J Mol Model (2010) 16:1597–1616
DOI 10.1007/s00894-010-0667-y

ORIGINAL PAPER

Docking and 3D-QSAR studies of diverse classes of human aromatase (CYP19) inhibitors

Partha Pratim Roy · Kunal Roy

Received: 12 November 2009 / Accepted: 18 January 2010 / Published online: 1 March 2010

Abstract Aromatase (cytochrome 19) inhibitors have emerged as promising candidates for treatment of breast cancer. In search of potent aromatase inhibitors, docking and three-dimensional quantitative structure - activity relationship (3D-QSAR) studies using molecular shape, spatial, electronic, structural and thermodynamic descriptors have been performed on a diverse set of compounds having human aromatase inhibitory activities. An attempt has also been made to include two-dimensional (2D) descriptors in the QSAR studies. The chemometric tools used for model development are genetic function approximation (GFA) and genetic partial least squares (G/PLS). The docking study shows that the important interacting amino acids in the active site cavity are Met374, Arg115, Ile133, Ala306, Thr310, Asp309, Val370 and Ser478. One or more hydrogen bond formation with Met374 is one of the essential requirements for the ligands for optimum aromatase inhibition. The binding is further stabilized by van der Waals interactions with a few non-polar amino acid residues in the active site. The developed QSAR models indicate the importance of different shape, Jurs parameters, structural parameters, topological branching index and E-state index for different fragments. The results obtained from the QSAR analysis are supported by our docking observations. There should be one or two hydrogen bond acceptor groups (like –NO_2, -CN) and optimal hydrophobicity for ideal aromatase inhibitors. A GFA model with spline option obtained using 3D descriptors was found to be the best model based on internal validation (Q^2=0.668) while the best (externally) predictive model was a GFA model with spline option using combined set (2D and 3D) descriptors ($R_{pred}{}^2$=0.687). Based on $r_m{}^2{}_{(overall)}$ criterion, the best model was a G/PLS model (using 3D descriptors) with spline option ($r_m{}^2{}_{(overall)}$=0.606).

Keywords CYP19 · Docking · GFA · G/PLS · QSAR

Introduction

Breast cancer is the second leading cause of cancer death in women in the United States. About 180,000 women in the United States were found to have invasive breast cancer in 2007. Approximately over 2 million women living in the United States have been treated for breast cancer [1]. In post menopausal women, the estrogens are synthesized from adrenal C_{19} steroids in peripheral tissues like liver, muscles [2]. The role of endogenous estrogens in the development of breast cancer has long been recognized [3] and estrogens are known to play pivotal role in the proliferation of cancer cells [4]. In endocrine therapy two main approaches have been devised to antagonize the action of these hormones. The approaches are either to act directly at the estrogen receptor by means of antagonists like tamoxifen or by blocking the key target (like enzyme) of the process [5]. Two-thirds of breast cancers are hormone-dependent, contain estrogen receptors (ERs), and require estrogen for tumor growth. These patients are, therefore, suitable candidates for hormonal therapy, which targets blocking estrogen stimulation of breast cancer cells [6, 7]. Aromatase (P450 arom) is a mitochondrial enzyme consisting of cytochrome P450 (CYP450) heme protein and a NADPH cytochrome reductase. Cytochrome P450 is a

P. P. Roy · K. Roy (✉)
Drug Theoretics and Cheminformatics Lab,
Division of Medicinal and Pharmaceutical Chemistry,
Department of Pharmaceutical Technology, Jadavpur University,
Kolkata 700 032, India
e-mail: kunalroy_in@yahoo.com
URL: http://www.geocities.com/kunalroy_in

family of more than 60 important metabolizing enzymes. Aromatase (CYP 19) is one of the subfamilies of cytochrome P450s. Aromatase converts androgens to estrogens and is a particularly attractive target in the treatment of estrogen receptor positive breast cancer. Inhibitors of this enzyme are potential therapeutics for estrogen dependant breast cancers [8]. Aromatase inhibitors can be both steroidal and non-steroidal compounds [9–11].

Historically, the first clinically used aromatase inhibitor (aminoglutethimide) was marketed in the late 1970s [12]. Several reports showed advantages of nonsteroidal aromatase inhibitors over tamoxifen in adjuvant treatment. Therefore, aromatase inhibitors represent an interesting alternative in the first line therapy. Third generation aromatase inhibitor (AIs) which include two triazole derivatives, anastrozole (Arimidex) [13], letrozole (Femara) [14] and one steroidal analogue, exemestine (Aromasin) [15] are currently used clinically for the treatment of hormone dependant breast cancer in postmenopausal women [16–19]. However, the occurrence of important side effects associated with the prolonged clinical use of AIs (like the onset of resistance in the long-term treatment of the breast cancer, and a reduced efficacy in the treatment of the more advanced forms of the tumor) calls for the search of new, potent, more selective, and less toxic cytochrome 19 (CYP19) inhibitors [20, 21].

The recently solved crystal structure of human placental aromatase enzyme (pdb code 3EQM) [22] helps to understand the molecular basis for structure function characterization of human aromatase enzyme. Due to non availability of three dimensional (3D) crystal structure of aromatase until then, several docking studies were carried out [23–27] taking a theoretical 3-D model of aromatase (for example: pdb code 1TQA).

One of the most important features for strong inhibitor binding to the CYP enzymes is the capability to interact as the ligand with the iron atom of the heme group. Most of the non steroidal aromatase inhibitors of therapeutic importance act by binding to the enzyme *via* a competitive mechanism that involves coordination with heme iron [28]. Exploration of the binding characteristics of aromatase inhibitors in the active site as well as the properties important for binding, are of importance in designing more selective aromatase inhibitors. To our knowledge, the binding mode of ligands to the aromatase enzyme using 3EQM has not been reported earlier. In this context we have performed molecular docking followed by QSAR studies with molecular shape analysis descriptors along with thermodynamic and structural descriptors and also with selected topological parameters on structurally diverse datasets of aromatase inhibitors to explore the important properties of potent and selective aromatase inhibitors [29–40].

Methods and materials

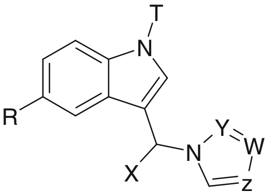
Dataset

Inhibitory activities of different classes of compounds toward human aromatase enzyme reported in the literature [29–40] have been used as the model data set for the present study (Tables 1 and 2). The experimental protocols for the determinations of enzyme inhibitory activities for all the compounds were the same. The quality of the data is good enough for QSAR studies as evidenced from small standard error values of individual observations. The inhibitory potencies of the compounds [$IC_{50}(\mu M)$] have been converted to the logarithmic scale [$pIC_{50}(mM)$] and then used for subsequent QSAR analyses as the response variable.

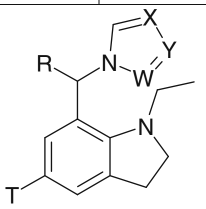
Docking

Crystal structure of human placental aromatase cytochrome P450 in complex with androstenedione (EC: 1.14.14.1, 3EQM.pdb) [22] has been obtained from the RCSB protein data bank (http://www.pdb.org). The enzyme is co-crystallized with androstenedione, protoporphirin IX containing Fe and phosphate ion. We have performed the docking studies by using LigandFit of receptor-ligand interactions protocol section of Discovery Studio 2.1 [41]. 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 has been selected and hydrogen atoms were added to it. The pH of the protein has been set in the range of 6.5 to 8.5. Then we have defined the aromatase enzyme as a total 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 a freshly prepared ligand (compound from the dataset in Table 1) prepared by us was placed. Then from the receptor- ligand interaction section LigandFit was chosen. We have used the preprocessed receptor and ligand as inputs. PLP1 was selected as the energy grid. The conformational search of the ligand poses was performed by 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 com-

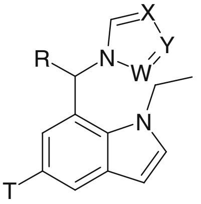
Table 1 Structural features of the diverse compounds [29–40] having aromatase inhibitory activity[a]



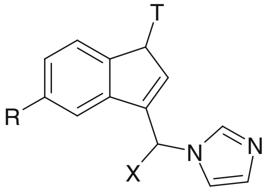
Sl	Isomerism	R	T	X	Y	W	Z
1	-	H	Et	H	C	N	C
2	-	H	Ph	H	C	N	C
3*	-	H	4-F-Ph	H	C	N	C
4*	-	H	Ph	H	N	C	N
5	-	H	4-F-Ph	H	N	C	N
6*	*R*	Br	Et	Ph	C	N	C
7	*R*	Br	Et	4-Cl-Ph	C	N	C
8*	*R*	Br	Et	Ph	N	C	N
9*	*R*	Br	Et	4-Cl-Ph	N	C	N

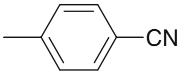
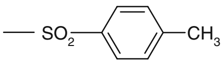


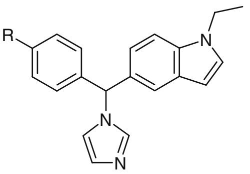
Sl	Isomerism	R	T	X	Y	W
10*	*R*	4-F-Ph	H	C	N	C
11	*R*	3-Cl-Ph	H	C	N	C
12	*R*	4-F-Ph	Br	C	N	C
13	*R*	4-F-Ph	Cl	C	N	C
14*	*R*	4-F-Ph	H	N	C	N



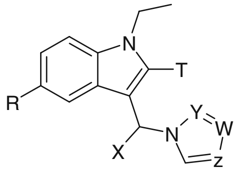
15	*R*	4-F-Ph	H	C	N	C
16	*R*	3-Cl-Ph	H	C	N	C
17	*R*	4-Cl-Ph	H	C	N	C
18	*R*	4-Br-Ph	H	C	N	C
19	*R*	4-F-Ph	Br	C	N	C



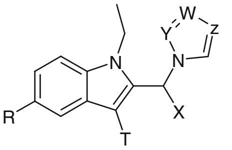
Sl	Isomerism	R	T	X
20	*R*	H	Et	4-F-Ph
21	*R*	Br	Me	4-F-Ph
22	-	H	2-Cl-benzyl	H
23	-	H		H
24	*R*	H		Ph



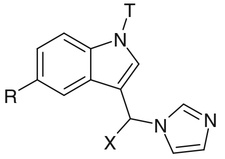
Sl	Isomerism	R
25	*R*	H
26	*R*	F



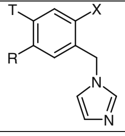
Sl	Isomerism	R	T	X	Y	W	Z
27*	*R*	H	Me	4-F-benzyl	C	N	C
28	*R*	H	Me	4-F-benzyl	N	C	N
29	*R*	Br	H	4-F-benzyl	C	N	C
30*	*R*	F	H	4-F-benzyl	C	N	C
31	*R*	CN	H	4-F-benzyl	C	N	C
32	*R*	Cl	H	4-F-benzyl	C	N	C



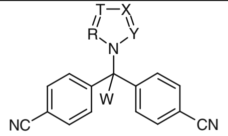
Sl	Isomerism	R	T	X	Y	W	Z
33*	*R*	H	Me	4-F-Ph	C	N	C
34	*R*	Br	H	4-F-Ph	C	N	C
35	*R*	Br	Me	4-Cl-Ph	C	N	C
36	*R*	Br	Me	Ph	C	N	C
37	*R*	Br	Me	3-Cl-Ph	C	N	C
38*	*R*	Br	Me	4-Cl-Ph	N	C	N
39	*R*	Br	Me	Ph	N	C	N
40	*R*	Br	H	4-F-Ph	N	C	N

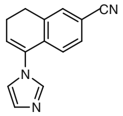
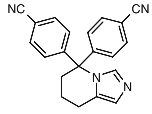
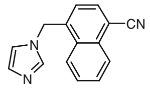
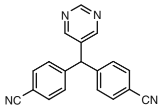
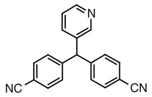
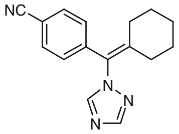
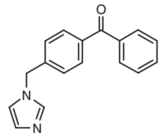
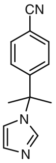
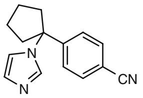
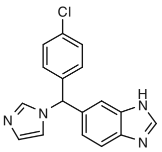
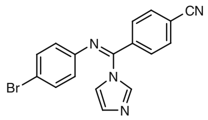


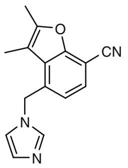
Sl	Isomerism	R	T	X
41	*R*	Br	*n*-Pr	4-F-Ph
42	*R*	Br	*i*-Pr	4-F-Ph

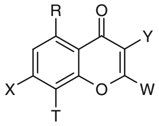


Sl	R	T	X
43	H	H	CN
44	H	Br	H
45	H	NO_2	H
46*	H	CN	H

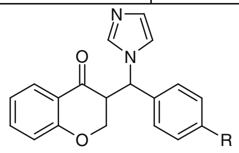


Sl	R	T	X	Y	W
47	C	N	C	C	H
48	N	C	N	C	H
49	N	N	C	C	H
50	N	N	C	N	H
51	N	C	N	C	Me
52	N	C	N	C	Et
53*	N	C	N	C	F
54*	N	N	N	C	F
55	N	N	C	N	F
56	C	N	C	C	F
57*					
58					
59					
60					
61					
62*					
63					
64					
65					
66*					
67					

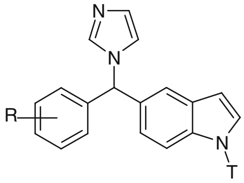
68*	



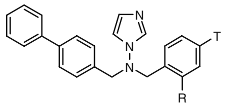
Sl	**R**	**T**	**X**	**Y**	**W**
69*	CN	-CH$_2$-Imidazol-1-yl	H	H	H
70	NO$_2$	-CH$_2$-Imidazol-1-yl	H	H	H
71	Br	-CH$_2$-Imidazol-1-yl	H	H	H
72	H	H	OMe	-CH$_2$-Imidazol-1-yl	Ph



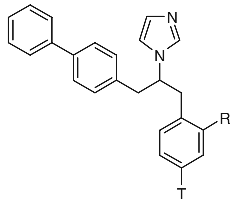
Sl	**Isomerism**	**R**
73	*R*	NO$_2$
74	*S*	NO$_2$
75	*R*	Br
76*	*S*	Br
77	*R*	CN
78	*S*	CN



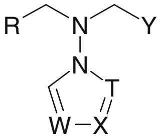
Sl	**Isomerism**	**R**	**T**
79	*R*	4-F	H
80	*R*	4-Cl	H
81	*S*	4-Cl	H
82	*R*	3-Cl	H
83*	*R*	4-Cl	Me
84	*R*	4-CN	H

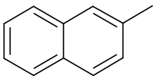
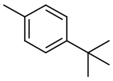
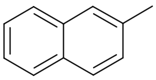
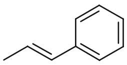
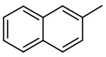
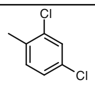
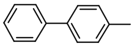
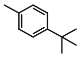
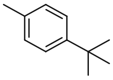
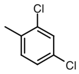
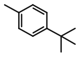
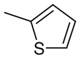
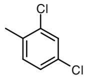
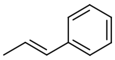
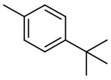
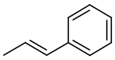
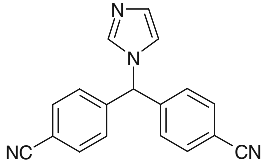
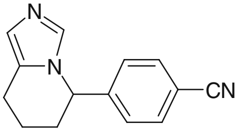


Sl	**R**	**T**
85*	H	H
86	Me	H
87*	Cl	H
88	F	H
89	H	Me
90	H	Cl
91*	H	F
92*	OMe	H
93	H	OMe
94	Cl	Cl
95	F	F



Sl	Isomerism	R	T
96*	*R*	H	*t*-Bu
97	*R*	H	H
98	*R*	Me	H
99	*R*	Cl	H
100	*R*	F	H
101	*R*	H	Me
102	*R*	H	F
103	*R*	OMe	H
104	*R*	H	OMe
105	*R*	Cl	Cl
106	*R*	F	F



Sl	R	T	X	Y	W
107		C	N		C
108		C	N		N
109*		C	N		C
110		C	N		C
111		C	N		C
112		C	N		C
113*		N	C		N
114		N	C		N
115*					
116	 (*S*)				

^a^Ph=Phenyl, Me= Methyl, Et=Ethyl, *R* = Rectus, *S* = Sinister

* indicates test set compounds

Table 2 Observed and calculated aromatase inhibitory activity of different classes of compounds

Sl	Obs[a]	Cal[b]	Cal[c]	Cal[d]	Cal[e]
Training set					
1	2.446	3.074	3.640	3.981	2.836
2	4.003	3.478	3.840	3.679	4.294
5	3.699	4.027	3.985	3.619	3.846
7	3.928	3.433	3.206	3.601	3.338
11	3.959	3.638	3.829	3.952	3.938
12	4.046	3.648	3.446	3.657	3.887
13	4.222	3.890	3.755	3.812	3.902
15	4.222	4.573	4.554	4.082	4.406
16	3.77	4.093	4.037	4.082	4.254
17	4.222	4.144	4.043	4.082	4.136
18	4.155	3.915	3.742	3.955	3.569
19	3.699	4.134	3.668	3.815	4.140
20	4.222	4.130	4.380	4.293	4.110
21	4.097	3.878	3.733	4.091	3.887
22	4.301	3.489	3.849	3.929	3.881
23	4.301	4.064	4.807	4.503	4.164
24	4.301	4.457	4.698	3.821	3.090
25	3.678	3.474	3.910	4.173	3.502
26	4.398	4.149	4.551	4.112	4.186
28	4.523	3.596	3.769	3.913	3.832
29	4.301	3.400	3.586	3.673	3.503
31	3.854	4.212	4.594	4.568	4.120
32	3.824	3.631	3.903	3.828	3.564
34	3.62	3.849	3.533	3.660	3.893
35	3.387	2.991	2.832	3.059	3.053
36	3.377	3.454	3.354	3.413	3.027
37	3.027	2.963	2.757	3.059	3.144
39	2.485	3.152	2.649	3.291	3.554
40	2.461	3.531	3.205	3.537	3.762
41	3.495	3.466	3.479	3.519	3.427
42	3.469	3.458	3.472	3.521	3.449
43	3.523	4.394	4.256	4.919	4.687
44	4.071	4.264	3.821	4.432	4.627
45	5.222	4.837	4.882	4.312	4.271
47	5.398	5.839	4.788	4.931	5.505
48	4.949	5.421	5.114	4.929	5.272
49	4.921	4.687	4.920	4.929	4.736
50	6	4.986	5.424	4.928	4.991
51	5.046	4.942	4.559	4.757	4.739
52	4.745	4.681	3.918	4.674	4.619
55	4.523	4.589	5.175	4.743	4.838
56	5.222	5.036	4.756	4.746	4.895
58	3.638	4.840	4.714	4.472	4.645
59	5.699	4.755	4.517	4.543	4.797
60	4.155	4.687	4.893	4.905	4.759
61	4.921	5.129	4.972	4.906	4.695
63	4.678	4.409	4.599	3.950	4.224
64	5.097	4.723	4.434	4.700	4.974
65	4.678	4.727	4.416	4.437	5.041
67	5.097	3.976	4.454	4.357	4.013
70	2.959	3.357	3.523	3.482	3.700
71	2.678	3.518	3.240	3.445	3.864
72	3.26	2.371	3.587	3.439	3.445
73	4.745	4.065	4.180	3.992	3.893
74	3.155	3.827	4.008	3.992	3.706
75	4.602	4.385	4.215	4.112	4.162
77	4.431	4.186	4.342	4.637	4.257
78	3.27	3.768	4.171	4.637	4.139
79	4.58	5.012	4.683	4.263	4.691
80	4.347	4.767	4.392	4.263	4.695
81	5.046	4.180	4.421	4.263	4.145
82	4.527	4.704	4.423	4.263	4.600
84	4.714	4.636	4.880	4.788	4.710
86	2.529	3.125	3.132	2.890	3.191
88	3.334	3.363	3.219	3.076	3.347
89	2.658	2.943	3.165	2.921	2.931
90	2.926	2.929	2.852	2.892	2.783
93	2.815	2.969	3.261	3.269	2.940
94	2.438	2.802	2.262	2.507	2.662
95	3.453	3.449	3.026	2.937	3.462
97	3.023	3.144	3.217	3.489	3.452
98	2.983	3.014	2.993	3.133	3.172
99	2.963	3.318	3.181	3.105	3.106
100	2.863	3.528	3.589	3.319	3.378
101	2.879	2.959	2.920	3.164	3.195
102	3.947	3.726	3.781	3.350	3.464
103	3.291	2.864	3.349	3.482	2.868
104	2.774	3.191	3.688	3.512	2.945
105	2.907	3.171	2.926	2.751	2.754
106	3.59	3.714	3.501	3.180	3.489
107	2.338	2.815	2.772	2.529	2.475
108	1.885	2.558	2.783	1.998	2.197
110	2.666	2.409	2.503	2.082	2.294
111	2.818	2.622	1.912	2.721	2.802
112	3.237	2.862	3.172	3.800	3.326
114	2.296	2.323	2.277	1.767	2.365
116	5.495	4.613	4.516	5.196	4.865
Test set					
3	4.144	4.271	4.466	3.619	4.111
4	3.509	3.682	4.019	3.679	4.054
6	4	3.606	3.796	3.955	3.573
8	2.52	3.364	3.295	3.903	3.524
9	3.162	3.191	3.295	3.549	3.125
10	4.222	4.100	4.334	4.082	4.142
14	3.301	3.877	3.785	4.082	3.905
27	4.523	3.704	4.348	3.962	3.862

Table 2 (continued)

Sl	Obs[a]	Cal[b]	Cal[c]	Cal[d]	Cal[e]
30	4.222	4.048	4.376	4.022	3.762
33	3.921	3.849	4.142	3.783	3.971
38	2.726	3.143	3.278	2.937	3.228
46	5	4.888	4.436	4.932	5.031
53	4.886	4.538	4.432	4.622	4.638
54	4.678	4.092	5.682	4.744	4.420
57	5.523	4.365	4.425	4.316	4.838
62	5	4.121	3.833	4.470	4.695
66	4.357	4.278	4.195	4.263	4.122
68	4.456	4.392	4.232	4.395	4.597
69	3.62	3.942	3.818	4.038	4.452
76	3.44	3.944	3.718	4.112	3.998
83	4.625	3.925	4.200	4.173	3.993
85	2.919	3.118	3.459	3.246	3.155
87	3.521	3.005	2.691	2.862	3.069
91	3.712	3.269	3.350	3.106	3.249
92	2.82	2.696	2.963	3.239	2.930
96	3.001	2.386	2.265	2.325	2.536
109	2.398	3.077	2.781	2.955	3.136
113	1.766	2.454	2.424	2.229	2.522
115	4.469	5.406	4.787	4.931	5.133

Obs[a] = [a]Observed aromatase inhibitory activity [29–40]; cal[b] = [b]Calculated from Eq. 1; Cal[c] = [c]Calculated from Eq. 2; Cal[d] = Calculated from Eq. 3; Cal[e] = Calculated from Eq. 4

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

Validation of the docking process

Validation is the essential part of docking studies. For validation purpose we have removed the preexisting co-crystallized ligand and 3D model of the ligand was freshly prepared (newly developed in silico model of the compound) and energy minimized. After that we have docked the energy minimized ligand and compared the binding site of preexisting co-crystallized ligand and that of the freshly prepared ligand. These steps are performed to determine whether the docked ligand binds with the same amino acid residues, as it got bound in the crystal structure of the enzyme, or it binds differently to the enzyme.

Descriptors

The analyses were performed using spatial (Radius of gyration, Jurs descriptors, Shadow indices, Area, PMI-mag, Density, Vm), shape (DiFFV, Fo, NCOSV, COSV, Shape RMS), thermodynamic (AlogP, AlogP98, Molref) and structural (MW, hydrogen bond donor, hydrogen bond acceptor, chiral centers, number 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 [42]. Definitions of all descriptors can be found at the Cerius2 tutorial available at the website htt://www.accelrys.com.

Model development

It was our priority to construct QSAR models which were statistically robust both internally as well as externally. The main target of any QSAR modeling is that the developed model should be robust enough to be capable of making accurate and reliable predictions of biological activities of new compounds. So, QSAR models which are developed from the training set should be validated using new chemical entities for checking the predictive capacity of the developed models. That is why the original data set is divided into training and test sets for QSAR model development and validation respectively. The ability of a model to predict accurately the target property of compounds that were not used for model development is based on the fact that a molecule which is structurally very similar to the training set molecules will be predicted well because the model has captured features that are common to the training set molecules and is able to find them in the new molecule [43]. On the other hand, a new molecule which has very little in common with the training set data should not be predicted very well, *i.e.*, the confidence in its prediction should be low. The selection of training and test sets should be based on the proximity of the representative points of the test set to representative points of the training set in the multidimensional descriptor space. In our study, the whole data set (n=116) was divided into training (n=87) and test (n=29) sets by *k*-means clustering techniques based on the standardized 2D variables [43]. This approach (clustering) ensures that the similarity principle can be employed for the activity prediction of the test set [44]. The splitting has been performed such that points representing both training and training sets are distributed within the whole descriptor space occupied by the entire dataset, and each point of the test set is close to at least one point of the training set. 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. For the development of the QSAR/QAAR models the statistical techniques used were genetic function approximation (GFA) and genetic partial least squares (G/PLS)

For the computation of shape analysis descriptors, the major steps are (1) generation of conformers and energy minimization; (2) hypothesizing an active conformer (global minimum of the most active compound, though we must acknowledge that minimum energy conformation of an isolated molecule may not be same as that of the molecule bound to the target site); (3) selecting a candidate shape reference compound (based on active conformation); (4) performing pairwise molecular superimposition using the maximum common subgroup [MCSG] method; (5) measuring molecular shape commonality using MSA descriptors; (6) determination of other molecular features by calculating spatial, electronic, and conformational parameters; (7) selection of conformers; and (8) generation of QSAR equations by genetic function algorithm (GFA). Optimal search was used as a conformational search method. The global minimum energy conformer of the most active compound [**50** having the highest pIC_{50}(mM) value] was selected as a shape reference to which all the structures in the study compounds were aligned through pairwise superpositioning. The method used for performing the alignment was a maximum common subgroup (MCSG) [42, 45]. This method looks at molecules as points and lines and uses the techniques of graph theory to identify patterns. It finds the largest subset of atoms in the shape reference compound that is shared by all the structures in the study table and uses this subset for alignment. A rigid fit of atom pairings was performed to superimpose each structure so that it overlays the shape reference compound. Finally additional electronic, spatial and thermodynamic descriptors were also calculated.

Genetic function approximation (GFA) technique [46, 47] was used to generate a population of equations rather than one single equation for correlation between biological activity and physicochemical properties. GFA involves the combination of multivariate adaptive regression splines (MARS) algorithm with genetic algorithm to evolve 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 modeling. 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 [48, 49] 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 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*) [50]. 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*) [51–53] and then were used for the prediction of enzyme inhibition activity values of the test set compounds. The prediction qualities of the models were judged by statistical parameters like predictive R^2 (R_{pred}^2), 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_{pred}^2 and r^2 might not be sufficient to indicate the external validation characteristics [54]. Thus, an additional parameter $r_{m}^{2}{}_{(test)}$ [defined as $r^2*(1-\sqrt{r^2-r_0^2})$], which penalizes a model for large differences between observed and predicted values of the test set compounds, was also calculated. Two other variants [55, 56] of r_m^2 parameter, $r_{m}^{2}{}_{(LOO)}$ [57] and $r_{m}^{2}{}_{(overall)}$, were also calculated. The parameter $r_{m}^{2}{}_{(overall)}$ is based on prediction of both training (LOO prediction) and test set compounds. It was previously shown [56] that $r_{m}^{2}{}_{(LOO)}$ and $r_{m}^{2}{}_{(test)}$ penalize a model more strictly than Q^2 and R_{pred}^2 respectively. Another parameter R_p^2 ($R_p^2 = R^2*\sqrt{R^2-R_r^2}$) ($R_r^2$ being squared mean correlation coefficient of random models) was also calculated [56] to check whether the models thus developed are not obtained by chance.

Results and discussion

Membership of compounds in different clusters generated using *k*-means clustering is shown in Table 3. The test set size was set to approximately 25% to the total data set size [58] and the test set members are shown in Table 3.

Docking

In the present study, to understand the interactions between human placental aromatase enzyme and its inhibitors, and

Table 3 *k*-Means clustering of compounds using standardized descriptors

Cluster No	No. of compounds in cluster	Compounds (Sl nos.) in different clusters																
1	14	**1**	**43**	**44**	**45**	**46**	**57**	**59**	**64**	**65**	**68**	**69**	**70**	**71**	**116**			
2	16	**2**	**3**	**4**	**5**	**15**	**23**	**25**	**26**	**63**	**66**	**79**	**80**	**81**	**82**	**83**	**84**	
3	68	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**24**
		27	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**72**
		73	**74**	**75**	**76**	**85**	**86**	**87**	**88**	**89**	**90**	**91**	**92**	**93**	**94**	**95**	**96**	**97**
		98	**99**	**100**	**101**	**102**	**103**	**104**	**105**	**106**	**107**	**108**	**109**	**110**	**111**	**112**	**113**	**114**
4	18	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**58**	**60**	**61**	**62**	**67**	**77**	**78**	**115**

to explore their binding mode, a docking study was performed using the LigandFit tool available in Discovery Studio 2.1 [41]. The specific cleft in which the ligands bind (within 4 Å) contains both polar (Arg115, Arg375, Asp309, Asp371, Ser478, Thr310, Asp371, Glu302) and non polar (Ala306, Ala307, Ile133, Ile305, Leu477, Met374, Phe134, Phe221, Trp224, Val369, Val370, Val373) amino acids and this is in agreement with previous reports [27, 59]. The crystal structure of human placental aromatase [22] shows that the bound ligand androgen makes a hydrogen bond with the backbone amide of Met374. Our docking study with LigandFit using the freshly prepared model of the ligand (androstenedione) also corroborates similar observation indicating the reliability of the docking procedure (Figs. 1 and 2). Figure 1 shows X-ray crystal structure of the protein along with the ligand (experimentally obtained) while Fig. 2 shows docked conformation of the ligand within the enzyme cavity. In both cases, the ligand forms hydrogen bond with Met374 and interacts with amino acids like Asp309, Ala306, Arg115, Leu477 and Leu 372.

The results obtained in the docking study indicates the important amino acids in the active site cavity responsible for important interactions are Met374, Arg115, Ile133, Ala306, Thr310, Asp309, Val370, Ser478. All the compounds in the high activity range from one or two hydrogen bond(s) with amide backbone of Met374 at a distance ranging from 1.58–2.30 Å. In case of compound **45,** the nitro (-NO_2) group forms two hydrogen bonds at 2.293 Å and 2.034 Å (Fig. 3). The same nitro group also forms another hydrogen bond with Arg115 (2.397 Å) (Fig. 3) and this compound (**45**) shows good inhibitory activity. Compound **59** forms two hydrogen bonds (Fig. 4), one between the –CN group of the ligand and the amide back bone of Met374 and the other between the NH fragment of the azole nucleus and the side chain hydroxyl group of Thr310. In spite of the steric bump formation with Ile133, this compound possesses good inhibitory activity due to the hydrogen bonds. In case of compound **116,** apart from the hydrogen bond with Met374 (using the –CN group), there is a steric bump formation with the polar amino acid Asp309 (Fig. 5). The docking results also suggest that apart from hydrogen formation with Met374 and/or Arg115, binding of different compounds with the active pocket is stabilized by van der Waals interactions with the non polar amino acids (Ala306, Thr310, Trp224, Val370, Ile133, Phe134, Leu372, Val373). It can also be mentioned that the ligands should contain hydrogen bond acceptor groups (like –NO_2, -CN) for hydrogen bond formation with Met374, Arg115 and/or Thr310 in the active site for good aromatase inhibition. The azoles family is going to hold an increasingly prominent position in development of aromatase inhibitors [13, 14]. The reason is that the azoles moiety is

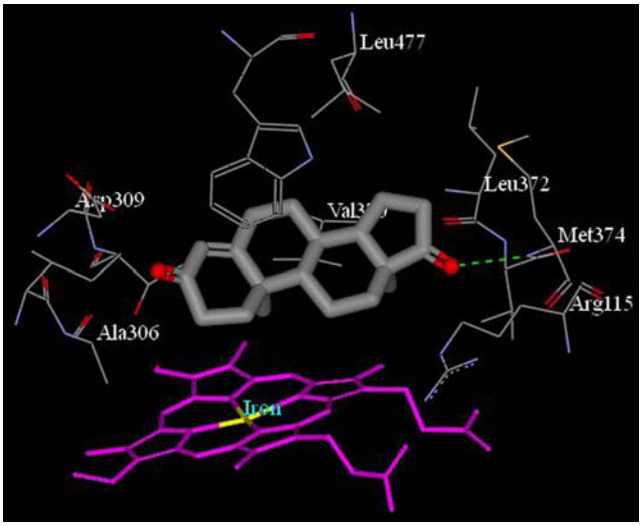


Fig. 1 Bound ligand (androstanedione) in the active site of human placental aromatase (X-ray crystal structure) [important interacting amino acids and iron in heme have been labeled]

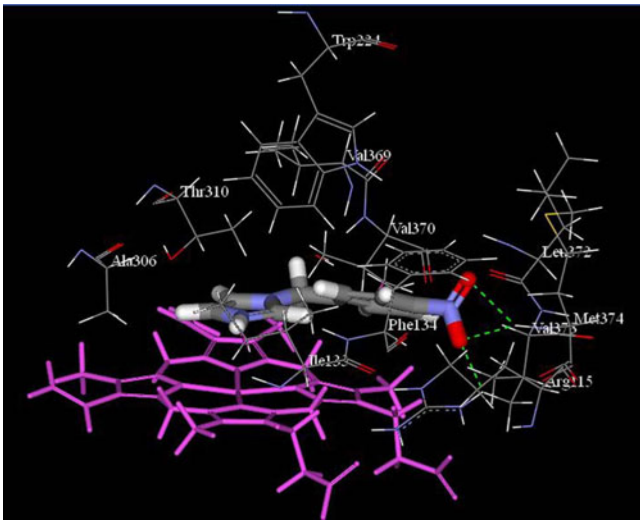


Fig. 3 Docked conformation of compound **45** along with the important amino acid residues of human placental aromatase: the nitro (-NO_2) group of **45** forms two hydrogen bonds at 2.293 Å and 2.034 Å; the same nitro group also forms another hydrogen bond with Arg115 (2.397 Å)

responsible for coordination with heme which is evident from the Figs. 3 and 4 [26, 28]. Considering the least active compounds (like compounds **107, 108, 109, 113, 114**) in the data set, the docking results show that a number of steric bumps with different amino acid residues occur in these cases. In the case of compound **113,** although one hydrogen bond formed with Met374, two steric bumps appear with the same amino acid residue (Fig. 6). Additional bumps have also occurred with amino acids Phe221, Ser478, Ala306, Thr310 and most importantly with the heme, thus resulting in poor inhibitory activity. Another compound in the list, compound **107,** shows poor inhibitory activity. The reason may be due to a number of bumps occurring with Asp309, Thr310, Met374, Arg115, Ser478, Val370 (Fig. 7). The volume of the active cavity of the enzyme is not more than 400 Å^3 [22]. The molecules in the least active range have molecular volume more than 300 Å^3 (322 Å^3 for **113** and 365 Å^3 for compound **107**) leading to

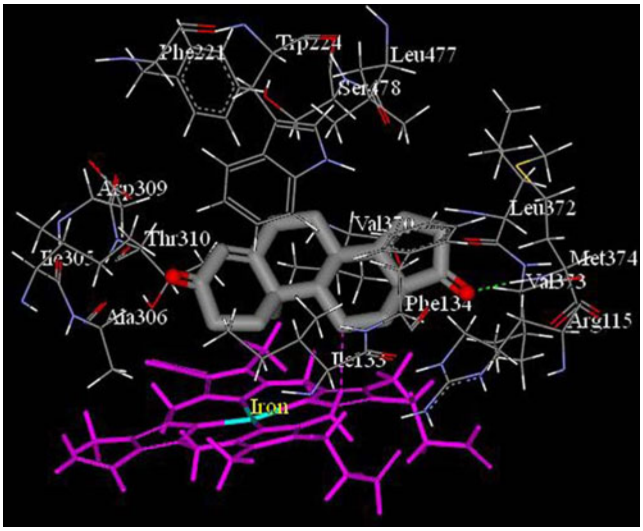


Fig. 2 Bound ligand (androstanedione) docked into the active site human placental aromatase [important interacting amino acids and iron in heme have been labeled]

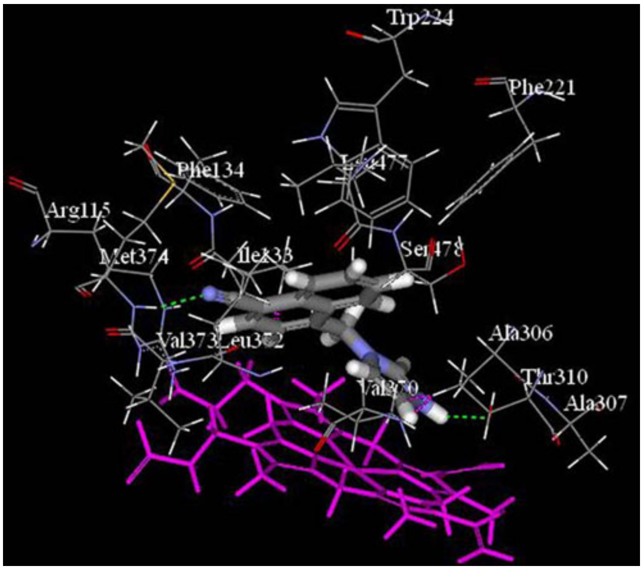


Fig. 4 Docked conformation of compound **59** along with the important amino acid residues of human placental aromatase: Compound **59** forms two hydrogen bonds one between the –CN group of the ligand and the amide back bone of Met374 and the other between the NH fragment of the azole nucleus and the side chain hydroxyl group of Thr310

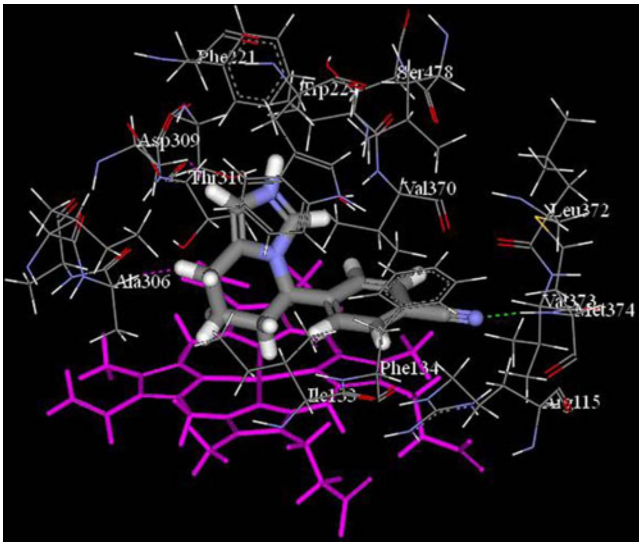


Fig. 5 Docked conformation of compound **116** along with the important amino acid residues of human placental aromatase: Apart from the hydrogen bond with Met374 (using the –CN group of the ligand), there is a steric bump formation with the polar amino acid Asp309

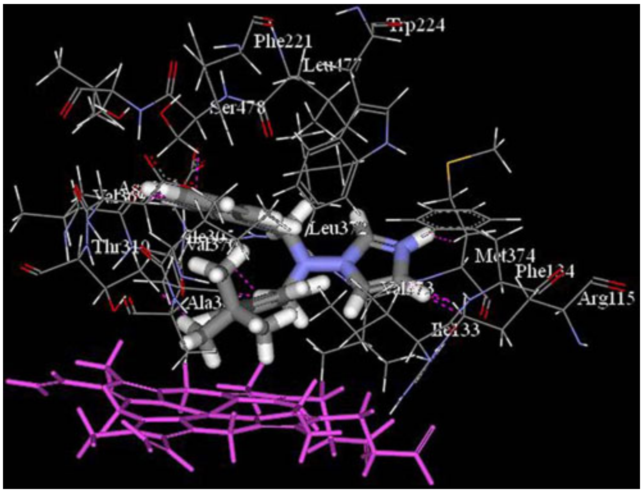


Fig. 7 Docked conformation of compound **107** along with the important amino acid residues of human placental aromatase: A number of bumps occur with Asp309, Thr310, Met374, Arg115, Ser478, Val370

formation of bumps. The ligands are somehow placed in the active cavity but the orientation of the molecules produces unfavorable steric interactions. One of the most important features of a strong inhibitor binding to CYP enzymes is the capability to interact as the ligand with the iron atom of the heme group [28]. From Figs. 3 and 4, it can be observed that the azole ring is in close proximity to the heme moiety. It is reported in the literature that azoles have the capacity to bind with heme iron of cytochromes [60]. This is supported by the results of our docking study.

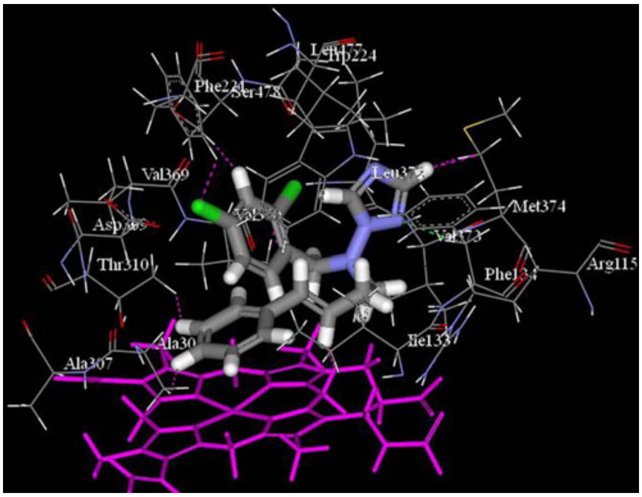


Fig. 6 Docked conformation of compound **113** along with the important amino acid residues of human placental aromatase: although one hydrogen bond has formed with Met374, two steric bumps appear with the same amino acid residue

Molecular shape analysis

The view of the aligned training set molecules is shown in Fig. 8. The following two equations (Eqs. 1 and 2) were among the best ones obtained from the genetic function approximation (5000 iterations) and genetic partial least squares (1000 crossovers, scaled variables, and other

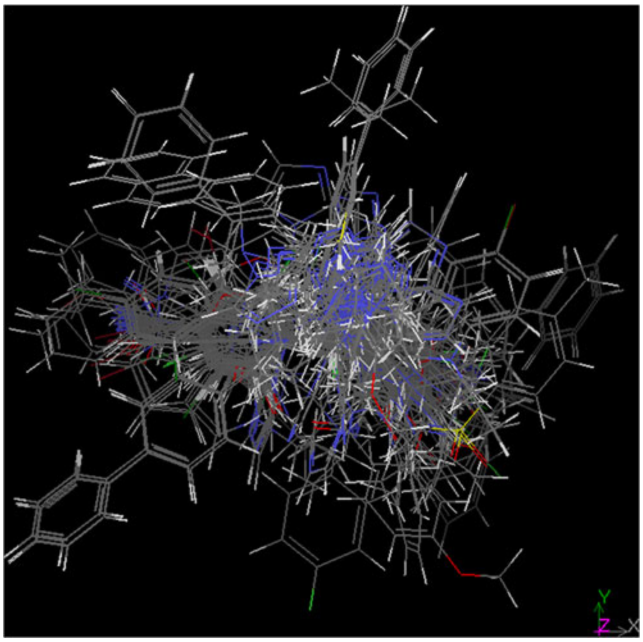


Fig. 8 Aligned geometry of training set molecules

default settings) respectively. Both linear and linear spline terms were used for development of the models.

$$\begin{aligned} pIC_{50} &= 6.856(\pm 0.236) - 53.500(\pm 6.793) < Jurs\_FNSA\_3 \\ &\quad +0.063 > -0.008(\pm 0.001)NCOSV - 0.461(\pm 0.081) \\ &\quad < Hbondacceptor - 2 > -0.472(\pm 0.095) < 4.134 \\ &\quad -A\log P > \\ n_{Training} &= 87, LOF = 0.309, R^2 = 0.713, \\ &\quad R_a^2 = 0.699, F = 50.91(df4, 82), Q^2 = 0.668, r^2_{m(LOO)} \\ &\quad = 0.496, n_{Test} = 29, R^2_{pred} = 0.639, r^2_{m(test)} \\ &\quad = 0.633, r^2_{m(overall)} = 0.510 \end{aligned} \tag{1}$$

The relative importance of the descriptors according to their standardized regression coefficients is in the following order: <*Jurs_FNSA_3*+0.063> >*NCOSV*> <*Hbondacceptor*-2> ><4.134-*AlogP*>.

The standard errors of regression coefficients are given within parentheses. Eq. 1 could explain 69.9% of the variance (adjusted coefficient of variation) while it could predict 66.8% of the variance (leave-one-out predicted variance). The difference between R^2 and Q^2 values is not very high (less than 0.3) [61]. When the equation was used to predict the CYP19 inhibition potency of the test set compounds, the predicted R^2 ($R_{pred}{}^2$) value was found to be 0.639. The $r_m{}^2$ values for the test, training and overall sets were found to be 0.633, 496 and 0.510 respectively.

All the terms in the equation have a negative contribution toward the inhibitory activity. The negative coefficient of the term <*Jurs_FNSA_3*+0.063> indicates that for optimal inhibitory activity the value of *Jurs_FNSA_3* should be more negative than -0.063. *Jurs_ FNSA_3* (functional charged partial negative surface area) is derived from the following equation

$$FNSA\_3 = \frac{PNSA\_3}{SASA},$$

where *PNSA_3* is atomic charge weighted negative surface area. It is the sum of products of atomic solvent accessible surface area and partial charges q_a^- over all negatively charged atoms, *i.e.*, $PNSA\_3 = \sum_{a-} q_a^- . SA_a^-$. SASA is the solvent accessible surface area.

Compounds like **1, 25, 86, 89, 98, 103, 107, 110, 114** show poor inhibitory activities because of less negative values of *Jurs_FNSA_3.* On the other hand compounds **24, 45, 48, 50, 51, 56, 60, 73, 77** having zero value of the term <*Jurs_FNSA_3*+0.063418> show activity in the higher range. Presence of heteroatoms (substituent groups like nitro, cyano) increases the negative value of Jurs_FNSA_3. This is supported by the docking study which shows that, for example, the nitro group of compound **45** and cyano group of compound **116** are involved in hydrogen bond formation with the active site.

The negative coefficient of the term *NCOSV* (non common steric overlap volume) shows its negative contribution. *NCOSV* indicates the non common steric overlap volume of each molecule to the shape reference compound **50**. Compounds with lower values of *NCOSV* (like **44, 45, 47, 48, 55, 64, 65, 79, 80, 82, 116**) show higher inhibitory activity than compounds having higher values of the parameter (**35, 37, 89, 98, 100, 101, 103, 104, 107, 108, 114**).

The term <*Hbondacceptor*-2> with negative regression coefficient indicates that the number of hydrogen bond acceptor groups should be 2 or less than 2 for optimum inhibitory activity. Compounds with more number of hydrogen bond acceptor groups (compounds like **39, 93, 105, 108, 114** containing *three* hydrogen bond acceptor groups, compounds like **40, 71, 94, 111** containing *four* hydrogen bond acceptor groups and compounds like **70** containing *five* hydrogen bond acceptor groups) show poor inhibitory activity. The docking study has indicated that one or two hydrogen bond(s) formed with amino acid Met374 is/are essential for all the highly active molecules and least active molecules as well. However, increase in hydrogen bond acceptor groups may not facilitate the inhibitory activity as other parts of the molecules (not involved in hydrogen bonding interactions) are stabilized by van der Waals interactions (*vide supra*). Figure 9 shows the docked geometry of compound **54** having 6 hydrogen bond acceptor groups. This compound forms two hydrogen bonds and also two steric bumps and the binding pose of this compound is different from that of others.

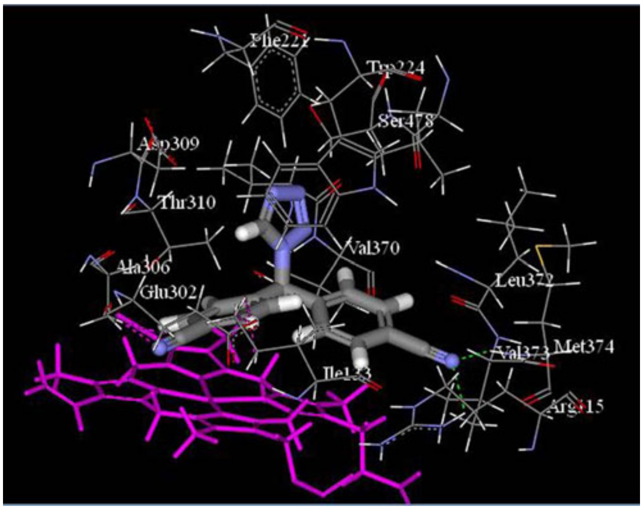


Fig. 9 Docked conformation of compound **54** along with the important amino acid residues of human placental aromatase: **54** forms two hydrogen bonds and also two steric bumps

The negative regression coefficient of the term <4.134-*AlogP*> indicates that the value of log of partition coefficient (AlogP) should be more than 4.134 for optimum inhibitory activity. This is supported by the docking study which suggests that binding of the compounds with the active pocket is stabilized by van der Waals interactions with the non polar amino acids (Ala306, Thr310, Trp224, Val370, Ile133, Phe134, Leu372, Val373).

$$\begin{aligned} pIC_{50} = {} & 5.561 - 0.679 < Hbondacceptor - 2 > -0.084 \\ & < Jurs\_PNSA\_3 + 34.086 > -0.553 < A\log P \\ & - 4.273 > -22.686 < Jurs\_FNSA\_1 - 0.414 \\ & > +0.139 Chiralcenters \\ n_{Training} = {} & 87, LSE = 0.266, R^2 = 0.691, R_a^2 = 0.676, \\ & F = 45.83(df4, 82), Q^2 = 0.630, r^2_{m(LOO)} = 0.605, \\ n_{Test} = {} & 29, R^2_{pred} = 0.630, r^2_{m(test)} = 0.608, \\ r^2_{m(overall)} = {} & 0.606 \end{aligned} \tag{2}$$

The above equation was found to be statistically significant with explained variance of 67.6% and leave-one-out predicted variance of 63.0%. When the equation is applied on the test set compounds the R_{pred}^2 value was found to be 0.630. Statistical significance of the model was also indicated by r_m^2 parameters listed in Table 4. According to the standardized values of the regression coefficients, the relative importance of the variables in the G/PLS equation is in the following order: <*Hbondacceptor*-2> ><*Jurs_PNSA_3* +34.086> > <*AlogP*-4.273> > <*Jurs_FNSA_1*-0.414>> *Chiralcenters.*

The negative coefficient of <*Jurs_PNSA_3* +34.086> indicates that compounds with the values of Jurs_PNSA_3 more negative than -34.086 possess significant inhibitory activity (for example **24, 45, 48, 51, 55, 56, 60)** than compounds with corresponding lower negative values of the parameter (**1, 25, 107**). Presence of heteroatoms (groups like nitro, cyano) increases the negative value of Jurs_PNSA_3. This is supported by the docking study which shows that, for example, the nitro group of compound **45** and cyano group of compound **116** are involved in hydrogen bond formation with the active site.

Jurs_FNSA_1 is the fractional charged partial negative surface area. The *Jurs_FNSA_1* values are obtained by dividing the product of partial negative solvent-accessible surface area and the total negative charge by the total molecular solvent-accessible surface area from the following equation

$$FNSA\_1 = \frac{PNSA_1}{SASA},$$

where PNSA_1 is the sum of the solvent accessible surface areas of all negatively charged atoms ($PNSA_1 = \sum_{a-} SA_a^-$). The negative coefficient of the term <*Jurs_FNSA_1*-0.414> indicates that the value of *Jurs_FNSA_1* should be less than 0.414 for better inhibitory activity (like compounds **24, 45, 47, 52, 77)**. The parameter FNSA_1 balances the term PNSA_3 in Eq. 2 as hydrophobicity and nonpolar surface area are also required for binding (*vide supra*).

The negative regression coefficient of the term <*AlogP*-4.273> indicates that the value of log of partition coefficient (AlogP) should be less than 4.273 for optimum inhibitory activity. As we have seen from the docking studies that the compounds are involved in both hydrogen bonding and van der Waals interactions, there will be a cut off higher limit of favorable hydrophobicity. Too much increase of molecular bulk (and hence hydrophobicity) may lead to unfavorable steric interactions.

The inhibitory activity is favored by increase in number of chiral centers as indicated by its positive regression coefficient. Compounds witha higher number of chiral centers (like **20, 21, 24, 81, 116)** show activity in the moderate range. Compounds without any chiral centers like **1, 86, 89, 94, 107, 108, 110, 114** show poor inhibitory activities. It has been observed that compounds without any chiral centers (**45, 47, 48, 51, 56, 64**) show activity in higher range due to favorable values of the other three parameters (<*Hbondacceptor*-2>, <*Jurs_PNSA_3* +34.086>, <*Jurs_FNSA_1*-0.414>).

Table 4 Statistical comparison of different models[a]

Type of descriptors	Type of statistical analysis	Equation no.	R^2	Q^2	R_{pred}^2	r_m^2 (test)	r_m^2 (LOO)	r_m^2 (overall)
MSA, Spatial, Electronic, Thermodynamic, Structural	GFA	(1)	**0.713**	**0.668**	0.639	0.633	0.496	0.510
	G/PLS	(2)	0.691	0.630	0.630	0.608	**0.605**	**0.606**
Topological, Structural, Thermodynamic	GFA	(3)	0.662	0.602	0.637	0.628	0.444	0.469
2D	GFA	(4)	0.680	0.621	**0.687**	**0.657**	0.454	0.489

[a] The best values of different metrics (see text for details) are shown in bold face.

Modeling with 2D descriptors

Eq. 3 is one of the best ones obtained from the genetic function approximation (5000 iterations). Both linear and linear spline terms were used for development of the models.

$$\begin{aligned} pIC_{50} &= 5.065(\pm 0.415) + 0.065(\pm 0.011) S\_tN \\ &\quad - 0.567(\pm 0.107) < A\log P - 4.701 > \\ &\quad + 0.644(\pm 0.139) Chiralcenters - 0.030(\pm 0.009) SC\_3P \\ &\quad - 0.367(\pm 0.126) < S\_dsCH - 1.553 > \\ n_{Training} &= 87, LOF = 0.374, R^2 = 0.662, R_a^2 = 0.641, \\ F &= 31.68(df\,5,81), Q^2 = 0.602, r^2_{m(LOO)} = 0.444, \\ n_{Test} &= 29, R^2_{pred} = 0.637, r^2_{m(test)} = 0.628, r^2_{m(overall)} = 0.469 \end{aligned} \tag{3}$$

The standard errors of regression coefficients are given within parentheses. The statistical quality of Eq. 3 is listed in Table 4. According to the standardized values of the regression coefficients, the relative importance of the variables is in the following order: *S_tN*> <*AlogP*-4.701> >*Chiralcenters* > *SC_3P* > <*S_dsCH*-1.553>.

The E-state index of fragment ≡*N* (*S_tN*) has positive contribution toward the inhibitory activity. Compounds (for example **47, 48, 50, 51, 56, 59)** with high values of the parameter possess significant inhibitory activity. Compounds having a cyano substituent have non-zero values of this parameter and it was found from the docking study that the cyano group of the compounds may be involved in the favorable hydrogen boning interactions with amino acid residues like Met374.

The negative regression coefficient of the term <*AlogP*-4.701> indicates that the value of log of partition coefficient (*AlogP*) should be less than 4.701 for optimum inhibitory activity. Considering Eqs. 1 and 3, we find that the range of AlogP should be from 4.134 to 4.701. Based on this range of AlogP values, compounds like **51, 52, 54, 55, 60** show good inhibitory activity. Other compounds in this range show poor activity due to absence of the ≡*N* fragment. In the docking study, it was found that binding of different compounds with the active pocket is stabilized by van der Waals interactions with the non polar amino acids (Ala306, Thr310, Trp224, Val370, Ile133, Phe134, Leu372, Val373).

In Eq. 3, number of chiral centers shows a positive contribution as also found in Eq. 2.

The parameter *SC_3P* is the number of third-order sub graphs in the molecular graph: the number of paths of length 3. It depends on the branching of molecules. The negative coefficient of the term indicates compounds with high values of the parameter (like **31, 35, 37, 58)** show activity in the lower range than compounds with low values of the parameter **(45, 64, 116**).

The parameter *S_dsCH* is the E-state index of fragment = CH -. The negative coefficient of the term <*S_dsCH*-1.553> indicates that for optimal inhibitory activity the value of the parameter should be less than 1.553. Almost all the compounds possess a zero value for the term *S_dsCH* except a few compounds. Compounds (like **70, 71, 108, 114**) with values of the parameter more than 1.553 show poor inhibitory activity. Compounds with a zero value for the parameter like **45, 47, 50, 51, 56, 64, 67, 116** show significant inhibitory activities. In this regard, compounds **94** and **107** show poor activity instead of zero value for the parameter due to lack of tertiary nitrogen atom (*S_tN*) and high *SC_3P* and *AlogP* values.

Modeling with combined set of descriptors

Eq. 4 is one of the best equations obtained from the genetic function approximation (5000 iterations) using combined set of descriptors. Both linear and linear spline terms were used for development of the models.

$$\begin{aligned} pIC_{50} &= 3.697(\pm 0.347) - 0.008(\pm 0.001) < Jurs\_TASA \\ &\quad - 494.777 > + 0.053(\pm 0.011) StN \\ &\quad - 12.944(\pm 3.624) < S\_aaaC - 2.520 > \\ &\quad - 0.208(\pm 0.063) Hbondacceptor \\ &\quad + 1.865(\pm 0.576) Fo \qquad (4) \\ n_{Training} &= 87, LOF = 0.354, \\ R^2 &= 0.680, R_a^2 = 0.660, F = 34.39(df\,5,81), \\ Q^2 &= 0.621, r^2_{m(LOO)} = 0.454, n_{Test} = 29, R^2_{pred} = 0.687, \\ r^2_{m(test)} &= 0.657, r^2_{m(overall)} = 0.489 \end{aligned}$$

According to the standardized regression coefficients, the relative importance of the descriptors is in the following order: <*Jurs_TASA*-494.777> >*S_tN*> < *S_aaaC* -2.520> >*Hbondacceptor*> *Fo*.

The negative coefficient of <*Jurs_TASA*-494.777> indicates that value of total hydrophobic surface area (TASA) should be less than 494.777. Jurs_TASA (total hydrophobic surface area) is defined as the sum of solvent accessible surface areas of atoms with absolute value of partial charges less than 0.2, *i.e.*,

$$TASA = \sum\nolimits_a SA_a$$

$$\forall a = |q_a| \langle 0.2$$

Compounds having lower values of this parameter have higher inhibitory activity. The presence of a number of polar groups or fragments upto the required limit in case of compounds like **45, 48, 58, 63, 64, 65, 73, 114** with TASA values less than 494.777 show significant favorable inhibitory activities whereas compounds (for example **103, 107, 108, 110, 114**) with corresponding higher values of the parameter show poor inhibitory activity. As we have already indicated in the docking studies that hydrogen bonding interactions are important apart from van der Waals interactions for this series of compounds, and hence, absence of required number of polar groups (leading to higher values of hydrophobic surface area) leads to poor inhibitory activity.

The E-state index of fragment ≡*N* (*S_tN*) has a positive contribution toward the inhibitory activity and this observation is similar to Eq. 3.

The term < *S_aaaC* -2.520> with negative regression coefficient indicates that the value of the E-state index of fragment (*S_aaaC*) should be less than 2.520. Compounds (**1, 25, 36,107, 108**) with higher values of the corresponding parameter show poor inhibitory activity. Compounds with zero and low values of the parameter like compounds **45, 48, 58, 64, 116** show good inhibitory activity and corresponding <*Jurs_TASA*-494.777> and *S_tN* parameters values for the mentioned compounds are within the favorable range as mentioned earlier.

The term *Hbondacceptor* shows a negative regression coefficient when the parameter S_tN shows a positive regression coefficient and this justifies the negative coefficient of the term <*Hbondacceptor*-2> Eqs. 1 and 2.

Common overlap volume ratio (*Fo*) is the ratio of common overlap steric volume to the volume of individual molecules. The positive coefficient of *Fo* indicates that molecules with similar common overlap steric volume to shape reference compounds will show good inhibitory activity as exemplified by the compounds like **47, 54, 48, 80**. Molecules (**72, 108, 110**) which are very dissimilar to the shape reference compounds show poor activity.

Randomization tests of the developed models

Further validation of the models was carried out using the Y scrambling technique. The process randomization test has been performed at 90% confidence level and the developed models were subjected to randomization test at 99% confidence interval. The Y column was permuted randomly and the average correlation coefficient (R_r) of all the randomized models was calculated. The process randomization is different from model randomization in that the descriptor selection process is repeated from the whole pool of descriptors in the former case while in the latter case only those descriptors present in the model are used. The values of R_r obtained for all the models were significantly lower than the squared correlation coefficient (R) of the non randomized model (Table 5). The metric R_p^2, which penalizes the model R^2 for small differences between R^2 and R_r^2, was calculated for all the

Table 5 Randomization test results for process and models

	Process randomization				Model randomization			
Eq. No.	1	2	3	4	1	2	3	4
Modeling technique	GFA (Spline)	G/PLS (Spline)	GFA (Spline)	GFA (Spline)	GFA (Spline)	G/PLS (Spline)	GFA (Spline)	GFA (Spline)
R from non random model	0.844	0.831	0.814	0.825	0.844	0.831	0.814	0.825
Confidence level	90%	90%	90%	90%	99%	99%	99%	99%
Mean value of R for random trials ± standard deviation	0.361±0.150	0.448±0.057	0.330±0.116	0.340±0.103	0.214±0.062	0.051±0.106	0.229±0.074	0.231±0.068
R_p^2	0.543	0.483	0.493	0.512	0.582	0.573	0.518	0.539

developed models. The results show that for all the equations the values of R_p^2 are above 0.5 or at least near 0.5 (for both process and model randomization tests) and this suggests that Eqs. 1–4 are robust and not obtained by chance.

Overview and conclusions

In order to explore the molecular shape features, properties and appropriate binding mode of aromatase inhibitors in the active site, molecular shape analysis (along with thermodynamic, structural and Jurs parameters and also with topological descriptors) and molecular docking studies were performed on a dataset of 116 structurally diverse compounds. For the QSAR studies, initially the dataset was divided into training (n=87) and test set (n=29) by *k*-means clustering techniques based on standardized topological, structural and thermodynamic descriptor matrix. The docking study indicates that the important interacting amino acids present in the active site are Met374, Arg115, Ile133, Ala306, Thr310, Asp309, Val370 and Ser478. One or more hydrogen bonds formed with Met 374 are one of the essential requirements of the ligands for optimum binding. Besides this, compounds in higher activity range form hydrogen bonds with Arg115 and/or Thr310. The amino acids responsible for hydrophobic interactions are Ala306, Thr310, Trp224, Val370, Ile133, Phe134, Leu372, Val373. There may be unfavorable steric clashes with Asp309, Thr310, Met374, Arg115, Ser478, Val370, Phe221 for compounds having undesirable substitution pattern. The developed QSAR models indicate that optimum number of *Hbondacceptor* groups (less than or equal to 2) is favorable for the binding and this is supported by our docking results. The developed QSAR model indicates the importance of a different shape (*NCOSV, Fo*) Jurs (*Jurs_FNSA_3, Jurs_PNSA_3, Jurs_FNSA_1, Jurs_TASA*) structural (*Hbond acceptors, Chiralcenters, AlogP*), topological branching index (*SC_3P*) and E-state index for different fragments (*S_tN, S_dsCH, S_aaaC*). Equations. (1), (2) and (3) indicate the optimal range of hydrophobicity of molecules. It was observed in the docking study that in compounds like **54, 56, 57, 116,** the –CN group (*S_tN* fragment) forms hydrogen bond with Met 374 and this is supported by the positive contribution of *S_tN* fragment in the QSAR models and this is also corroborated by the published literature [26]. All four reported QSAR models have been subjected to validation using multiple strategies like internal validation, external validation and Y-randomization. The statistical quality in terms of external validation of the model with 2D descriptors is almost comparable with that of the MSA models. However, internal validation results of the model with 2D descriptors are inferior to the MSA models. However, the advantage of 2D descriptors is that these do not require conformational analysis and alignment unlike MSA. For aromatase inhibition, the GFA model (MSA) with spline option (Eq. 1) was found to be the best model based on internal validation (Q^2=0.668) and the best predictive model (external validation) was the GFA model with spline option using combined set of descriptors (Eq. 4; R_{pred}^2=0.687). Based on $r_{m\ (overall)}^2$ criterion, the best model among the four models (Table 4) was the G/PLS model (MSA) with spline option (Eq. 2; $r_{m\ (overall)}^2$=0.606). So, it can be concluded that for ideal aromatase inhibitors, there should be at least one or two hydrogen bond acceptor groups (like $–NO_2$, -CN) and optimal hydrophobicity.

Acknowledgments This work is 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. Cancer facts and figures (2007) American Cancer Society: Atlanta, GA, 2007. http://www.cancer.org/downloads/STT/CAFF2007PWsecured.pdf (accessed on Nov 11, 2009)
2. Labrie F (1991) Intracrinology. Mol Cell Endocrinol 78:C113–C118. doi:10.1016/0303-7207(91)90116-A
3. Cuzick J, Wang DY, Bulbrook RD (1986) The prevention of breast cancer. Lancet 8472:83–86. doi:10.1016/S0140-6736(86)90729-4
4. Clemons M, Goss P (2001) Mechanisms of disease: estrogen and the risk of breast cancer. N Engl J Med 344:276–285. doi:10.1056/NEJM200101253440407
5. Osborne CK, Yochmowitz MG, Knight WA, McGuire WL (1980) The value of estrogen and progesterone receptors in the treatment of breast cancer. Cancer 46:2884–2888. doi:10.1002/1097-0142(19801215)46:12+<2884::AID-CNCR2820461429>3.0.CO;2-U
6. Brueggemeier RW, Hackett JC, Diaz-Cruz ES (2005) Aromatase inhibitors in the treatment of breast cancer. Endocr Rev 26:331–345. doi:10.1210/er.2004-0015
7. Trunet PF, Vreeland F, Royce C, Chaudri HA, Cooper J, Bhatnagar AS (1997) Clinical use of aromatase inhibitors in the treatment of advanced breast cancer. J Steroid Biochem Mol Biol 61:241–245. doi:10.1016/S0960-0760(96)00249-X
8. Brodie AMH, Njar VCO (1998) Aromatase inhibitors in advanced breast cancer: mechanism of action and clinical implications. J Steroid Biochem Mol Biol 66:1–10. doi:10.1016/S0960-0760(98)00022-3
9. Banting L, Nicholls PJ, Shaw MA, Smith HJ (1989) Recent developments in aromatase inhibition as a potential treatment for oestrogen-dependent breast cancer. Prog Med Chem 26:253–298. doi:10.1016/S0079-6468(08)70242-X
10. Banting L (1996) Inhibition of aromatase. Prog Med Chem 33:147–184. doi:10.1016/S0079-6468(08)70305-9
11. O'Reilly JM, Brueggemeier RW (1996) 7alpha-arylaliphatic androsta-1,4-diene-3,17-diones as enzyme-activated irreversible inhibitors of aromatase. J Steroid Biochem Mol Bio l59:93–102. doi:10.1016/S0960-0760(96)00087-8
12. Santen RJ, Samojlik E, Lipton A, Harvey H, Ruby EB, Wells SA, Kendall J (1977) Kinetic, hormonal and clinical studies with

aminoglutethimide in breast cancer. Cancer 39:2948–2958. doi:10.1002/1097-0142(197706)39:6<2948::AID-CNCR2820390681>3.0.CO;2-9

13. Plourde PV, Dyroff M, Dowsett M, Demers L, Yates R, Webster A (1995) ARIMIDEX: a new oral, once-a-day aromatase inhibitor. J Steroid Biochem Mol Biol 53:175–179. doi:10.1016/0960-0760(95)00045-2
14. Lipton A, Demers LM, Harvey HA, Kambic KB, Grossberg H, Brady C et al (1995) Letrozole (CGS 20267). A phase I study of a new potent oral aromatase inhibitor of breast cancer. Cancer 75:2132–2138. doi:10.1002/1097-0142(19950415)75:8<2132:AID-CNCR2820750816>3.0.CO;2-U
15. Evans TR, Di Salle E, Ornati G, Lassus M, Benedetti MS, Pianezzola E et al (1992) Phase I and endocrine study of exemestane (FCE 24304), a new aromatase inhibitor, inpostmenopausal women. Cancer Res 52:5933–5939
16. Goss PE, Ingle JN, Martino S, Robert NJ, Muss HB, Piccart MJ et al (2003) A randomized trial of letrozole in postmenopausal women after five years of tamoxifen therapy for early-stage breast cancer. N Engl J Med 349:1793–1802. doi:10.1056/NEJMoa032312
17. Coombes RC, Hall E, Gibson LJ, Paridaens R, Jassem J, Delozier T et al (2004) A randomized trial of exemestane after two to three years of tamoxifen therapy in postmenopausal women with primary breast cancer. N Engl J Med 350:1081–1092. doi:10.1056/NEJMoa040331
18. Baum M, Budzar AU, Cuzick J, Forbes J, Houghton JH, Klijn JG et al (2002) Anastrozole alone or in combination with tamoxifen versus tamoxifen alone for adjuvant treatment of postmenopausal women with early breast cancer: first results of the ATAC randomised trial. Lancet 359:2131–2139. doi:10.1016/S0140-6736(02)09088-8
19. Nabholtz JM, Buzdar A, Pollak M, Harwin W, Burton G, Mangalik A et al (2000) Anastrozole is superior to tamoxifen as first-line therapy for advanced breast cancer in postmenopausal women: results of a north american multicenter randomized trial. arimidex study group. J Clin Oncol 18:3758–3767
20. Arora A, Potter JF (2004) Aromatase inhibitors: current indications and future prospects for treatment of postmenopausal breast cancer. J Am Geriatr Soc 52:611–616. doi:10.1111/j.1532-5415.2004.52171.x
21. Goss PE (1999) Risks versus benefits in the clinical application of aromatase inhibitors. Endocr Relat Cancer 6:325–332. doi:10.1677/erc.0.0060325
22. Ghosh D, Griswold J, Erman M, Pangborn W (2009) Structural basis for androgen specificity and estrogen synthesis in human aromatase. Nature 457:219–223. doi:10.1038/nature07614
23. Favia AD, Cavalli A, Masetti M, Carotti A, Recanatini M (2006) Three-dimensional model of the human aromataseenzyme and density functional parameterization of the iron-containing protoporphyrin IX for a molecular dynamics study of heme-cysteinato cytochromes. Proteins 62:1074–1087. doi:10.1002/prot.20829
24. Hong Y, Yu B, Sherman M, Yuan YC, Zhou D, Chen S (2007) Molecular basis for the aromatization reaction and exemestane-mediated irreversible inhibition of human aromatase. Mol Endocrinol 21:401–414. doi:10.1210/me.2006-0281
25. Hong Y, Cho M, Yuan Y, Chen S (2008) Molecular basis for the interaction of four different classes of substrates and inhibitors with human aromatase. Biochem Pharmacol 75:1161–1169. doi:10.1016/j.bcp. 2007.11.010
26. Castellano S et al (2008) CYP19 (aromatase): Exploring the scaffold flexibility for novel selective inhibitors. Bioorg Med Chem 16:8349–8358. doi:10.1016/j.bmc.2008.08.046
27. Karkola S, Wähälä K (2009) The binding of lignans, flavonoids and coumestrol to CYP450 aromatase: A molecular modelling study. Mol Cell Endocrinol 301:235–244. doi:10.1016/j.mce.2008.10.003
28. Cole PA, Robinson CH (1990) Mechanism and Inhibition of Cytochrome P-450 Aromatase. J Med Chem 33:2933–2942. doi:10.1021/jm00173a001
29. Le Borgne M, Marchand P, Duflos M, Delevoye-Seiller B, Piessard-Robert S, Le Baut G, Hartmann RW, Palzer M (1997) Synthesis and in vitro evaluation of 3-(1-azolylmethy1)-1H-indoles and 3-(1-azolyl-l-phenylmethyl)-1H-indoles as inhibitors of P450 arom. Arch Pharm 330:141–145. doi:10.1002/ardp. 19973300506
30. Marchand P, Le Borgne M, Palzer M, Le Baut G, Hartmann RW (2003) Preparation and pharmacological profile of 7-(α-Azolylbenzyl)-1H-indoles and indolines as new aromatase inhibitors. Bioorg Med Chem Lett 13:1553–1555. doi:10.1016/S0960-894X(03)00182-3
31. Le Borgne M, Marchand P, Delevoye-Seiller B, Robert JM, Le Baut G, Hartmann RW, Palzer M (1999) New selective nonsteroidal aromatase inhibitors: synthesis and inhibitory activity of 2, 3 or 5-(α-azolylbenzyl)-1H-indoles. Bioorg Med Chem Lett 9:333–336. doi:10.1016/S0960-894X(98)00737-9
32. Hartmann RW, Palusczak A, Lacan F, Ricci G, Ruzziconi R (2004) CYP 17 and CYP 19 Inhibitors. Evaluation of fluorine effects on the inhibiting activity of regioselectively fluorinated 1-(Naphthalen-2-ylmethyl) imidazoles. J Enzyme Inhib Med Chem 19:145–155. doi:10.1080/147563604200196222
33. Sonnet P, Guillon J, Enguehard C, Dallemagne P, Bureau R, Rault S, Auvray P, Moslemi S, Sourdaine P, Galopin S, Séralini GE (1998) Design and synthesis of a new type of non steroidal human aromatase inhibitors. Bioorg Med Chem Lett 8:1041–1044. doi:10.1016/S0960-894X(98)00157-7
34. Recanatini M, Bisi A, Cavalli A, Belluti F, Gobbi S, Rampa A, Valenti P, Palzer M, Palusczak A, Hartmann RW (2001) A new class of nonsteroidal aromatase inhibitors: design and synthesis of chromone and xanthone derivatives and inhibition of the P450 enzymes aromatase and 17r-Hydroxylase/C17, 20-Lyase. J Med Chem 44:672–680. doi:10.1021/jm000955s
35. Cavalli A, Bisi A, Bertucci C, Rosini C, Paluszcak A, Gobbi S, Giorgio E, Rampa A, Belluti F, Piazzi L, Valenti P, Hartmann RW, Recanatini M (2005) Enantioselective nonsteroidal aromatase inhibitors identified through a multidisciplinary medicinal chemistry approach. J Med Chem 48:7282–7289. doi:10.1021/jm058042r
36. Leze MP, Le Borgne M, Pinson P, Palusczak A, Duflos M, Le Baut G, Hartmann RW (2006) Synthesis and biological evaluation of 5-[(aryl)(1H-imidazol-1-yl)methyl]-1H-indoles: Potent and selective aromatase inhibitors. Bioorg Med Chem Lett 16:1134–1137. doi:10.1016/j.bmcl.2005.11.099
37. Setzu MG, Stefancich G, Colla PL, Castellano S (2002) Synthesis and antifungal properties of N-[(1, 1?-biphenyl)-4-ylmethyl]-1H-imidazol-1-amine derivatives. Il Farmaco 57:1015–1018. doi:10.1016/S0014-827X(02)01294-6
38. Castellano S, Stefancich G, Chillotti A, Poni G (2003) Synthesis and antimicrobial properties of 3-aryl-1-(1, 1?-biphenyl-4-yl)-2-(1H-imidazol-1-yl)propanes as ‘carba-analogues’ of the Narylmethyl-N-[(1, 1?-biphenyl)-4-ylmethyl])-1H-imidazol-1-amines, a new class of antifungal agents. Il Farmaco 58:563–568. doi:10.1016/S0014-827X(03)00094-6
39. Castellano S, Colla PL, Musiu C, Stefancich G (2000) Azole antifungal agents related to naftifine and butenafine. Arch Pharm 333:162–166. doi:10.1002/1521-4184(20006)333:6<162::AID-ARDP162>3.0.CO;2-S
40. Castellano S, Stefancich G, Musiu C, Colla PL (2000) A new class of antifungal agents. Synthesis and antimycotic activity of disubstituted *N*-azolylamines. Archiv der Pharmazie 333:299–304. doi:10.1002/1521-4184(20009)333:9<299::AID-ARDP299>3.0.CO;2-F
41. Discovery Studio 2.1 is a product of Accelrys Inc, San Diego, CA, USA

42. Cerius2 Version 4.10 is a product of Accelrys Inc, San Diego, USA. http://www.accelrys.com/cerius2
43. Leonard JT, Roy K (2006) On selection of training and test sets for the development of predictive QSAR models. QSAR Comb Sci 25:235–251. doi:10.1002/qsar.200510161
44. Roy K, Mandal AS (2008) Development of linear and nonlinear predictive QSAR models and their external validation using molecular similarity principle for anti-HIV indolyl aryl sulfones. J Enz Inh Med Chem 23:980–995. doi:10.1080/14756360701811379
45. Hopfinger AJ, Tokarsi JS (1997) Three-dimensional Quantitative structure acticity relationship analysis. In: Charifson PS (ed) Practical Applications of Computer-Aided Drug Design. Dekker, New York, pp 105–164
46. Fan Y, Shi LM, Kohn KW, Pommier Y, Weinstein JN (2001) Quantitative structure-antitumor activity relationships of camptothecinanalogues: cluster analysis and genetic algorithm-based studies. J Med Chem 44:3254–3263. doi:10.1021/jm0005151
47. Rogers D, Hopfinger AJ (1994) Application of genetic function approximation to quantitative structure - activity relationship and quantitative structure - property relationship. J Chem Inf Comput Sci 34:854–866. doi:10.1021/ci00020a020
48. Dunn WJ III, Rogers D (1996) Genetic partial least squares in QSAR. In: Devillers J (ed) Genetic algorithms in molecular modeling. Academic, London, pp 109–130
49. Hasegawa K, Miyashita Y, Funatsu K (1997) GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists. J Chem Inf Comput Sci 37:306–310. doi:10.1021/ci960047x
50. Snedecor GW, Cochran WG (1967) Statistical methods. Oxford & IBH, New Delhi
51. Wold S (1995) PLS for Multivariate Linear Modeling. In: van de Waterbeemd H (ed) Chemometric methods in molecular design. VCH, Weinheim, pp 195–218
52. Debnath AK (2001) In: Ghose AK, Viswanadhan VN (eds) Combinatorial library design and evaluation. Dekker, New York, pp 73–129
53. Roy K (2007) On Some aspects of validation of predictive QSAR models. Expert Opin Drug Discov 2:1567–1577. doi:10.1517/17460441.2.12.1567
54. Roy PP, Roy K (2008) On some aspects of variable selection for partial least squares regression models. QSAR Comb Sci 27:302–313. doi:10.1002/qsar.200710043
55. Roy K, Roy PP (2008) Comparative QSAR studies of CYP1A2 inhibitor flavonoids using 2D and 3D descriptors. Chem Biol Drug Des 72:370–382. doi:10.1111/j.1747-0285.2008.00717.x
56. Roy PP, Paul S, Mitra I, Roy K (2009) On two novel parameters for validation of predictive QSAR models. Molecules 14:1660–1701. doi:10.3390/molecules14051660
57. Mitra I, Roy PP, Kar S, Ojha P, Roy K (2010) On further application of r_m^2 as a metric for validation of QSAR models. J Chemometrics 24:22–33. doi:10.1002/cem.1268
58. Roy PP, Leonard JT, Roy K (2008) Exploring the impact of the size of training sets for the development of predictive QSAR models. Chemom Intell Lab Sys 90:31–42. doi:10.1016/j.chemolab.2007.07.004
59. Murthy JN, Nagaraju M, Sastry GM, Rao AR, Sastry GN (2006) Active site acidic residues and structural analysis of modelled human aromatase: a potential drug target for breast cancer. J Comput Aided Mol Des 19:857–870. doi:10.1007/s10822-005-9024-0
60. Vanden Bossche H, Koymans L (1998) Cytochromes P450 in fungi. Mycoses 41:32–38. doi:10.1111/j.1439-0507.1998.tb00581.x
61. Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect 111:1361–1375. doi:10.1289/ehp. 5758