J Mol Model (2012) 18:2885–2895
DOI 10.1007/s00894-011-1304-0

ORIGINAL PAPER

Receptor-guided 3D-QSAR approach for the discovery of c-kit tyrosine kinase inhibitors

Anna Maria Almerico · Marco Tutone · Antonino Lauria

Received: 28 June 2011 / Accepted: 7 November 2011 / Published online: 30 November 2011

Abstract Studies of the the three-dimensional quantitative structure–activity relationships for ninety-five c-kit tyrosine kinase inhibitors were performed. Based on a co-crystallized compound (1 T46), known inhibitors were aligned with c-kit by induced-fit docking, and multiple training/test set splitting was performed to validate the selected pharmacophore model. The best pharmacophore model consisted of five features: one hydrogen-bond donor and four aromatic rings. Reliable statistics were obtained (R^2=0.95, $R_{pred}{}^2$=0.75), and the model was validated by using it to select c-kit inhibitors from a database; 82.1% of the hits it retrieved were active. Accordingly, our model can be reliably used to identify new c-kit inhibitors, and can provide useful information when designing new inhibitors.

Keywords C-kit · 3D-QSAR · Kohonen maps · Induced-fit docking

Introduction

The *c-kit* proto-oncogene encodes a transmembrane tyrosine kinase receptor that is activated by the stem cell factor (SCF), its natural ligand. C-kit protein plays a critical role in modulating histamine release from mast cells [1, 2] following its binding with SCF, which leads to dimerization and autophosphorylation at specific tyrosine residues. Moreover, signaling by c-kit plays an important role in cellular transformation and differentiation, including proliferation, survival, adhesion, and chemotaxis [3]. The overexpression of the *c-kit* proto-oncogene has been reported in hematopoietic cells, small cell lung cancer, and gastrointestinal stromal tumors [4–6]. Furthermore, it has been demonstrated that mutations of *c-kit* protect human colon adenocarcinoma cells against apoptosis and enhance their invasive potential [7]. The clinical importance of *c-kit* expression in tumors has led to research focused on finding inhibitors of this tyrosine kinase. Imatinib (Gleevec®) was the first such compound to be used in therapy, but mutations of *c-kit* led to reduced efficacy or a complete lack of efficacy of this treatment. Other compounds are likely to be effective against mutants, such as sunitinib (Sutent®), but the need for new and more effective inhibitors is still critical. In this paper, we report a three-dimensional quantitative structure–activity relationship (3D-QSAR) analysis of 95 known c-kit inhibitors that were initially docked into the crystal structure of c-kit by means of a mixed approach including molecular dynamics and docking (induced fit). The model obtained revealed interesting features that should be considered during the design and development of new potentially active candidates targeting this kinase, which could be useful as anticancer agents.

Materials and methods

Dataset for analysis

A dataset of 95 compounds (2-aminobenzoxazole derivatives [8], 3-aminobenzoisoxa(thi)azole derivatives [9, 10], thienopyrimidine derivatives [11], and anilinophthalazine

Electronic supplementary material The online version of this article (doi:10.1007/s00894-011-1304-0) contains supplementary material, which is available to authorized users.

A. M. Almerico (✉) · M. Tutone · A. Lauria
Dipartimento di Scienze e Tecnologie Molecolari e Biomolecolari (STEMBIO), Sezione di Chimica Farmaceutica e Biologica,
Università di Palermo,
Via Archirafi, 32,
90123 Palermo, Italy
e-mail: annamaria.almerico@unipa.it

derivatives [12]) was selected for the present study (Table 1). Their pIC_{50} values were used as a dependent variable in the QSAR models. In order to validate each model, the inhibitors were split into a training set and a test set. Two different splitting methods were applied. The first method involved automated random selection, while the second involved a Kohonen map artificial neural network (ANN) or self-organizing maps (SOM) [13, 14]. Due to their clustering capabilities, Kohonen maps ensure that both sets are homogeneously distributed within the entire area of descriptor space. The test compounds were selected by evaluating the minimum distance from the centroid of each cell in the top map. Selection in this manner allows predictions to be made by interpolation and not extrapolation from the domain of the particular QSAR model [15]. Descriptor calculations were carried out using the CODESSA software package [16], and descriptor space was explored using Kohonen maps for autoscaled data. This structural information was used to build a Kohonen map (five per five neurons, 300 epochs) [17]. After 300 epochs of net training, similar compounds are clustered together in the multi-dimensional descriptor space. In each splitting method, 76 of the 95 inhibitors were chosen for a training set and 19 were selected as a test set (Table 1).

Ligand preparation

The 95 ligands were processed with the LigPrep software package in order to assign the appropriate protonation states to them at physiological pH(7.2±0.2), employing the Ionizer option. Conformers were generated through MacroModel torsional sampling using the OPLS_2005 force field [18].

Induced-fit docking

Generated conformers were docked into the c-kit crystallographic structure (PDB: 1 T46), which was originally complexed with imatinib, and the best score poses for each ligand were used to generate pharmacophore hypotheses. The mixed molecular docking/dynamics protocol called induced-fit docking (IFD) [19] was used. In an interative manner, this approach combines ligand-docking techniques with those used to model receptor conformational changes. The Glide docking software package [20] was used for ligand flexibility, while the refinement module in the Prime program [21] was used to account for receptor flexibility: the degrees of freedom of side chains were mainly sampled, while minor backbone movements were allowed through minimization. The strategy used was to first dock ligands into a rigid receptor using a softened energy function such that steric clashes do not prevent at least one pose from assuming a conformation close to the correct one (the "ligand sampling step"). The degrees of freedom of the receptor were then sampled, and global ligand/receptor energy minimization was performed for many ligand poses; this attempted to identify low free-energy conformations of the whole complex (the "protein sampling step"). After that, a second ligand docking step was performed on the refined protein structures, using a hard potential function to sample the ligand's conformational space within the refined protein environment (the "ligand resampling step"). Finally, a composite score function was applied to rank the complexes; this accounted for the receptor/ligand interaction energy as well as strain and solvation energies (the "scoring step"). The composite score, which was used to perform the final ranking of the compounds, was derived as follows:

$$\mathrm{IFScore} = \mathrm{GlideScore} + 0.05\mathrm{PrimeEnergy}. \quad (1)$$

3D-QSAR pharmacophore modeling

The 3D-QSAR study was performed using the Phase software package [22]. Phase utilizes fine-grained conformational sampling and a range of scoring techniques to identify common pharmacophore hypotheses. These convey characteristics of 3D chemical structures that are reported to be critical for binding. The pharmacophore model was developed by using a set of pharmacophore features to generate sites for all of the compounds. Each structure was represented by a set of points in 3D space that coincided with various chemical features which facilitated noncovalent binding between the ligand and its binding pocket. Phase provides a standard set of six pharmacophore features: hydrogen-bond acceptor (A), hydrogen-bond donor (D), hydrophobic group (H), negatively ionizable (N), positively ionizable (P), and aromatic ring (R). Hypotheses were generated by systematically varying the number of sites (n_{sites}) and the number of matching active compounds (n_{act}). With $n_{act}=n_{act\_tot}$ initially (n_{act_tot} is the total number of active compounds in the training set), n_{sites} was decreased from its initial value of 7 until at least one hypothesis was found and scored successfully. In this study, with $n_{sites}=5$ and $n_{act}=n_{act\_tot}=8$, common pharmacophores were examined using a scoring protocol to identify the pharmacophore from each surviving n-dimensional box that yielded the best alignment of the active-set ligands. The scoring protocol allows the different hypotheses to be ranked so that the most appropriate can be chosen for further investigation. Inactive molecules were also scored, in order to observe the alignment of these molecules with the pharmacophore hypotheses and to select the best ones. The larger the difference between the scores of the actives and inactives, the better the hypothesis is at discriminating

Table 1 Actual and predicted pIC_{50} values of compounds included in the study

	Random selection		Kohonen map			Random	Kohonen
Compound	Actual pIC_{50}	Pred. pIC_{50}	Actual pIC_{50}	Pred. pIC_{50}	Pharm set	QSAR set	QSAR set
1	5.602	5.59	5.602	5.63	Inactive	Test	Training
2	6.000	5.83	6.000	5.76	Inactive	Training	Test
3	5.854	5.76	5.854	5.39	Inactive	Training	Training
4	5.523	5.58	5.523	5.70	Inactive	Training	Test
5	5.102	5.12	5.102	5.09	Inactive	Training	Training
6	6.155	5.79	6.155	5.90		Test	Training
7	5.456	5.56	5.456	5.45	Inactive	Training	Training
8	5.699	5.83	5.699	5.86	Inactive	Test	Training
9	5.569	5.63	5.569	5.81	Inactive	Training	Training
10	7.495	7.65	7.495	7.20		Training	Training
11	8.046	7.92	8.046	8.04	Active	Training	Training
12	7.553	7.58	7.553	7.47		Training	Test
13	8.222	7.95	8.222	8.00	Active	Training	Training
14	7.167	7.32	7.167	7.10		Training	Training
15	7.081	7.08	7.081	7.28		Training	Test
16	7.301	7.17	7.301	7.54		Test	Training
17	7.721	7.07	7.721	7.00		Test	Test
18	7.824	7.73	7.824	7.54		Training	Training
19	7.569	7.12	7.569	7.17		Test	Test
20	7.481	7.64	7.481	7.62		Test	Training
21	6.959	6.67	6.959	6.90		Training	Training
22	7.658	6.80	7.658	7.72		Test	Training
23	7.538	7.55	7.538	7.46		Training	Training
24	7.301	7.51	7.301	7.91		Training	Test
25	7.319	7.35	7.319	7.25		Training	Training
26	6.854	6.88	6.854	7.16		Training	Training
27	7.367	7.37	7.367	7.39		Training	Training
28	7.721	7.80	7.721	7.24		Training	Training
29	6.757	6.62	6.757	6.76		Training	Training
30	6.740	6.71	6.740	6.86		Training	Training
31	6.796	6.46	6.796	6.98		Test	Training
32	6.398	6.36	6.398	6.53		Training	Training
33	6.032	6.06	6.032	6.14		Training	Training
34	6.237	6.28	6.237	6.14		Training	Training
35	6.237	6.23	6.237	5.80		Training	Training
36	6.097	6.22	6.097	6.12		Training	Training
37	7.086	7.28	7.086	7.29		Training	Training
38	6.102	6.06	6.102	5.97		Training	Training
39	6.018	5.97	6.018	5.90		Training	Training
40	7.432	7.38	7.432	7.24		Training	Training
41	6.444	6.33	6.444	6.26		Training	Training
42	7.585	7.67	7.585	7.33		Training	Training
43	6.092	6.03	6.092	6.79		Training	Test
44	7.569	7.58	7.569	7.78		Training	Training
45	6.699	7.19	6.699	6.87		Test	Training
46	6.745	6.92	6.745	7.04		Training	Training
47	7.420	7.45	7.420	7.58		Training	Training
48	7.268	7.25	7.268	7.51		Training	Test

Table 1 (continued)

	Random selection		Kohonen map			Random	Kohonen
Compound	Actual pIC_{50}	Pred. pIC_{50}	Actual pIC_{50}	Pred. pIC_{50}	Pharm set	QSAR set	QSAR set
49	7.553	7.55	7.553	7.79		Training	Training
50	7.167	7.22	7.167	6.98		Training	Test
51	7.523	7.46	7.523	7.47		Training	Training
52	7.523	7.47	7.523	6.76		Training	Test
53	4.830	4.74	4.830	5.37	Inactive	Training	Test
54	5.352	5.98	5.352	5.32	Inactive	Test	Training
55	7.638	7.59	7.638	7.32		Training	Test
56	6.943	7.02	6.943	7.07		Training	Training
57	6.102	6.10	6.102	6.00		Training	Training
58	5.580	5.28	5.580	5.55	Inactive	Training	Training
59	7.796	7.44	7.796	8.07		Training	Test
60	7.796	7.33	7.796	8.04		Test	Training
61	7.745	8.04	7.745	7.85		Training	Test
62	7.678	7.86	7.678	7.65		Training	Training
63	6.780	6.80	6.780	6.77		Training	Training
64	4.630	4.98	4.630	5.43	Inactive	Training	Training
65	7.921	7.73	7.921	7.73		Training	Training
66	8.398	8.45	8.398	8.08	Active	Training	Training
67	7.854	7.88	7.854	7.56		Training	Training
68	7.959	7.75	7.959	7.96		Test	Training
69	8.155	7.97	8.155	7.77	Active	Training	Training
70	8.000	7.84	8.000	7.75	Active	Training	Training
71	8.301	8.08	8.301	8.09	Active	Training	Training
72	7.569	7.75	7.569	7.76		Training	Training
73	8.301	8.28	8.301	7.75	Active	Training	Test
74	8.155	7.90	8.155	7.96	Active	Test	Training
75	6.991	7.08	6.991	7.08		Training	Training
76	7.432	6.85	7.432	7.09		Test	Training
77	6.470	6.79	6.470	6.68		Test	Training
78	7.921	7.80	7.921	7.70		Training	Training
79	7.208	7.42	7.208	7.37		Training	Training
80	7.444	7.57	7.444	7.59		Training	Training
81	7.854	7.93	7.854	7.74		Training	Test
82	7.620	7.63	7.620	7.64		Training	Training
83	7.301	7.36	7.301	7.63		Training	Training
84	7.367	7.22	7.367	7.32		Training	Training
85	7.252	7.08	7.252	6.86		Test	Training
86	7.143	7.28	7.143	7.33		Training	Training
87	7.638	7.66	7.638	7.63		Training	Test
88	7.194	7.10	7.194	7.35		Training	Training
89	6.666	6.95	6.666	6.95		Training	Training
90	6.532	6.49	6.532	7.27		Training	Test
91	6.943	6.89	6.943	6.93		Training	Training
92	6.943	6.91	6.943	7.00		Test	Training
93	7.149	7.25	7.149	7.34		Training	Training
94	7.149	6.70	7.149	7.32		Test	Training
95	8.097	7.89	8.097	8.21	Active	Training	Training

active from inactive molecules. For QSAR development, models of the pharmacophore features of training-set molecules were placed into a regular grid of cubes, with each cube allotted zero or more "bits" to account for the different types of pharmacophore features in the training-set molecule that occupy the cube (1 Å). This representation gave rise to binary-valued occupation patterns that could be used as independent variables to create partial least-squares (PLS) factor 3D-QSAR models. Statistics on the correlation of the predicted with the actual activity data were collated for the hypothesis. Phase QSAR models can be either atom-based or pharmacophore-based, the difference being whether all atoms are taken into account, or whether only the pharmacophore sites that can be matched to the hypothesis are considered. The choice of model depends largely on whether or not the training-set molecules are sufficiently rigid and congeneric. If the structures contain a small number of rotatable bonds and have some of their structural framework in common, then an atom-based model may work quite well [22]. The selected dataset did not have many rotatable bonds, so the atom-based QSAR model was used.

Results and discussion

We started our work by redocking imatinib into the protein via the IFD approach in order to evaluate the reliability of the computational algorithm before carrying out receptor-guided alignment of the inhibitors. Training alignment of imatinib gave good results in terms of both IF score (crystallized score=−12.28; docked=−12.51), descriptor contributions (crystallized: Hbond=−1.5, vdW=−66.1, Coul=−9.1; docked: Hbond=−1.4, vdW=−66.8, Coul=−9.1), and root mean square deviation (RMSD=0.43) (Fig. 1).

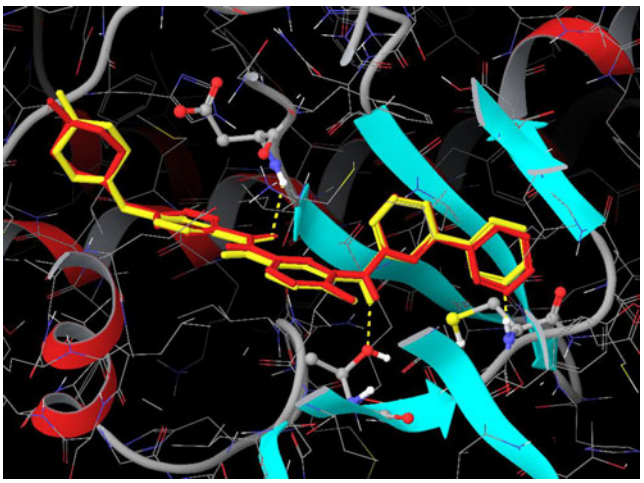

Fig. 1 Superposition of co-crystallized imatinib (*yellow*) and IFD (*red*)

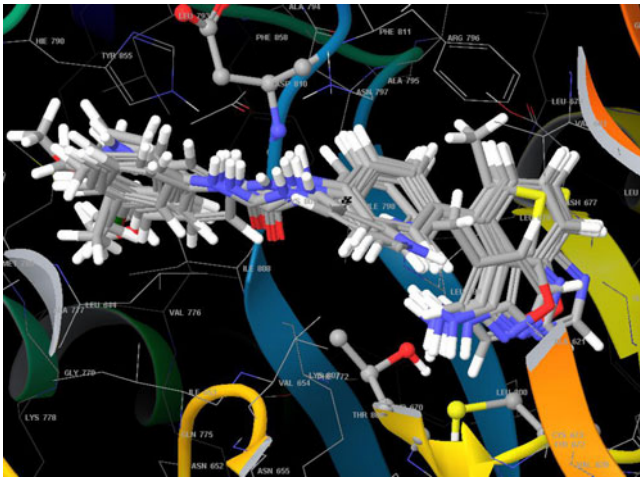

Fig. 2 Docked structures of the most active compounds

Thus, it was possible to align the selected inhibitors using the same approach. The top-ranked binding poses, selected by IFD score and by visual inspection, were used to generate the pharmacophore hypotheses (Fig. 2). The 3D-QSAR studies were carried out using the Phase package from Schrödinger LLC. To find the common pharmacophore hypothesis, the dataset was split into an active set and an inactive set (the "pharm" set) (Table 1). Compounds with $pIC_{50}>8.00$ were considered to be active, those with $pIC_{50}<6.00$ were considered to be inactive, whereas those in-between were considered to be moderately active. Three hundred sixty-one hypotheses were identified (Table 2). Generated hypotheses were submitted to a Phase scoring procedure consisting of three scores: *survival* was calculated solely on the basis of the active set; *surv-inactive* was calculated on the basis of the active and inactive sets; *post-hoc* was calculated on the

Table 2 Identified pharmacophore hypotheses

Variant	Max no. of hypotheses
ARRRR	5
ADHRR	7
ADDRR	97
DDRRR	45
AADRR	58
DRRRR	21
AADHR	4
AADDR	49
AARRR	5
ADRRR	70
	361

A H-acceptor, *D* H-donor, *H* hydrophobic, *R* aromatic ring

Table 3 Best pharmocophore hypotheses

Hypothesis	Survival	Surv-inact	Post-hoc	Site	Vector	Volume	Selectivity
DDRRR.174	3.690	2.161	6.180	0.90	0.994	0.795	1.852
DDRRR.178	3.690	2.161	6.180	0.90	0.994	0.795	1.852
DRRRR.240	3.686	2.177	6.177	0.90	0.986	0.805	1.981
DRRRR.241	3.686	2.177	6.177	0.90	0.986	0.805	1.981
AADRR.48	3.641	2.635	6.161	0.86	0.994	0.788	1.582
ADDRR.182	3.752	2.247	6.152	0.92	0.995	0.838	1.662
DDRRR.175	3.742	2.187	6.142	0.91	0.996	0.837	1.852
DDRRR.179	3.742	2.187	6.142	0.91	0.996	0.837	1.852
ADRRR.252	3.647	2.241	6.138	0.88	0.973	0.793	1.844
DRRRR.113	3.643	2.004	6.134	0.89	0.972	0.786	1.963
DRRRR.173	3.643	2.004	6.134	0.89	0.972	0.786	1.963
DDRRR.13	3.607	2.078	6.127	0.89	0.991	0.731	1.828
DDRRR.23	3.607	2.078	6.127	0.89	0.991	0.731	1.828
ADRRR.262	3.726	2.310	6.126	0.90	0.990	0.838	1.884
AADRR.1	3.605	2.457	6.124	0.85	0.994	0.758	1.460
ADRRR.59	3.604	2.251	6.123	0.85	0.989	0.761	1.642
ADRRR.138	3.633	2.379	6.123	0.88	0.987	0.767	1.673
ADRRR.139	3.633	2.379	6.123	0.88	0.987	0.767	1.673
ADRRR.251	3.672	2.123	6.118	0.88	0.982	0.810	1.846
DRRRR.242	3.716	2.038	6.116	0.89	0.989	0.837	1.985
DRRRR.243	3.716	2.038	6.116	0.89	0.989	0.837	1.985
ADDRR.209	3.715	2.376	6.115	0.89	0.989	0.836	1.724
ADDRR.219	3.713	2.151	6.113	0.88	0.991	0.840	1.771
AADRR.186	3.712	2.601	6.112	0.90	0.995	0.822	1.591
ADDRR.77	3.712	2.480	6.112	0.89	0.995	0.829	1.605
ARRRR.13	3.591	2.197	6.111	0.81	0.985	0.801	1.800
DDRRR.211	3.711	2.107	6.111	0.88	0.990	0.837	1.955
AADRR.62	3.710	2.696	6.110	0.90	0.995	0.817	1.473
AADRR.175	3.663	2.583	6.110	0.88	0.99	0.797	1.544
ADRRR.160	3.710	2.535	6.110	0.88	0.989	0.841	1.745
AADDR.140	3.706	2.568	6.106	0.88	0.997	0.827	1.420
DDRRR.201	3.706	2.059	6.106	0.88	0.990	0.839	1.958
AADDR.134	3.703	2.505	6.103	0.89	0.995	0.814	1.391
ADDRR.183	3.635	1.972	6.102	0.87	0.991	0.774	1.661
AADDR.124	3.652	2.559	6.099	0.87	0.992	0.795	1.387
DRRRR.22	3.699	2.123	6.099	0.87	0.990	0.840	1.957
DRRRR.25	3.699	2.123	6.099	0.87	0.990	0.840	1.957
ADDRR.189	3.652	2.275	6.098	0.87	0.982	0.800	1.700
ADDRR.75	3.696	2.379	6.097	0.88	0.996	0.817	1.605
AADRR.2	3.650	2.664	6.096	0.88	0.992	0.779	1.466
ADDRR.81	3.696	2.150	6.095	0.87	0.998	0.823	1.621
ADDRR.84	3.696	2.150	6.095	0.87	0.998	0.823	1.621
ADRRR.10	3.695	2.243	6.095	0.89	0.996	0.811	1.666
DDRRR.14	3.691	2.139	6.090	0.90	0.991	0.795	1.829
DDRRR.24	3.691	2.139	6.090	0.90	0.991	0.795	1.829
AADDR.128	3.643	2.438	6.089	0.88	0.989	0.774	1.357
DRRRR.115	3.686	1.960	6.086	0.88	0.983	0.826	1.968
DRRRR.175	3.686	1.960	6.086	0.88	0.983	0.826	1.968
AADRR.196	3.682	2.522	6.082	0.86	0.990	0.834	1.606

Table 3 (continued)

Hypothesis	Survival	Surv-inact	Post-hoc	Site	Vector	Volume	Selectivity
AADRR.207	3.682	2.334	6.082	0.85	0.992	0.838	1.671
AARRR.41	3.680	2.384	6.080	0.85	0.990	0.842	1.705
AADRR.176	3.673	2.586	6.073	0.87	0.989	0.811	1.545
AADDR.154	3.669	2.292	6.069	0.85	0.991	0.830	1.440
ADDRR.8	3.666	2.414	6.066	0.86	0.990	0.814	1.591
AADDR.125	3.665	2.570	6.065	0.86	0.991	0.818	1.388
ARRRR.14	3.664	2.140	6.064	0.84	0.989	0.836	1.810
ARRRR.15	3.664	2.140	6.064	0.84	0.989	0.836	1.810
AADDR.129	3.657	2.444	6.057	0.88	0.989	0.791	1.358
AADDR.32	3.609	2.537	6.056	0.84	0.994	0.779	1.412
ADRRR.20	3.656	2.134	6.056	0.88	0.983	0.797	1.712
ADRRR.100	3.653	2.290	6.053	0.87	0.988	0.791	1.646
AARRR.8	3.605	2.404	6.051	0.81	0.983	0.810	1.698
AADDR.30	3.595	2.347	6.041	0.84	0.990	0.767	1.383
AADRR.187	3.641	2.476	6.041	0.85	0.987	0.803	1.610
AADDR.136	3.640	2.373	6.040	0.85	0.988	0.803	1.438
ARRRR.51	3.562	2.123	6.029	0.79	0.981	0.793	1.815
ARRRR.56	3.562	2.123	6.029	0.79	0.981	0.793	1.815
AARRR.42	3.575	2.221	6.022	0.80	0.975	0.805	1.714
AARRR.43	3.616	2.265	6.016	0.81	0.979	0.832	1.714

basis of the active and inactive sets, with a reward assigned based on the pIC_{50} of each compound of the data set. When a post-hoc scoreof >6.00 was taken to be the cut-off value, 68 hypotheses survived and were then used in the generation of QSAR models (Table 3). All molecules in the dataset were then aligned, matching to at least three pharmacophore features. Two different approaches to splitting the dataset into training and test sets were used: in the first, the compounds were randomly divided into a 76-member training set and 19-member test set, biasing both sets in order to give structural diversity to them both, and employing the standard 4:1 training/test ratio for a QSAR study. The other approach was based on a Kohonen map artificial neural network or SOM (Fig. 3). Descriptor calculations and autoscaling of the descriptor matrix were the starting point when performing the Kohonen clustering approach. The selected test set members were characterized by the minimum distance from the centroid of each cell in the top map. The standard ratio (76/19 sets) was used in this approach too.

Both approaches to test-set selection selected the same 3D pharmacophore model, so this model was the only one evaluated any further. This best model was DRRRR.115, in which all active molecules in the active set matched the hypothesis.

The Phase statistical analysis for each test-set selection method is shown in Table 4. A statistical analysis that included the R^2 versus RMSE/SD plot was employed to choose the best PLS model for each set selection method. The best model was chosen on the basis of the PLS factor model's minimum RMSE/SD value, where the R^2 value was still higher than 0.9. The principle of the 5:1 training set/PLS factor ratio was respected. The validity of each model was tested based on the calculated correlation coefficient for the test set. The squared correlation (random

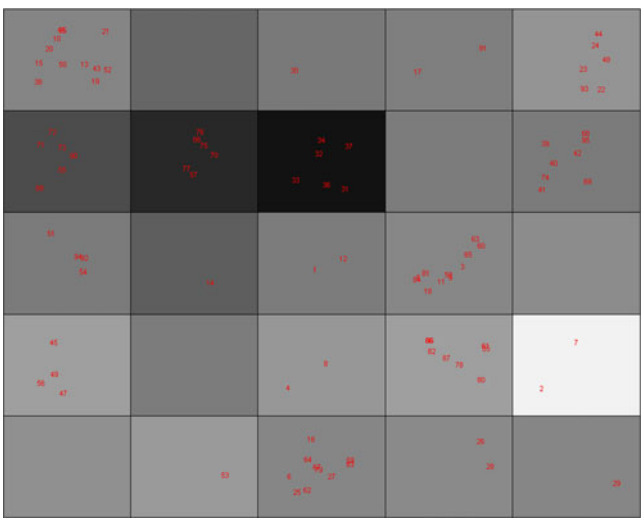

Fig. 3 Kohonen top map

Table 4 Summary of the 3D-QSAR results

Statistical parameters	Random selection of test set	Kohonen map for test set selection
R^2	0.96	0.93
Number of molecules in training set	76	76
Number of molecules in test set	19	19
Optimum number of components	6	5
SD	0.1869	0.2328
F value	251.3	181.6
Pearson R	0.88	0.87
RMSE	0.4241	0.4477
R_{pred}^2	0.72	0.75

selection R_{pred}^2=0.72), the Pearson R (=0.88), and the root mean square error for test set predictions (RMSE=0.42) all confirm the good predictive capabilities of the final QSAR model for the test set. In the case of the Kohonen map ANN, the model with five PLS factors was chosen as the optimum one. The R_{pred}^2(=0.75), Pearson R(=0.87), and RMSE(=0.44) values confirm the validity of this model (Table 5; Fig. 4).

Analysis of the atom-based 3D-QSAR model

The best model, DRRRR.115, consists of four aromatic rings and one hydrogen-bond donor. The spatial arrangement of the pharmacophore sites shows that three aromatic rings, two of which are very close to each other, occupy the hydrophobic pocket created by the residues Val603, Leu595, Leu799, Phe811, and Tyr672. The hydrogen-bond donor site is located between the three aromatic rings and the fourth one, which is 8–10 Å away from the others (Fig. 5). The docking modes of the most active inhibitors resemble the docking pose of imatinib. In fact, for both the thienopyrimidines and aminobenzoisoxazoles, fused aromatic rings lie in the hydrophobic pocket created by the residues mentioned above. It should be emphasized that in each case the interactions of the target with the inhibitor are stabilized by at least two H-bonds. While the residues involved are Asp810, Cys673, and Thr670 in the case of imatinib, Glu640 and Glu671 are involved in the inhibitors from the dataset. Only in the case of compound 71 does the number of H-bonds with Cys673 rise up to four. For derivative 95, the residues involved are the same as those for imatinib. The substantial difference between imatinib and the inhibitors from the dataset is that the former acts as an H-acceptor for two sites and as an H-donor for one site, while the inhibitors from the dataset have only donor sites. It should be underlined that, although just one H-bond donor region and four aromatic features are present in the best model (DRRRR.115), analysis of the docking modes of the most active inhibitors indicates the formation of more extensive interaction patterns. However, in the pharmacophore hypothesis, the only features common to all active inhibitors in the data set are already taken into account.

Table 5 Results of 3D-QSAR analysis with the random selection method (a) and the Kohonen map method (b)

No. of factors	SD	R^2	F	RMSE	R_{pred}^2	Pearson R	RMSE/SD	Δ	Opt.model
(a)									
1	0.6638	0.4078	51	0.5589	0.5155	0.7336	0.911	-	-
2	0.5351	0.6203	59.6	0.4644	0.6654	0.833	0.997	0.086	-
3	0.4438	0.7424	69.2	0.5139	0.5903	0.8199	0.860	−0.137	-
4	0.34	0.8509	101	0.4521	0.683	0.8621	1.307	0.447	-
5	0.2321	0.9315	190	0.4605	0.6711	0.8565	1.923	0.616	-
6	0.1869	0.9562	251	0.4241	0.7209	0.8852	2.577	0.654	√
7	0.1445	0.9742	367	0.4165	0.7309	0.885	3.300	0.723	-
(b)									
1	0.6187	0.4587	63.6	0.5635	0.5988	0.7999	0.842	-	-
2	0.5262	0.6137	58.8	0.5244	0.6525	0.8141	0.868	0.026	-
3	0.4256	0.7507	73.3	0.366	0.8307	0.9349	1.158	0.29	-
4	0.3149	0.8654	116	0.4115	0.7861	0.8886	1.329	0.171	-
5	0.2328	0.9275	182	0.4477	0.7468	0.865	1.984	0.655	√
6	0.1745	0.9598	279	0.4497	0.7445	0.863	2.269	0.285	-
7	0.1373	0.9755	392	0.4531	0.7407	0.8608	2.882	0.613	-

Δ Difference from preceding

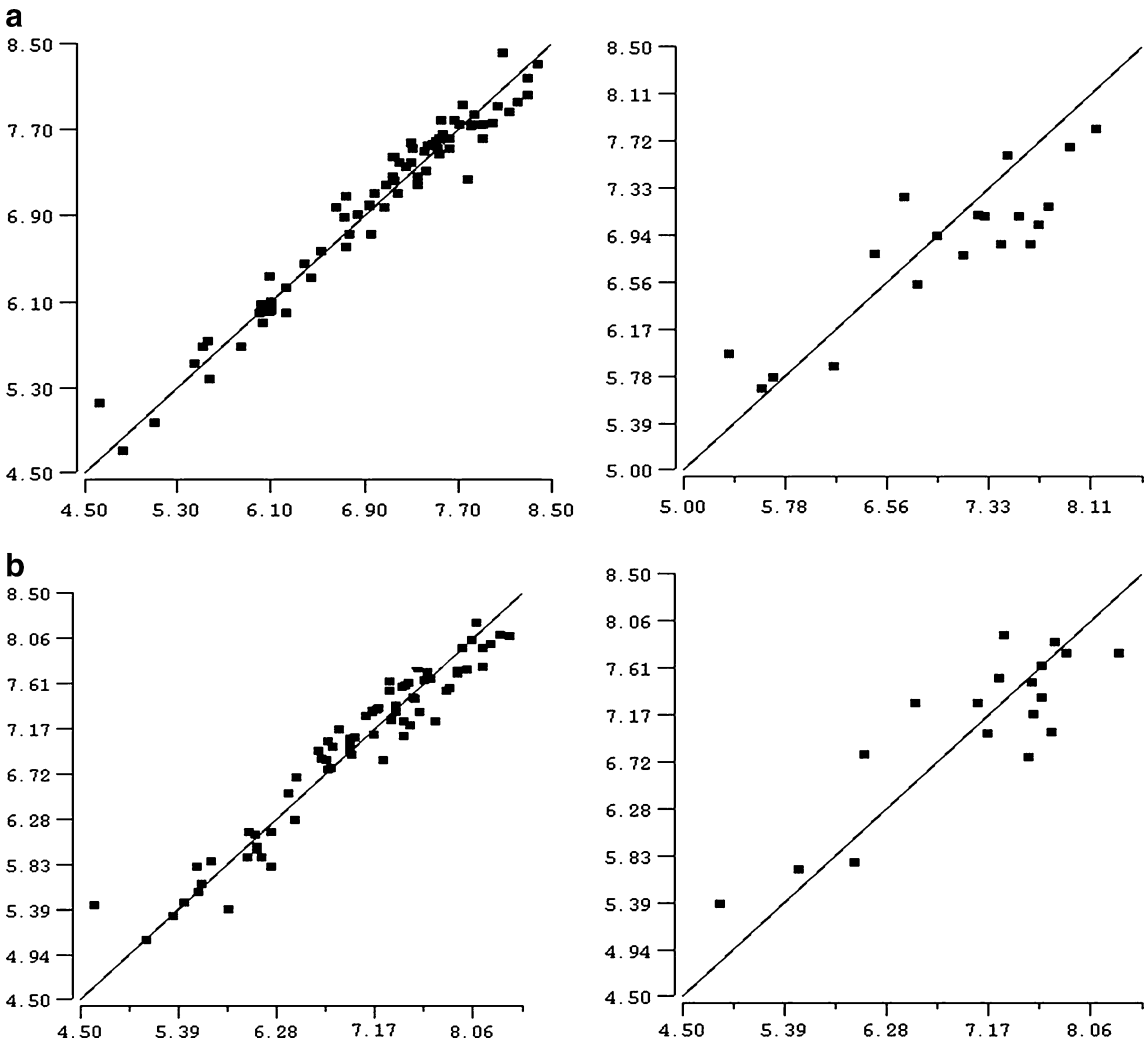

Fig. 4 Actual versus predicted values for training and test sets: **a** random, **b** Kohonen

This means also that, for example, several compounds present more than four aromatic rings, but only four are common to all inhibitors. Figure 6 shows the volume occlusion maps for the atom-based 3D-QSAR model (donor, aromatic ring, electron-withdrawing features). These maps represent the regions of favorable and

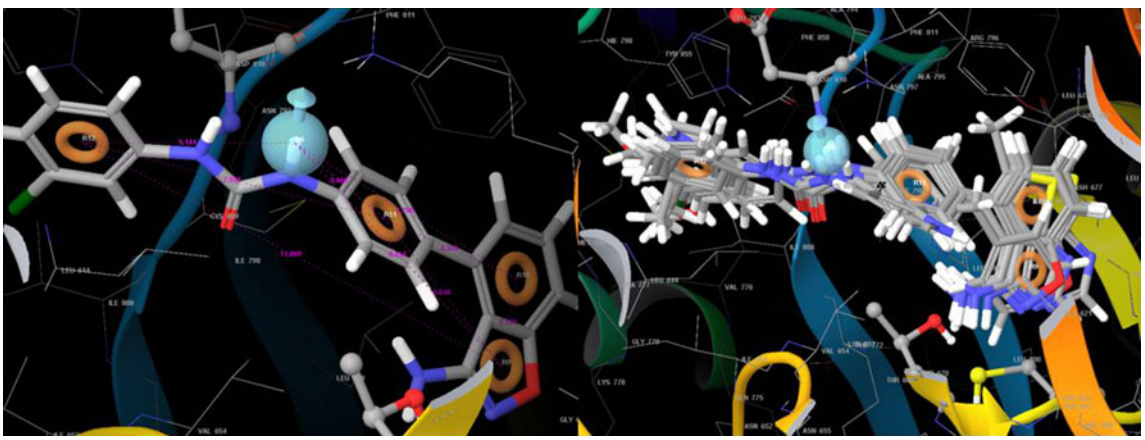

Fig. 5 Pharmacophore mapping of the most active compounds (*left*); superposition of active compounds on the pharmacophore

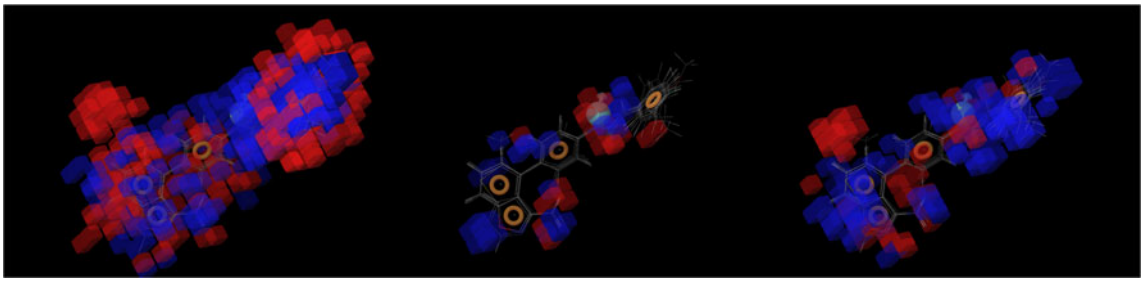

Fig. 6 Occlusion maps. *Left*: hydrophobic; *middle*: donor; *right*: electron-withdrawing features

unfavorable interactions (shown in blue and red, respectively). The volume occlusion map for the H-bond donor describes the favorable 3D arrangement of hydrogen-bonding interactions with acceptor groups of the protein. The occlusion map surrounding the active molecules shows blue cubes opposite the oxazole or pyrimidine ring, which describe a favorable H-bond with an acceptor group of the protein.

The volume occlusion map for the electron-withdrawing groups indicates that the most suitable position for this kind of group is in the external aromatic ring, while the presence of these groups in the inner aromatic ring appears to be unfavorable. This analysis indicates that improvements in binding affinity can be achieved by adding electron-withdrawing groups to the external aromatic ring. The hydrophobic volume occlusion map shows mixed coloration, indicating that an increase in the activity can be expected if the marked hydrophobicity of inhibitors is balanced by the presence of hydrophilic features or a reduction in the hydrophobicity.

The model was further validated by assessing its ability to pick out c-kit inhibitors in a known database aside from the model dataset, using the goodness-of-hit (GH score) approach [23]. For this external validation, a database of 234 compounds was created which included 34 known c-kit inhibitors, the structures of which are provided in the "Electronic supplementary material." The structures of the known inhibitors were taken from the binding database. The other compounds employed in the test set used for external validation were selected from among the structures included in the ZINC database, with care taken to ensure that none of the molecules was structurally correlated with those in the original training and test sets. A summary of the results is provided in Table 6. When the 3D-QSAR model DRRRR.115 was used to query this database, 28 molecules were retrieved as hits (H_t); among these, 23 molecules were known active inhibitors (H_a). The calculated GH score and enrichment factor for the model were 0.74 and 4.83, respectively. The numbers of false positives and false negatives were 11 and 5, respectively. Thus, 82.14% of the hits retried by the model from the database were active inhibitors.

Conclusions

In summary, the goal of this study was to establish a strong relationship between structural features and inhibitory activity. Using a selected set of c-kit inhibitors, a 3D-QSAR pharmacophore model was obtained, consisting of one donor site and four aromatic rings. Volume occlusion maps demonstrated that inhibitory activity can be increased by modulating the donor abilities of the nitrogen or oxygen atoms in the fused aromatic rings that are involved in the H-bond interactions with the binding site of the receptor. This model was validated by predicting the active inhibitors in a test set prediction, and then by using it to pick active inhibitors from a known database. The model generated can be used to query databases and to provide guidelines when designing more potent inhibitors.

Table 6 GH-score values

Property	
Total number of compounds in database (D)	234
Total number of active inhibitors in database (A)	34
Total number of hits (H_t)	28
Number of hits that were active inhibitors (H_a)	23
% Yield of active inhibitors	82.14
% Ratio of active inhibitors to hits	67.64
Enrichment factor (E)	4.83
Number of false negatives	11
Number of false positives	5
GH score	0.75

References

1. Eklund KK (2007) Mast cells in the pathogenesis of rheumatic diseases and as potential targets for anti-rheumatic therapy. Immunol Rev 217:38–52
2. Marshall J (2004) Mast-cell responses to pathogens. Nat Rev Immunol 4:787–799
3. Linnekin D (1999) Early signaling pathways activated by c-kit in hematopoietic cells. Int J Biochem Cell Biol 31:1053–1074
4. Hirota S, Isozaki K, Moriyama Y et al (1998) Gain-of-function mutations of c-kit in human gastrointestinal stromal tumors. Science (Washington, DC) 279:577–580
5. Wang WL, Healy ME (2000) Growth inhibition and modulation of kinase pathways of small cell lung cancer cell lines by the novel tyrosine kinase inhibitor STI 571. Oncogene 19:3521–3528
6. Heinrich MC, Griffith DJ, Druker BJ, Wait CL, Ott KA, Zigler AJ (2000) Inhibition of c-kit receptor tyrosine kinase activity by STI 571, a selective tyrosine kinase inhibitor. Blood 96:925–932
7. Bellone G, Carbone A, Sibona N et al (2001) Aberrant activation of c-kit protects colon carcinoma cells against apoptosis and enhances their invasive potential. Cancer Res 61:2200–2206
8. Potashman MH, Bready J et al (2007) Design, synthesis, and evaluation of orally active benzimidazoles and benzoxazoles as vascular endothelial growth factor-2 receptor tyrosine kinase inhibitors. J Med Chem 50:4351–4373
9. Kunz RK, Rumfelt S (2008) Discovery of amido-benzisoxazoles as potent c-kit inhibitors. Bioorg Med Chem Lett 18:5115–5117
10. Ji Z, Ahmed AA, Albert DH et al (2008) 3-Amino-benzo[*d*] isoxazoles as novel multitargeted inhibitors of receptor tyrosine kinases. J Med Chem 51:1231–1241
11. Dai Y, Guo Y, Frey RR et al (2005) Thienopyrimidine ureas as novel and potent multitargeted receptor tyrosine kinase inhibitors. J Med Chem 48:6066–6083
12. Bold G, Altmann KH, Frei J et al (2000) New anilinophthalazines as potent and orally well absorbed inhibitors of the VEGF receptor tyrosine kinases useful as antagonists of tumor-driven angiogenesis. J Med Chem 43:2310–2323
13. Zupan J, Novic M, Ruisánchez I (1997) Kohonen and counter propagation artificial neural networks in analytical chemistry. Chemom Int Lab Syst 38:1–23
14. Gasteiger J, Zupan J (1993) Neural networks in chemistry. Angew Chem Int Ed 32:503–527
15. Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci 22:69–76
16. Lobanov V, Karelson M, Katriztky AR (1996) Codessa v.2.21 software package. Center for Heterocyclic Compounds, University of Florida, Gainesville
17. Ballabio D, Consonni V, Todeschini R (2009) The Kohonen and CP-ANN toolbox: a collection of MATLAB modules for self organizing maps and counterpropagation artificial neural networks. Chemom Intell Lab Syst 98:115–122
18. Schrödinger LLC (2005) MacroModel 91 reference manual. Schrödinger LLC, New York
19. Scherman W, Day T, Jacobson MP, Friesner RA, Farid R (2006) Novel procedure for modeling ligand/receptor induced fit effects. J Med Chem 49:534–553
20. Schrödinger LLC (2009) Glide, v.5.5. Schrödinger LLC, New York
21. Schrödinger LLC (2009) Prime, v.2.1. Schrödinger LLC, New York
22. Dixon SL, Smondyrev AM, Knoll EH, Rao SN, Shaw DE, Friesner RA (2006) PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening. 1. Methodology and preliminary results. J Comput Aided Mol Des 20:647–671
23. Güner OF, Henry DR (1998) Formula for determining the "goodness of hit lists" in 3D database searches. Accelrys/MDL Information Systems, Inc., San Diego/San Leandro. http://www.netsci.org/Science/Cheminform/feature09.html