to investigate the type of interaction.

Table 1 Observed and calculated inhibitory activity of a series of androstenedione analogues towards human CYP19 (pIC50)

Series No.	pIC50 (mM)						
	Obs[a]	Cal[b]	Cal[c]	Cal[d]	Cal[e]	Cal[f]	Cal[g]
Training set							
1	3.44	3.16	3.13	3.47	3.31	3.76	3.49
2	3.41	3.64	3.62	3.64	3.61	3.56	3.59
3	3.50	3.77	3.75	3.70	3.68	3.54	3.56
4	3.56	3.68	3.64	3.76	3.73	3.51	3.42
5	3.66	3.57	3.49	3.75	3.69	3.51	3.41
6	3.70	3.43	3.41	3.58	3.52	3.51	3.35
8	3.13	3.14	3.19	3.37	3.31	3.68	3.55
9	4.38	3.63	3.71	3.50	3.49	3.56	3.66
10	4.00	4.01	4.08	3.56	3.62	3.60	3.72
11	3.60	3.92	3.97	3.88	3.79	3.57	3.74
14	3.64	3.49	3.61	3.72	3.55	3.53	3.68
15	3.77	3.31	3.25	3.54	3.41	3.24	3.14
17	2.85	3.10	3.03	2.61	2.60	2.90	2.9
19	2.66	2.48	2.55	2.31	2.46	3.01	3.24
21	3.02	2.97	2.92	3.10	2.98	2.93	2.85
22	3.00	3.12	3.07	2.87	2.85	2.69	2.66
23	2.26	2.44	2.46	2.52	2.51	2.34	2.23
24	3.52	3.05	3.07	3.51	3.34	3.25	3.24
25	3.12	2.69	2.69	2.87	2.73	3.02	3.09
26	2.85	2.84	2.85	2.65	2.58	2.82	2.91
27	2.22	2.15	2.24	2.33	2.33	2.40	2.45
28	3.44	3.08	3.11	3.73	3.56	3.10	3.07
31	2.33	2.21	2.30	2.54	2.38	2.26	2.19
32	3.96	3.84	3.99	3.49	3.71	3.80	4.07
34	3.24	3.31	3.28	3.46	3.45	3.78	3.37
36	4.05	3.72	3.70	3.47	3.42	4.10	3.79
37	2.29	2.73	2.87	2.73	2.48	2.70	2.74
38	3.22	3.03	3.18	3.18	3.30	2.72	3.40
42	1.92	2.56	2.58	2.66	2.82	2.04	2.64
45	2.85	3.19	3.11	3.04	3.00	3.26	2.97
46	2.66	3.34	3.27	2.68	2.75	3.00	2.72
47	3.32	3.63	3.56	3.29	3.27	3.51	3.30
48	1.82	2.02	1.95	1.54	1.38	2.07	1.46
49	1.93	2.36	2.29	2.36	2.35	2.12	2.24
50	2.22	2.61	2.51	2.38	2.30	2.33	2.24
51	2.75	2.51	2.60	2.31	2.42	2.67	2.74
52	2.35	2.46	2.53	2.50	2.55	2.38	2.77
53	2.26	2.38	2.41	2.57	2.58	2.22	2.66
54	3.18	2.93	2.70	2.80	2.84	2.52	2.78
55	2.44	1.96	1.78	2.18	2.02	2.45	1.92
56	2.82	3.15	2.89	2.94	2.91	2.71	2.84
57	2.75	2.91	2.79	2.79	2.67	2.97	3.03
58	3.20	2.60	3.13	3.06	3.02	3.52	3.30
59	2.62	2.75	2.67	2.99	3.05	2.76	2.78
Test set							
7	3.37	3.29	3.32	3.36	3.29	3.42	3.26
12	3.85	3.8	3.81	3.95	3.83	3.53	3.80
13	3.81	3.65	3.72	3.88	3.74	3.53	3.77
16	3.10	2.95	2.88	2.84	2.75	3.13	3.11
18	2.39	2.39	2.42	2.28	2.39	2.55	2.44
20	3.80	3.33	3.29	3.71	3.55	3.30	3.04
29	2.88	2.72	2.74	3.06	2.88	2.88	2.76
30	2.68	2.87	2.89	2.85	2.67	2.64	2.61
33	4.00	3.84	3.99	3.65	3.82	3.77	4.06
35	3.15	3.61	3.63	3.50	3.60	3.45	3.65
39	3.60	3.26	3.20	3.38	3.36	3.69	3.36
40	3.24	3.26	3.20	3.33	3.33	3.69	3.24
41	1.96	2.26	2.24	2.24	2.13	2.02	2.09
43	3.26	3.32	3.27	3.72	3.59	3.53	3.38
44	3.01	2.93	2.83	2.94	2.77	2.96	2.47

[a]Observed CYP19 inhibitory activity;[26,40–44] [b]calculated from Equation 1; [c]calculated from Equation 2; [d]calculated from Equation 3; [e]calculated from Equation 4; [f]calculated from Equation 5 (equation not shown); [g]calculated from Equation 6 (equation not shown).

Validation of the docking process

Validation is the essential part of docking studies. For validation purposes we removed the pre-existing co-crystallized ligand and a 3D model of the ligand was freshly prepared and energy minimized. After that we docked the energy-minimized ligand and compared the binding site of pre-existing co-crystallized ligand and that of the freshly prepared ligand. These steps were performed to determine whether the docked ligand bound with the same amino-acid residues, as it got bound in the crystal structure of the enzyme, or bound differently to the enzyme.

Descriptors for QSAR

The analyses were performed using spatial (radius of gyration, Jurs descriptors, area, PMI-mag, density, Vm), thermodynamic (LogP, ALogP, ALogP98, MR, Molref) and structural (MW, hydrogen bond donor, hydrogen bond acceptor, chiral centers, No. of rotatable bonds) and topological descriptors, including E-state descriptors. For the calculation of 3D descriptors, multiple conformations of each molecule were generated using the optimal search as a conformational search method. Each conformer was subjected to an energy minimization procedure using smart minimizer under open force field (OFF) to generate the lowest energy conformation for each structure. The charges were calculated according to the Gasteiger method. All the descriptors were calculated using Descriptor+ module of the Cerius2 version 4.10 software running on a Silicon Graphics workstation.[46] Definitions of all descriptors can be found in the Cerius2 tutorial (available at http://www.accelrys.com). The terms that appear in the reported QSAR equations are explained in the Results and Discussion section. It may be mentioned here that the QSAR studies were carried out independently of the observations made in the docking study. No parameters obtained from the docking experiment were used as inputs for the QSAR study. Lists of important 2D and 3D descriptors are given in Tables S1 and S2 in the Supplementary Materials section.

QSAR model development

To begin the model development process, the data set ($n = 59$) was classified into clusters by using k-means cluster based on standardized topological, thermodynamic and structural descriptor matrix.[47] The numbers of compounds for the training and test sets were 44 and 15, respectively. QSAR models were developed using the training set compounds (optimized by Q^2), and then the developed models were validated (externally) using the test set compounds. QSAR models were generated separately for 2D descriptors and 3D descriptors. We used thermodynamic (physicochemical) descriptors both with 2D and 3D descriptors. Finally we developed QSAR models taking the combined set of descriptors. The chemometric tools used for QSAR model development were GFA (genetic function approximation) and G/PLS (genetic partial least squares).

The GFA technique[48,49] was used to generate a population of equations rather than one single equation for correlation between biological activity and properties. GFA involves the combination of multivariate adaptive regression splines (MARS) algorithm with genetic algorithm to evolve a population of equations that best fit the training set data. It provides an error measure, called the lack of fit (LOF) score, that automatically penalizes models with too many features. It also inspires the use of splines as a powerful tool for non-linear modelling. A distinctive feature of GFA is that it produces a population of models (e.g. 100), instead of generating a single model, as do most other statistical methods. The range of variations in this population gives added information on the quality of fit and importance of the descriptors.

The genetic partial least squares (G/PLS) algorithm[50,51] may be used as an alternative to a GFA calculation. G/PLS is derived from two QSAR calculation methods: GFA and partial least squares (PLS). The G/PLS algorithm uses GFA to select appropriate basis functions to be used in a model and PLS regression as the fitting technique to weigh the basis functions' relative contributions in the final model. Application of G/PLS thus allows the construction of larger QSAR equations while still avoiding overfitting and eliminating most variables.

Statistical qualities of QSAR models and model validation

The statistical qualities of the equations were judged by the parameters such as squared correlation coefficient (R^2) and variance ratio (F) at specified degrees of freedom (df).[52] For G/PLS equations, least-squares error (LSE) was taken as an objective function to select an equation, while lack-of-fit (LOF) was noted for the GFA-derived equations. The generated QSAR equations were validated by leave-one-out cross-validation R^2 (Q^2) and predicted residual sum of squares (*PRESS*)[53–55] and then were used for the prediction of enzyme inhibitory potency values of the test set compounds. The prediction qualities of the models were judged by statistical parameters such as predictive R^2 (R^2_{pred}), squared correlation coefficient between observed and predicted values of the test set compounds with (r^2) and without (r_0^2) intercept. It was previously shown that use of R^2_{pred} and r^2 might not be sufficient to indicate the external validation characteristics.[56] Thus, an additional parameter $r_{m\,(\text{test})}^2$ (defined as $r^2 \times \left(1-\sqrt{r^2-r_0^2}\right)$), which penalizes a model for large differences between observed and predicted values of the test set compounds, was also calculated. Two other variants[57,58] of r_m^2 parameter, $r_{m\,(\text{LOO})}^2$[59] and $r_{m\,(\text{overall})}^2$, were also calculated. The parameter $r_{m\,(\text{overall})}^2$ is based on prediction of both training (LOO prediction) and test set compounds. It was previously shown[58] that $r_{m\,(\text{LOO})}^2$ and $r_{m\,(\text{test})}^2$ penalize a model more strictly than Q^2 and R^2_{pred}, respectively. Another parameter R_p^2 ($R_p^2 = R^2 \times \sqrt{R^2 - R_r^2}$) ($R_r^2$ being squared mean correlation coefficient of random models) was also calculated[58] to check that the models thus developed were not obtained by chance.

Results and Discussion

Membership of the compounds in different clusters generated using *k*-means clustering technique is shown in Table S3 in the Supplementary Materials section. The test set size was set to approximately 25% of the total data set size[60] and the test set members are shown in Table 1.

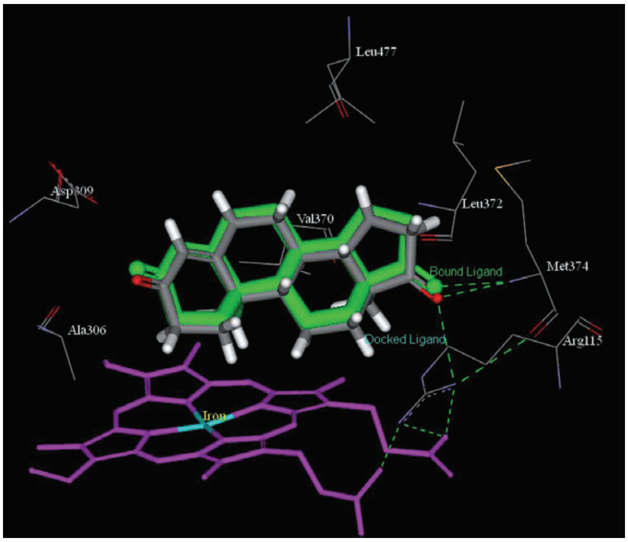


Figure 2 Superimposition of docked ligand and bound ligand (androstenedione) in the active site of human aromatase enzyme.

Molecular docking

To obtain an insight into the interactions between the human placental aromatase enzyme and its inhibitors and to explore the binding modes, a docking study was performed using the LigandFit tool available in Discovery Studio 2.1. The docking study indicated that all the androstenedione analogues bind in a similar fashion in the same binding mode. The important amino acids in the active site cavity (within 4Å) were Arg115, Asp309, Ser478 and Thr310 (polar amino acids) and Ala306, Ala307, Ile133, Ile305, Leu477, Met374, Phe134, Phe221, Trp224, Val369, Val370 and Val373 (non-polar amino acids) and this observation is in agreement with previous reports.[36,39,61] The reliability of the docking procedure was indicated by the low RMSD value (0.56Å) obtained between the bound ligand in the crystal structure and computationally freshly prepared docked ligand (Figure 2).

The amino acids responsible for important interactions with the ligands within the active site are Met374, Arg115, Ile133, Ala306, Thr310, Asp309, Val370, Leu477 and Ser478. All the compounds (high and low activity) form at least one hydrogen bond with Met 374. But the difference in the inhibitory activity between different compounds depends on the steric clashes of the compounds in the active site with important amino-acid residues as well as most importantly with the iron atom of the heme moiety. In the case of the most active compound in the dataset (compound **9**), it was observed that the17-keto oxygen forms two hydrogen bonds with the amide backbone of Met374 and one of the NH groups of Arg115 at a distance of 1.696 Å and 2.459 Å, respectively (Figure 3). A few steric clashes were observed between the ligand (compound **9**) with amino acids such as Ser478, Phe221 and Val373.

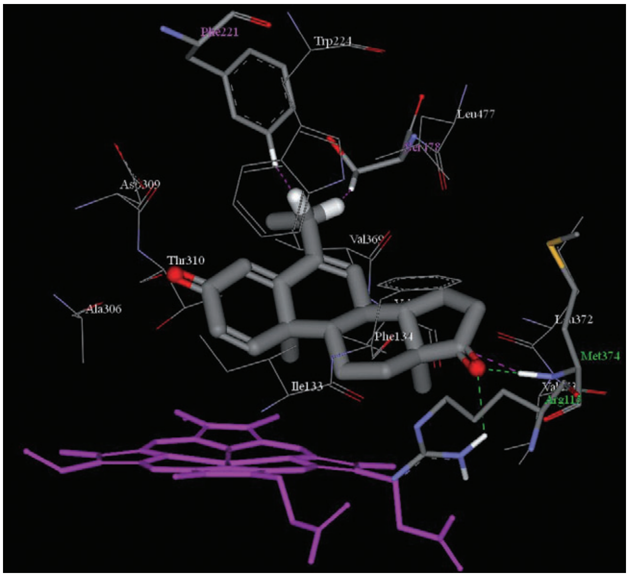


Figure 3 Docked conformation of compound **9** along with the important amino acid residues of human placental aromatase. Green dashed line indicates hydrogen bond formation; magenta dashed line indicates bump formation with amino acid residues.

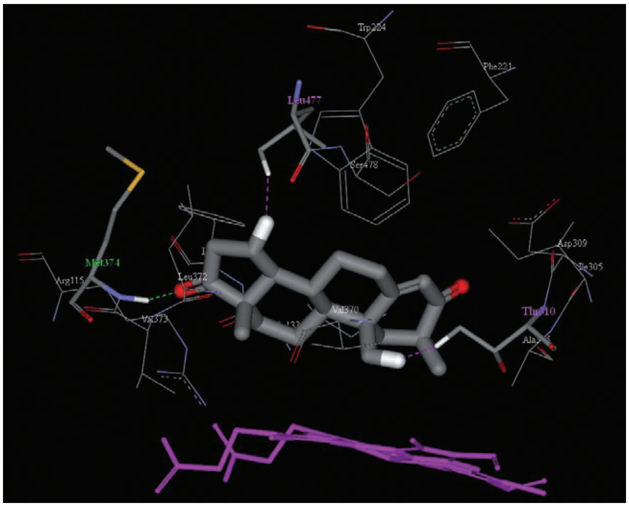


Figure 4 Docked conformation of compound **36** along with the important amino acid residues of human placental aromatase. Green dashed line indicates hydrogen bond formation; magenta dashed line indicates bump formation with amino acid residues.

In the case of another analogue in the high activity range (compound **36**), the 17-keto oxygen forms a hydrogen bond with the amide group of Met374 at a distance of 1.846 Å, while two steric bumps with Leu477 and Thr310 residues were observed (Figure 4). It was observed for both the compounds (**9**, **36**) that the 2β-hydrogen is close to Asp309 and orientation of the residue in that position facilitates enzyme acid–base-catalysed enolization process to selectively remove the 2β-hydrogen.[62]

In the case of the least active compound in the series (compound **48**), the 17-keto oxygen group forms a hydrogen bond with the –NH group of Met374 at a distance of 2.088 Å and a steric bump with Val373 (Figure 5). The unfavorable steric interactions occur with Thr310, Phe134 and Val370. The introduction of an –OH group at the C2 position changes the orientation of the molecule in such a fashion that a lot of unfavourable steric interactions with the heme moiety were observed, leading to decreased inhibitory activity.

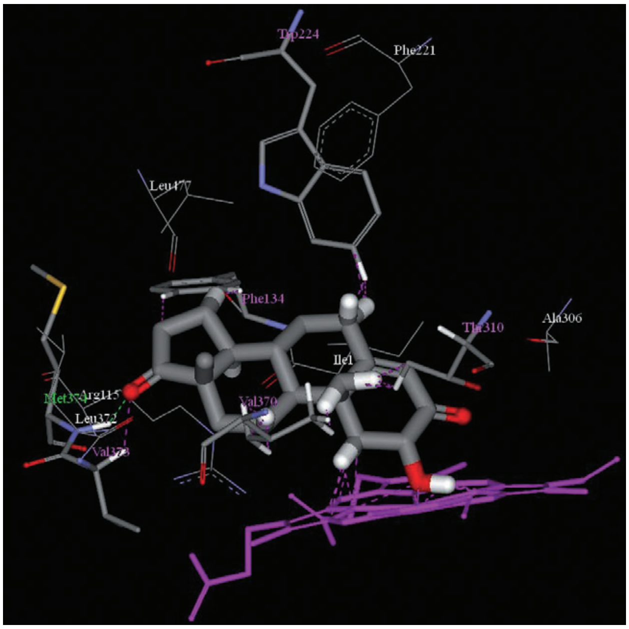


Figure 5 Docked conformation of compound **48** along with the important amino acid residues of human placental aromatase. Green dashed line indicates hydrogen bond formation; magenta dashed line indicates bump formation with amino acid residues.

Compound **37** is structurally closely related to compound **36**, although the former compound showed poor inhibitory activity. One of the C19 methyl hydrogens is replaced with an –OH group in compound **37**. Similarly to other compounds, the 17-keto oxygen of compound **37** forms a hydrogen bond with the amide backbone of Met374 at a distance of 1.864 Å (Figure 6). Steric bumps were observed with residues such as Leu477, Ala306, Ile133 and Thr310. The introduction of an –OH group at one C19 methyl hydrogen leads to detrimental interaction with the heme moiety, in turn leading to poor inhibitory activity. It was also observed that introduction of electronegative groups (–I, –Cl, –Br) at one of the C19 methyl hydrogens (e.g. **27**, **16** and **17**, respectively) changes the orientation of the molecules in the active site and produces unfavourable interactions with the iron atom of the heme moiety (data not shown).

Apart from the hydrogen bond formation with the 17-keto oxygen present on the cyclopentano ring system, other parts of the molecules are stabilized by hydrophobic interactions with the non-polar amino acids (Ala306, Trp224, Val369, Val370, Ile133, Phe134). This is in agreement with our previous observations with non-steroidal aromatase inhibitors.[39]

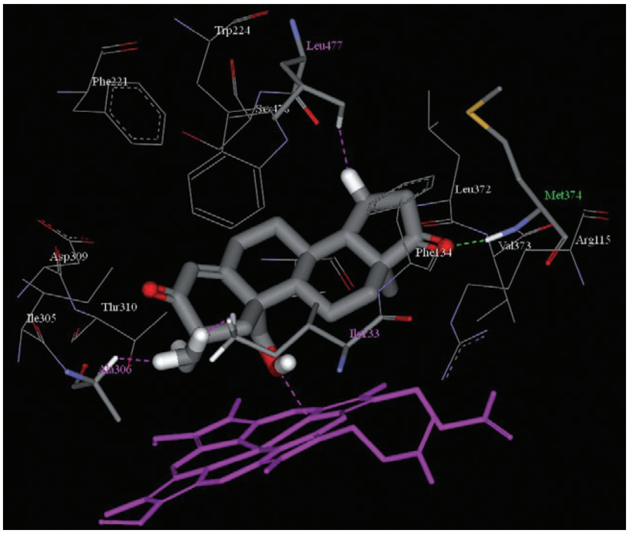


Figure 6 Docked conformation of compound **37** along with the important amino acid residues of human placental aromatase. Green dashed line indicates hydrogen bond formation; magenta dashed line indicates bump formation with amino acid residues.

The most important criterion for potent and selective inhibitors of CYP19 is coordination of the ligand with the iron atom of the heme moiety.[63] It has been reported in the literature that the C19 methyl hydrogens bind with the heme group within 4 Å distance. In our docking results for highly active compounds, the distance from the C19 methyl hydrogens to the heme group was within 4 Å. For the least active compounds (**48**, **37**), it was evident that the C19 methyl hydrogens were far away from the heme group or there were steric clashes with heme group.

Modelling with 2D descriptors

The following two equations (1 and 2) were among the best ones obtained from the GFA (5000 iterations) and G/PLS (1000 crossovers, linear terms, scaled variables and other default settings), respectively. Both linear and linear spline terms were used for development of the models.

$$\begin{aligned} &pIC50 = -1.565\,(\pm 1.040) - 0.976\,(\pm 0.114) <4.453 - \\ &ALogP98> - 0.482\,(\pm 0.085) S\_sI + 3.084\,(\pm 0.0622) JX \\ &- 2.694\,(\pm 0.563) <S\_ssssC - 0.266> n_{Training} = 44, \\ &LOF = 0.176,\ R^2 = 0.732,\ R^2{}_a = 0.704,\ F = 26.59(df\,4, 39), \\ &Q^2 = 0.648,\ r^2{}_{m(LOO)} = 0.482,\ n_{Test} = 15,\ R^2{}_{pred} = 0.847, \\ &r^2{}_{m(test)} = 0.779,\ r^2{}_{m(overall)} = 0.529 \qquad (1) \end{aligned}$$

The standard errors of regression coefficients are given within parentheses. Equation 1 could explain 70.4% of the variance (adjusted coefficient of variation) while it could predict 64.8% of the variance (leave-one-out predicted variance). The difference between R^2 and Q^2 values is not very high (<0.3).[64] When the equation was used to predict the CYP19 inhibition potency of the test set compounds, the predicted R^2 (R^2_{pred}) value was found to be 0.847. The r_m^2 values for the test, training and overall sets were found to be 0.779, 0.482 and 0.529, respectively. The relative importance of the descriptors according to their standardized regression coefficients is in the following order: $<4.453 - ALogP98> > S\_sI > JX > < S\_ssssC - 0.266>$.

The term $<4.453 - ALogP98>$ with negative regression coefficient indicates that the value of *ALogP98* should be more than 4.453 for good inhibitory activity. *ALogP98* (measure of hydrophobicity) is an atom-type-based *LogP* method using the published set of parameters.[65] The magnitude of this descriptor decreases with an increase of polar atoms in the molecule. It was observed that compounds such as **3**, **5**, **6**, **10**, **11**, **14**, **32** and **36** with *ALogP98* values greater than 4.453 showed better inhibitory activity than compounds with lower *ALogP98* values (compounds **37**, **48**, **55**). It was observed that introduction of the polar groups (=O, –OH) (as in compounds **19**, **37**, **48**, **55**) decreased the *ALogP98* value and these compounds showed poor inhibitory activity. This observation was supported by our docking study, which suggested that the compounds in the active site cavity are stabilized by hydrophobic interactions with the non-polar amino acids (Ala306, Trp224, Val370, Ile133, Phe134).

The next term with negative contribution is the E-state index of the fragment –I (*S_sI*). Only three compounds (**23**, **27**, **31**) in the training dataset had this fragment and they showed poor inhibitory activity. The position of the fragment –I is at C19. The docking study indicated that the presence of the group at this position produces an unfavourable interaction with the heme moiety by changing the orientation of the molecules in the active site. Therefore, docking results and QSAR-model-derived observations were in close agreement with each other.

The Balaban J index (*JX*), which characterizes the shape of the molecule based on covalent radii, is defined as follows:

$$J = \frac{q}{\mu + 1} . \sum \bar{V}_{Di} . \bar{V}_{Dj}$$

In this equation, i and j are the adjacent vertices, q is the number of edges, μ represents the number of cycles (i.e. $\mu = 0$ for linear graphs) and $\bar{V}_{Di}$ and $\bar{V}_{Dj}$ are the average distance sum of the vertex i and j, respectively.

The positive coefficient of the term indicates that compounds (such as **9** and **10**) with high values of the parameter have higher inhibitory activity than compounds (**53**, **56**, **57**) with low values of *JX*. It was observed that the value of *JX* decreased as the number of rings increased and the inhibitory potency of these compounds (like compound **58**) was also lower.

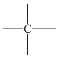
The term $< S\_ssssC - 0.266 >$ with negative coefficient indicates that for optimal inhibitory activity the value of the E-state index of fragment  (*S_ssssC*) should be less than 0.266. Compounds **51**, **52** and **53**, with *S_ssssC* values greater than 0.266, showed poor inhibitory activity. Compounds **9**, **10**, **11**, **14**, **28**, **32** and **36**, with *S_ssssC* values less than 0.266, showed good inhibitory activity. Compounds such as **19**, **21**, **31**, **37**, **42**, **48** and **55** showed poor inhibitory activity due to low values of *ALogP98* in spite of their low *S_ssssC* value.

Table 2 Comparison of statistical qualities of different models

Descriptor	Chemometric tool	Equation No.	R^2	Q^2	R^2_{pred}	$r_m^2{}_{(test)}$	$r_m^2{}_{(LOO)}$	$r_m^2{}_{(overall)}$
2D + Thermodynamic	GFA	1	0.732	0.648	0.847	0.779	0.482	0.529
	G/PLS	2	0.758	**0.705**	0.826	0.799	**0.674**	**0.710**
3D + Thermodynamic	GFA	3	0.763	0.673	0.850	**0.836**	0.498	0.547
	G/PLS	4	0.754	0.653	**0.864**	0.825	0.616	0.666
Combined	GFA	5 (not shown)	0.756	0.691	0.759	0.657	0.516	0.537
	GFA	6 (not shown)	**0.772**	0.728	0.825	0.789	0.547	0.585

The best values of different metrics are shown in bold print. GFA, Genetic function approximation; G/PLS, genetic partial least squares.

$$pIC50 = -4.803 - 0.962 <4.453 - ALogP98> + 4.302JX + 0.453 <2.505 - S\_sI> - 0.069 <25.007 - S\_dO> + 0.110 <SC\_3P - 66>$$
$$n_{Training} = 44,\ LSE = 0.094,\ R^2 = 0.758,\ R^2{}_{\alpha} = 0.726,\ F = 41.67(df\,3, 40),\ Q^2 = 0.705,\ r^2{}_{m(LOO)} = 0.674,\ n_{Test} = 15,\ R^2{}_{pred} = 0.826,\ r^2{}_{m(test)} = 0.799,\ r^2{}_{m(overall)} = 0.710 \quad (2)$$

The relative order of importance of the descriptors is: <4.453 – *ALogP98*> > *JX* > <2.505 – *S_sI*> > <25.007 – *S_dO*> > <*SC_3P* – 66>. The statistical quality of Equation 2 is listed in Table 2 along with that of other models.

The terms <4.453 – *ALogP98*> and *JX* have negative and positive regression coefficients in Equation 2, similarly to Equation 1.

The E-state index of the fragment –I (*S_sI*) should be less than 2.505 for good inhibitory activity as the term <2.505 – *S_sI*> shows positive contribution towards the activity. Similar observation was made also in equation 1 for the term *S_sI*.

The term <25.007 – *S_dO*> has negative impact on the inhibitory activity. This indicates that the E-state index of the fragment =O (*S_dO*) should be greater than 25.007 for the desired biological activity. Compounds like **51**, **52**, **53** and **55** have relatively small values of *S_dO* and significantly lower inhibition potential than compound **32** having a high value of *S_dO*. It may be mentioned here that the docking study showed that the 17-keto oxygen of the ligands is responsible for the formation of a hydrogen bond with Met374. However, compounds like **19**, **38** and **42** with *S_dO* values greater than 25.007 showed poor inhibitory activity due to low values of *ALogP98*.

The values of number of third-order subgraphs in a molecular graph (*SC_3P*) should be less than 66 for higher inhibitory activity as the term <*SC_3P* – 66> has a positive coefficient. Compounds like **3**, **4**, **9**, **10**, **11**, **15**, **24** and **36** with *SC_3P* values less than 66 showed a better inhibitory activity than compounds **37**, **57** and **58** with higher values of *SC_3P*.

Modelling with 3D descriptors

The following two equations (Equations 3 and 4) were among the best ones obtained from the GFA (5000 iterations) and G/PLS (1000 crossovers, scaled variables and other default settings), respectively. Both linear and linear spline terms were used for development of the models.

$$pIC50 = 16.625\,(\pm 1.771) - 27.108\,(\pm 3.603)Jurs\_RNCG - 0.003\,(\pm 0.0004)PMI\_mag - 0.181\,(\pm 0.028) <-26.205 - Jurs\_PNSA\_3> - 7.648\,(\pm 1.225)Jurs\_FPSA\_1 + 0.261\,(\pm 0.043)LogP$$
$$n_{Training} = 44,\ LOF = 0.169,\ R^2 = 0.763,\ R^2{}_{\mathrm{a}} = 0.725,\ F = 23.66\,(df\,5,38),\ Q^2 = 0.673,\ r^2{}_{m(LOO)} = 0.498,\ n_{Test} = 15,\ R^2{}_{pred} = 0.850,\ r^2{}_{m(test)} = 0.836,\ r^2{}_{m(overall)} = 0.547 \quad (3)$$

The relative importance of the descriptors according to their standardized regression coefficients is in the following order: *Jurs_RNCG* > *PMI_mag* > <– 26.205 – *Jurs_PNSA_3*> > *Jurs_FPSA_1* > *LogP*. The standard errors of regression coefficients are given within parentheses.

The negative coefficient of *Jurs_RNCG* indicates that it is detrimental to the inhibitory potency. The relative negative charge (RNCG) is defined as the partial charge of the most negative atom divided by total negative charge in the following manner: *Jurs_RNCG* = Q^-_{max}/Q^-, where Q^-_{max} is the charge of the most negative atom and Q^- is the total negative charge.

Compounds like **17**, **26**, **27**, **48**, **51**, **52**, **53** and **59** showed poor inhibitory activity due to high values of *Jurs_RNCG*. It was observed that the introduction of electronegative groups like –Br (compounds **17**, **26**), –I (**27**) or –OH (**48**) increased the value of *Jurs_RNCG* and contributed to poor inhibitory activity, supporting the docking result. On the other hand, compounds **5**, **6** and **14** with low values of *Jurs_RNCG* showed good inhibitory activity.

The principle moment of inertia (*PMI_mag*) has a detrimental effect on the inhibitory activity. *PMI_mag* is the moment of inertia, resultant of the moment of inertia of three axes, which are calculated for a series of straight lines through the centre of mass. Compounds like **1**, **15**, **24** and **28** with low values of *PMI_mag* showed better inhibitory activity than compounds **57** and **58** with high values of *PMI_mag*. But it is not true that all compounds with low values of the parameter have high activity: it was observed that compounds **51** and **59** showed poor inhibitory activity due to high values of *Jurs_RNCG*.

Jurs_PNSA_3 is the atomic charge weighted negative surface area and is calculated as $Jurs\_PNSA\_3 = \sum_{a-} q_a^- .SA_a^-$ (SA_a^- = atomic solvent accessible surface area of all negatively charged atoms and q_a^- = charge of overall negatively charged atoms). It is evident that this parameter is mainly governed by the total negative charge of the molecules. The negative coefficient of the term

Table 3 Randomization results for process of model development and developed models

Type of model	Chemometric tool	Equation no.	Process randomization			Model randomization		
			at 90% confidence level			at 99% confidence level		
			R^2	R_r^2	R_p^2	R^2	R_r^2	R_p^2
2D + Thermodynamic	GFA	1	0.732	0.177	0.545	0.732	0.073	0.594
	G/PLS	2	0.758	0.379	0.466	0.758	0.008	**0.657**
3D + Thermodynamic	GFA	3	0.763	0.127	**0.608**	0.763	0.110	0.617
	G/PLS	4	0.754	0.397	0.451	0.754	0.002	0.654
Combined	GFA	5	0.756	0.195	0.566	0.756	0.076	0.624
	GFA	6	0.772	0.158	0.605	0.772	0.078	0.643

The best value of the R_p^2 metric in each type of randomization test is shown in bold face. GFA, Genetic function approximation; G/PLS, genetic partial least squares.

<– 26.205 – *Jurs_PNSA_3*> indicates that when *Jurs_PNSA_3* has a value less negative than 26.205, it has detrimental effect on the inhibitory activity. Thus the absolute numerical value of *Jurs_PNSA_3* should be more than 26.205 for optimal inhibitory activity. Compounds like **9**, **11**, **14**, **34**, **36** and **47**, having absolute numerical values of *Jurs_PNSA_3* more than 26.205, showed good inhibitory activity while compounds like **19**, **31**, **42** and **48** with absolute values of *Jurs_PNSA_3* less than 26.205 had poor inhibitory activity. As the parameter *Jurs_PNSA_3* depends on the negatively charged atoms, the presence of an additional keto group at the 6 position (compound **19**) , –I, –CHO groups at the C19 position (compound **31**, **42**) and –OH group at the 2 position (compound **48**) decreased the inhibitory activity.

Jurs_FPSA_1 (fractional charged partial positive surface area) is obtained by dividing sum of the solvent-accessible surface areas of all positively charged atoms by the total molecular solvent-accessible surface area as follows:

$$Jurs\_FPSA\_1 = \frac{Jurs\_PPSA\_1}{SASA},$$

where $Jurs\_PPSA\_1 = \sum_{a+} SA_a^+$ is the sum of the solvent accessible surface areas of positively charged atoms.

Compounds with higher *Jurs_FPSA_1* values had detrimental effect on the inhibitory potency. For example, compounds **51**, **52**, **53**, **55** and **57,** having high values of *Jurs_FPSA-1*, had less CYP19 inhibitory activity.

The n-octanol/water partition coefficient (*LogP*) has a positive impact on the inhibitory activity as indicated by the positive regression coefficient. *LogP* is related to the hydrophobic character of the molecule. Compounds (like **5**, **9**, **10**, **11**, **14**, **15**, **32** and **36**) with high *LogP* value showed good inhibitory activity, while those (e.g. compounds **27**, **42**, **48**, **49** and **50**) with lower values had poor inhibitory activity. This observation is supported by our docking study, which suggests that the compounds in the active site cavity are stabilized by hydrophobic interactions with the non-polar amino acids (Ala306, Trp224, Val370, Ile133, Phe134).

Equation 3 was found to be statistically significant with explained variance of 73.1% and leave-one-out predicted variance of 67.3%. When the equation is applied on the test set of compounds, the R^2_{pred} value was found to be 0.850. Statistical significance of the model was also indicated by r_m^2 parameters listed in Table 2.

$$\begin{aligned} pIC50 = {} & 12.944 - 27.023 Jurs\_RNCG - 79.632 <- 0.050 - Jurs\_FNSA\_3> \\ & - 0.002 PMI\_mag - 0.010 <174.887 - Jurs\_PNSA\_1> \\ & - 1.409 Jurs\_FPSA\_2 + 0.165 LogP \\ & n_{Training} = 44,\ LSE = 0.097,\ R^2 = 0.754,\ R^2{}_{a} = 0.728, \\ & F = 29.90(df\,4,39),\ Q^2 = 0.653,\ r^2{}_{m(LOO)} = 0.616,\ n_{Test} = 15, \\ & R^2{}_{pred} = 0.864,\ r^2{}_{m(test)} = 0.825,\ r^2{}_{m(overall)} = 0.666 \end{aligned} \qquad (4)$$

The relative order of importance of the descriptors according to the standardized regression coefficient is: *Jurs_RNCG* > <– 0.050 – *Jurs_FNSA_3*> > *PMI_mag* > < 174.887 – *Jurs_PNSA_1*> *Jurs_FPSA_2* > *LogP*. Equation 4 was found to be statistically significant according to the internal and external validation parameters as listed in Table 2.

In Equation 4, the terms *Jurs_RNCG*, *PMI_mag* with negative coefficient and *LogP* with positive coefficient show similar results as in Equation 3.

The negative coefficient of the term <– 0.050 – *Jurs_FNSA_3*> indicates that the absolute value of *Jurs_FNSA_3* should be more than 0.050 for ideal aromatase inhibitors. For example, compounds **2**, **3**, **4**, **5**, **6**, **9** and **10** with absolute *Jurs_FNSA_3* values more than 0.050 showed good inhibitory activity, while compounds like **19**, **21**, **23**, **27**, **31**, **42** and **48**, having absolute *Jurs_FNSA_3* values less than 0.050, showed inferior inhibitory activity.

The partial negative charge surface area (*Jurs_PNSA_1*) is the sum of solvent accessible surface areas of all negatively charged atoms and is derived from the following equation:

$Jurs\_PNSA\_1 = \sum_{a-} SA_a^-$, where the sum is restricted to negatively charged atoms *a*–. The value of *Jurs_PNSA_1* should be less than 174.887 for optimum inhibitory activity as the term <174.887 – *Jurs_PNSA_1*> bears a negative regression coefficient. Compounds **23**, **27** and **31** with *Jurs_PNSA_1* values more than 174.887 showed poor inhibitory activity.

The results of the developed QSAR models with different combination of descriptors are listed in Table 2. As the qualities of the models developed with 2D and 3D descriptors in combination (Equations 5 and 6, not shown) were not superior to the models developed from 2D and 3D descriptors separately, the former equations are not described here. A comparison among various models shows that Equation 2 based on 2D descriptors shows the best internal validation characteristics while Equations 3 and 4 based on 3D descriptors are better than the 2D models in external validation characteristics. However, based on the $r_m{}^2{}_{(overall)}$ criterion, Equation 2 is the best one. Scatter plots of the observed vs calculated/predicted values of the training/test set compounds are shown in Figure S1 in the Supplementary Materials section.

The results of process and model randomization tests are shown in Table 3. The process randomization results of the G/PLS derived models (Equations 2 and 4) do not fulfill the required criterion of $R_p{}^2$ (the values being less somewhat lower than 0.5), although $R_p{}^2$ values for the model randomization are above the recommended cut off value. Based on the results on randomization tests, the GFA-derived Equation 3 is found to be more reliable than the other reported equations.

Overview and Conclusions

To explore the binding characteristics and important interactions of aromatase inhibitors in the active site, docking studies were carried out taking the crystal structure of human placental aromatase enzyme (pdb code: 3EQM).[31] In addition, QSAR studies were carried out taking spatial, thermodynamic, structural and topological descriptors to find out the properties of interest for ideal inhibitors. For the QSAR study the whole dataset ($n = 59$) was divided into training (n = 44) and test ($n = 15$) sets by *k*-means clustering techniques. The docking study indicates the presence of polar (Arg115, Asp309, Ser478, Thr310) and non-polar (Ala306, Ala307, Ile133, Ile305, Leu477, Met374, Phe134, Phe221, Trp224, Val369, Val370, Val373) amino acids in the active site. The important interacting amino acids were Met374, Arg115, Ile133, Ala306, Thr310, Asp309, Val370, Leu477 and Ser478. All the compounds form at least a single hydrogen bond with the amide backbone of the amino acid Met374. During the docking study, the positions of the steroidal nucleus found to be important were C19 and C2, supporting previously published reports.[31,39] It was observed that introduction of –OH or halogens at one of the methyl hydrogens or a –CHO group at C19 position changes the orientation of the molecules leading to unfavourable interactions with the heme moiety. This was also corroborated by the QSAR study. Both docking and QSAR studies indicate that hydrophobicity is an important determinant of the aromatase inhibitory activity. The developed QSAR models indicate the importance of different Jurs parameters (*Jurs_FNSA_3, Jurs_PNSA_3, Jurs_FPSA_2, Jurs_RNCG, Jurs_FPSA_1, Jurs_PNSA_1*), thermodynamic parameters (*ALogP98, LogP*), topological branching index (*SC_3P*), *JX* and E-state index for different fragments (*S_sI, S_ssssC, S_dO*), which are in close agreement with our previous study.[39] Modelling with both 2D and 3D descriptors indicates the importance of hydrophobicity and its optimal range for ideal aromatase inhibitors. The –I fragment at position C19 produces unfavourable molecular orientation of the ligands in the active site and in the QSAR study it shows detrimental effect on the inhibitory activity. The majority of the Jurs descriptors indicate the importance of electronegative hetero atoms in the molecules and their limiting range. It was evident that introduction of more negative atoms changes the polarity of the molecules and finally reduces the hydrophobicity, leading to poor inhibitory activity. The G/PLS models with 2D and 3D descriptors were found to be the best model according to internal and external validation statistics, respectively (Equation 2: $Q^2 = 0.705$, Equation 4: $R^2_{pred} = 0.864$). The best model among the comparable models based on the r^2_m $_{(overall)}$ criterion was the G/PLS model with 2D descriptors (Equation 2: $r^2_{m\ (overall)} = 0.710$). The results of the randomization test indicate the models were not obtained by chance.

Declarations

Conflict of interest

The Author(s) declare(s) that they have no conflicts of interest to disclose.

Funding

This work was supported by a Major Research Project of the University Grants Commission (UGC), New Delhi. PPR thanks the UGC, New Delhi for a fellowship.

References

1. Murthy NS, Mathew A. Cancer epidemiology, prevention and control. *Curr Sci* 2004; 86: 518–527.
2. Garcia M *et al*. *Global Cancer Facts & Figures 2007*, Atlanta, USA: American Cancer Society, 2007.
3. [Ministry of Health official in the Health National Cancer Institute, Brazil. Ministério da Saúde. Secretaria de Atenção à Saúde. Instituto Nacional de CânceCoordenação de Prevenção e Vigilância de Câncer. *Coordination of Prevention and Monitoring of Cancer. Estimativas 2008: Incidência de Câncer no Brasil. Estimates 2008: Incidence of Cancer in Brazil*. Rio de Janeiro: INCA, 2007. Rio de Janeiro: INCA, 2007] [in Portuguese].
4. Weigel NL, Rowan BG. Estrogen and progesterone action. In: DeGroot LJ *et al*., eds. *Endocrinology*. Philadelphia: WB Saunders, 2001: 2053–2060.
5. Cuzick J *et al*. The prevention of breast cancer. *Lancet* 1986; 8472: 83–86.
6. Clemons M, Goss P. Mechanisms of disease: estrogen and the risk of breast cancer. *N Engl J Med* 2001; 344: 276–285.
7. McGuire WL. An update on estrogen and progesterone receptors in prognosis for primary and advanced breast cancer. In: Iacobelli S *et al*. eds. *Hormones and Cancer*. New York: Raven Press, 1980: 337–344.
8. Séralini GE, Moslemi S. Aromatase inhibitors: past, present and future. *Mol Cell Endocrinol* 2001; 178: 117–131.
9. Simpson ER *et al*. Aromatase cytochrome P450, the enzyme responsible for estrogen biosynthesis. *Endocr Rev* 1994; 15: 342–355.
10. Hong Y *et al*. Molecular characterization of aromatase. *Ann NY Acad Sci* 2009; 1155: 112–120.
11. Kellis JT Jr, Vickery LE. Purification and characterization of human placental aromatase cytochrome P-450. *J Biol Chem* 1987; 262: 4413–4420.
12. Yoshida N, Osawa Y. Purification of human placental aromatase cytochrome P-450 with monoclonal antibody and its characterization. *Biochemistry* 1991; 30: 3003–3010.

13. Hemsell DL *et al*. Plasma precursors of estrogen: II. Correlation of the extent of conversion of plasma androstenedione to estrone with age. *J Clin Endocrinol Metab* 1974; 38: 476–479.
14. Buzdar A, Howell A. Advances in aromatase inhibition: clinical efficacy and tolerability in the treatment of breast cancer. *Clin Cancer Res* 2001; 7: 2620–2635.
15. Zhou D *et al*. Gene regulation studies of aromatase expression in breast cancer and adipose stromal cells. *J Steroid Biochem Mol Biol* 1997; 61: 273–280.
16. Díaz-Cruz ES *et al*. Cyclooxygenase inhibitors suppress aromatase expression and activity in breast cancer cells. *J Clin Endocrinol Metab* 2005; 90: 2563–2570.
17. Osborne CK *et al*. The value of estrogen and progesterone receptors in the treatment of breast cancer. *Cancer* 1980; 46: 2884–2888.
18. Brodie AMH, Njar VCO. Aromatase inhibitors in advanced breast cancer: mechanism of action and clinical implications. *J Steroid Biochem Mol Biol* 1998; 66: 1–10.
19. Banting L *et al*. Recent developments in aromatase inhibition as a potential treatment for oestrogen-dependent breast cancer. *Prog Med Chem* 1989; 26: 253–298.
20. Banting L. Inhibition of aromatase. *Prog Med Chem* 1996: 33: 147–184.
21. O'Reilly JM, Brueggemeier RW. 7 Alpha-arylaliphatic androsta-1,4-diene-3,17-diones as enzyme-activated irreversible inhibitors of aromatase. *J Steroid Biochem Mol Biol* 1996; l59: 93–102.
22. Guallar V *et al*. Peripheral heme substituents control the hydrogen-atom abstraction chemistry in cytochromes P450. *Proc Natl Acad Sci USA* 2003; 100: 6998–7002.
23. Plourde PV *et al*. ARIMIDEX: a new oral, once-a-day aromatase inhibitor. *J Steroid Biochem Mol Biol* 1995; 53: 175–179.
24. Lipton A *et al*. Letrozole (CGS 20267). A phase I study of a new potent oral aromatase inhibitor of breast cancer. *Cancer* 1995; 75: 2132–2138.
25. Evans TR *et al*. phase I and endocrine study of exemestane (FCE 24304), a new aromatase inhibitor, in postmenopausal women. *Cancer Res* 1992; 52: 5933–5939.
26. Numazawa M *et al*. 6-Alkylandrosta-4,6- diene-3,17-diones and their 1,4,6-triene analogs as aromatase inhibitors. *Steroids* 1997; 62: 595–602.
27. Harper-Wynne C, Dowsett M. Recent advances in the clinical application of aromatase inhibitors. *J Steroid Biochem Mol Biol* 2001; 76: 179–186.
28. Lønning PE. Pharmacology of new aromatase inhibitors. *The Breast* 1996; 5: 202–208.
29. Arora A, Potter JF. Aromatase inhibitors: current indications and future prospects for treatment of postmenopausal breast cancer. *J Am Geriatr Soc* 2004; 52: 611–616.
30. Goss PE. Risks versus benefits in the clinical application of aromatase inhibitors. *Endocr Relat Cancer* 1999; 6: 325–332.
31. Ghosh D *et al*. Structural basis for androgen specificity and estrogen synthesis in human aromatase. *Nature* 2009; 457: 219–223.
32. Favia AD *et al*. Three-dimensional model of the human aromatase enzyme and density functional parameterization of the iron-containing protoporphyrin IX for a molecular dynamics study of heme-cysteinato cytochromes. *Proteins* 2006; 62: 1074–1087.
33. Hong Y *et al*. Molecular basis for the aromatization reaction and exemestane-mediated irreversible inhibition of human aromatase. *Mol Endocrinol* 2007; 21: 401–414.
34. Hong Y *et al*. Molecular basis for the interaction of four different classes of substrates and inhibitors with human aromatase. *Biochem Phamacol* 2008; 75: 1161–1169.
35. Castellano S *et al*. CYP19 (aromatase): Exploring the scaffold flexibility for novel selective inhibitors. *Bioorg Med Chem* 2008; 16: 8349–8358.
36. Karkola S, Wähälä K. The binding of lignans, flavonoids and coumestrol to CYP450 aromatase: a molecular modelling study. *Mol Cell Endocrinol* 2009; 301: 235–244.
37. Nagar S *et al*. Pharmacophore searching of benzofuran derivatives for selective CYP19 aromatase inhibition. *Lett Drug Des Discov* 2009; 6: 38–45.
38. Nagar S *et al*. Pharmacophore mapping of flavone derivatives for aromatase inhibition. *Mol Divers* 2008; 12: 65–76.
39. Roy PP, Roy K. Docking and 3D-QSAR studies of diverse classes of human aromatase (CYP19) inhibitors. *J Mol Model* 2010 (in press). doi: 10.1007/s00894-010-0667-y.
40. Numazawa M *et al*. Studies directed towards a mechanistic evaluation of inactivation of aromatase by the suicide substrates androsta-1,4-diene-3,17- diones and its 6-ene derivatives. Aromatase inactivation by the 19-substituted derivatives and their enzymic aromatization. *J Steroid Biochem Mol Biol* 2007; 107: 211–219.
41. Numazawa, M *et al*. Synthesis and biochemical properties of 6-bromoandrostenedione derivatives with a 2,2-dimethyl or 2-methyl group as aromatase inhibitors. *Biol Pharm Bull* 2004; 27: 1878–1882.
42. Numazawa M *et al*. Probing the active site of aromatase with 2-methyl-substituted androstenedione analogs. *Steroids* 2003; 68: 503–513.
43. Numazawa, M *et al*. Structure–activity relationships of 2α-substituted androstenedione analogs as aromatase inhibitors and their aromatization reactions. *J Steroid Biochem Mol Biol* 2005; 97: 353–359.
44. Nagaoka M *et al*. Structure–activity relationships of 3-deoxy androgens as aromatase inhibitors. synthesis and biochemical studies of 4-substituted 4-ene and 5-ene steroids. *Steroids* 2003; 68: 533–542.
45. Discovery Studio 2.1 is a product of Accelrys Inc, San Diego, CA, USA.
46. Cerius2 version 4.10 is a product of Accelrys, Inc., San Diego, USA, http://www.accelrys.com/cerius2.
47. Leonard JT, Roy K. On selection of training and test sets for the development of predictive QSAR models. *QSAR Comb Sci* 2006; 25: 235–251, 45–48
48. Fan Y *et al*. Quantitative structure-antitumor activity relationships of camptothecin analogues: cluster analysis and genetic algorithm-based studies. *J Med Chem* 2001; 44: 3254–3263.
49. Rogers D, Hopfinger AJ. Application of genetic function approximation to quantitative structure–activity relationship and quantitative structure – property relationship. *J Chem Inf Comput Sci* 1994; 34: 854–866.
50. Dunn III WJ, Rogers D. Genetic partial least squares in QSAR In: Devillers J, eds. *Genetic Algorithms in Molecular Modeling*. London: Academic Press, 1996: 109–130.
51. Hasegawa K *et al*. GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists. *J Chem Inf Comput Sci* 1997; 37: 306–310.
52. Snedecor GW, Cochran WG. *Statistical Methods*. New Delhi: Oxford & IBH Publishing Co. Pvt. Ltd, 1967.
53. Wold S. PLS for multivariate linear modeling. In: van de Waterbeemd H, ed. *Chemometric Methods in Molecular Design*. Weinheim: VCH, 1995: 195–218.
54. Debnath AK. Quantitative structure–activity relationship (QSAR): a versatile tool in drug design. In: Ghose AK, Viswanadhan VN, eds. *Combinatorial Library Design and Evaluation*. New York: Marcel Dekker, Inc., 2001: 73–129.

55. Roy K On some aspects of validation of predictive QSAR models. *Expert Opin Drug Discov* 2007; 2: 1567–1577.
56. Roy PP, Roy K. On some aspects of variable selection for partial least squares regression models. *QSAR Comb Sci* 2008; 27: 302–313.
57. Roy K, Roy PP. Comparative QSAR studies of CYP1A2 inhibitor flavonoids using 2D and 3D descriptors. *Chem Biol Drug Des* 2008; 72: 370–382.
58. Roy PP *et al*. On two novel parameters for validation of predictive QSAR models. *Molecules* 2009; 14: 1660–1701.
59. Mitra I *et al*. On further application of r_m^2 as a metric for validation of QSAR models. *J Chemometrics* 2010; 24: 22–33.
60. Roy PP *et al*. Exploring the impact of the size of training sets for the development of predictive QSAR models. *Chemom Intell Lab Sys* 2008; 90: 31–42.
61. Murthy JN *et al*. Active site acidic residues and structural analysis of modelled human aromatase: a potential drug target for breast cancer. *J Comput-Aided Mol Des* 2006; 19: 857–870.
62. Graham-Lorence S *et al*. A three-dimensional model of aromatase cytochrome P450. *Protein Sci* 1995; 4: 1065–1080.
63. Cole PA, Robinson CH. Mechanism and inhibition of cytochrome P-450 aromatase. *J Med Chem* 1990; 33: 2933–2942.
64. Eriksson L *et al*. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ Health Perspect* 2003; 111: 1361–1375.
65. Ghose AK *et al*. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. *J Phys Chem* 1998; 102: 3762–3772.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1 Scatter plots of the observed vs calculated/predicted values of the training/test set compounds according to Equation 2
Table S1 List of values of important 2D descriptors
Table S2 List of values of important 3D descriptors
Table S3 *k*-Means clustering of compounds using standardized descriptors

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.