Journal of Molecular Graphics and Modelling 30 (2011) 135–147

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Multi-conformation 3D QSAR study of benzenesulfonyl-pyrazol-ester compounds and their analogs as cathepsin B inhibitors

Zhigang Zhou*, Yanli Wang, Stephen H. Bryant

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health Bethesda, MD 20894, USA

ARTICLE INFO

Article history:
Received 28 April 2011
Received in revised form 17 June 2011
Accepted 30 June 2011
Available online 7 July 2011

Keywords:
Multi-conformation QSAR
4D QSAR
Docking
CoMFA
CoMSIA
Cathepsin B
Inhibitors
Regression
Partial least squares
PubChem

ABSTRACT

Cathepsin B has been found being responsible for many human diseases. Inhibitors of cathepsin B, a ubiquitous lysosomal cysteine protease, have been developed as a promising treatment for human diseases resulting from malfunction and over-expression of this enzyme. Through a high throughput screening assay, a set of compounds were found able to inhibit the enzymatic activity of cathepsin B. The binding structures of these active compounds were modeled through docking simulation. Three-dimensional (3D) quantitative structure–activity relationship (QSAR) models were constructed using comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) based on the docked structures of the compounds. Strong correlations were obtained for both CoMFA and CoMSIA models with cross-validated correlation coefficients (q^2) of 0.605 and 0.605 and the regression correlation coefficients (r^2) of 0.999 and 0.997, respectively. The robustness of these models was further validated using leave-one-out (LOO) method and training-test set method. The activities of eight (8) randomly selected compounds were predicted using models built from training set of compounds with prediction errors of less than 1 unit for most compounds in CoMFA and CoMSIA models. Structural features for compounds with improved activity are suggested based on the analysis of the CoMFA and CoMSIA contour maps and the property map of the protein ligand binding site. These results may help to provide better understanding of the structure–activity relationship of cathepsin B inhibitors and to facilitate lead optimization and novel inhibitor design. The multi-conformation method to build 3D QSAR is very effective approach to obtain satisfactory models with high correlation with experimental results and high prediction power for unknown compounds.

Published by Elsevier Inc.

1. Introduction

Cathepsins, as lysosomal peptidases, are responsible for intracellular as well as extracelluar proteolysis in mammalian cells. Several cathepsins (such as cathepsin B, D, H, K, L, and S) have been found to be involved in the biological processes related to a variety of diseases and disorders [1–4]. Cathepsin B, a member of the papain superfamily [3], is one of the ubiquitous lysosomal cysteine proteases. Researchers have found that over expression of this enzyme is related to several important human diseases, including neurodegenerative disorder, cardiovascular disease, cancer, inflammation, rheumatoid arthritis, and Alzheimer's disease [5–12]. Cathepsin B has also been shown to play an important role in tumor metastasis and angiogenesis by facilitating cell migration and dissolving the extracellular barriers [13–15]. Furthermore, it is an essential factor to viral entry and replication of several viruses, such as Ebola and SARS in human cells through its proteolysis functions [16–18]. Since cathepsin B has been shown to play an important role in various human diseases, this enzyme was selected as a therapeutic molecular target for drug development. A number of inhibitors have been developed for the disease treatment [19–21], including several inhibitors that were effective against rheumatoid arthritis in animal models [5–10,12,19–21].

Biological studies have demonstrated that many cathepsin inhibitors bind to the enzyme irreversibly [22,23]. The structural biology studies revealed that such inhibitors bind to the active site by forming a covalent bond with the catalytic thiol group (Cys30) of the enzyme [22,23]. The reported irreversible cathepsin inhibitors include dipeptidyl nitriles [23], vinyl sulfones, expoxysuccinates, acyloxymethyl ketones, fluoromethyl ketones, hydrazides, and *bis*-α-amidoketones [24]. However, some other inhibitors have been reported to form reversible covalent bond with the enzyme, such as α-ketoamides and aldehydes [24]. The experimental three dimensional structure shows that inhibitor dipeptidyl nitrile (DPN) binds at the active site, formed by amino acids Gln24, Cys27-Trp31, Gly69, Asp70, Asn73-Pro77, Ala174, Gly199-Ala201, and Glu245, close to the interface of the protein dimer. The inhibitor forms a covalent sulfur bond with protein amino acid Cys30 [22,23].

* Corresponding author. Tel.: +1 301 435 7792; fax: +1 301 480 9241.
E-mail addresses: zhougeor@ncbi.nlm.nih.gov (Z. Zhou), bryant@ncbi.nlm.nih.gov (S.H. Bryant).

1093-3263/$ – see front matter. Published by Elsevier Inc.
doi:10.1016/j.jmgm.2011.06.013

In addition to experimental studies, computational methods have been used to study the binding structures of active inhibitors with the protein with the purpose to characterize interaction features and to interpret inhibition mechanism of small molecules for several members of the cathepsin class [25–27] including cathepsin L [28,29] and cathepsin S [29]. The authors also performed a computational study to establish the interaction and activity mode between these inhibitors and cathepsin B [30,31].

CoMFA and CoMSIA 3D QSAR analysis have been applied in numerous molecular systems to extract useful information from compounds with bio-activity for lead optimization and design. The 3D modeling process requires a predetermined conformation for each compound and an overall molecular alignment among the compounds. Several strategies for building compound alignments have been suggested and successfully applied to 3D QSAR research. Yet, the determination of "active" conformation still remains challenging. Therefore, docking simulation has been applied to determinate the "active" conformation and to perform compound alignment. However, multi-conformations were often obtained for a single compound, thus usually a conformation had to be picked based on simplified rules or assumptions. In the light of the emergence of the 4D QSAR concept [32], an approach involving multi-conformations QSAR was developed and evaluated in the work for the construction of 3D QSAR model. This approach utilized multi-conformations for each small molecule in QSAR model construction and optimization with the purpose of eliminating manual conformation selection and obtaining best correlations.

2. Materials and methods

2.1. Molecular structures and bioactivity

The three-dimensional structure coordinate of cathepsin protein with bound dipeptidyl nitrile (DPN) was obtained from the protein databank (PDB code: 1GMY [23]). The PDB file for the crystallographic structure complex contains three chains (A, B, and C) of cathepsin B and 3 copies of the small molecule ligand. Chains A and B form a dimer and the coordinates of this dimer were used in this study for the 3 dimension (3D) quantitative structure–activity relationship (QSAR) model construction.

The two dimension (2D) chemical structures and the biological activities of the small molecules were taken from the PubChem database website (http://pubchem.ncbi.nlm.nih.gov). The biological activity of these compounds was measured based on a dose response confirmatory screening experiment (PubChem BioAssay ID: 820 and 523 at) [33], where the inhibitory activities for a total of 75 and 27 compounds were tested in assay 820 and 523, respectively, by the Center for Molecular Discovery of University of Pennsylvania. In the confirmatory biological screenings, compounds were reported as active if they have inhibitory activity against the catalytic functionality of the human liver cathepsin B in a multiple-dose response experiment by measuring the release of the fluorophore aminomethyl coumarin (AMC) from the hydrolysis of an AMC-labeled dipeptide. Detailed information for the assay protocol and materials is available on the PubChem website (http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=820, 523). These tested compounds were identified in a HTS assay for the inhibition activity of over 60,000 compounds against cathepsin B where a single dose of compound concentration was tested. Thirty-seven (37) were identified as active compounds, 35 compounds were reported as inactive compounds, while 3 compounds were suggested as "inconclusive" in the results of assay 820. Meanwhile, the numbers of active, inactive, and "inconclusive" compounds reported in assay 523 are 10, 16, and 1, respectively. Six (6) active compounds were reported in both assays. So the number of unique compounds reported as active are 41. Through further examining these active compounds with different protocols, the experimental authors suggested than some of them could be artifacts. To eliminating the potential artifacts, the authors examined these active compounds using several independent bioassay tests (different biological experiments) that were affected by the similar causes of artifacts. The examination resulted in the elimination of five active compounds from our active compound list as they bear greater chance as being false positives. The final 36 active compounds were included in the work. To expend the activity range of the modeled compounds, 7 inactive compounds with the similar sizes and structures were added in the modeling. In training-test experiments, 8 compounds were randomly selected as test set of compounds while the rest 34 compounds were used as training set of compounds to build models.

The 3D structures of these compounds were constructed from their 2D structures using the Molecular Operating Environment (MOE) program (Version 2007.09, developed by Chemical Computing Group, Montreal, Canada). The structure of each organic molecule was minimized until the root mean square (RMS) deviation of energy is smaller than 0.01 using optimized potential for liquid simulation (OPLS) force field [34–37]. The structure and biological activity of these active compounds are listed in Table 1.

2.2. Small molecule alignment by docking simulation

Docking simulation was used to determine the "active" conformation of each compound and align the small molecules according to the structural and physicochemical properties of the binding site.

The cathepsin B structure was prepared following the standard protocol of GLIDE docking program (version 8 in FirstDiscovery suite) [38,39] by adding hydrogen atoms and assigning partial charges using the OPLS-AA force field [34–37]. A coarse energy minimization with a small number of steps was performed to the complex structure to relax side chain amino acids and the added hydrogen atoms. A docking grid was generated based on the refined protein structure with a geometric center at the center of the original ligand and the dimension of the docking grid was defined to enclose the whole binding site.

The ligand compounds were docked into the binding pocket of the protein receptor using the GLIDE program. Conformational flexibility of the ligand was considered by generating rotamers from explicit rotation of single-bond torsion angles. All rotamers for a given ligand were docked and oriented in the site, and grid-based refinement was then executed for 5000 poses. The 800 top refined poses for each ligand were subjected to 400 steps of conjugate gradient energy minimization. Poses of a given compound were saved if they differed from those obtained previously by more than 0.5 Å RMS deviations in heavy atom coordinates, or involved atomic displacements by greater than 1 Å. The flexibility of the receptor was not included in GLIDE docking. The docking poses were ranked by Gscore (GLIDE score) and the top five poses were selected and used for the QSAR model construction. The docking simulation followed the similar dockings reported previously in detail [30,40,41].

2.3. Alignment and multi-conformation selection

Docking simulation was performed to determine the "active" conformation of each ligand and to align the compounds together according to their and the receptor's structural and physicochemical properties and the top five poses were selected for each ligand from the docking results, which were included in the initial dataset for the QSAR model construction. A cluster of 5 conformers (multi-conformation) was initially included in the QSAR modeling. The selection of the final conformation for each compound was done by identifying the conformation that correlates the best with other

Table 1
The molecular structure, property, and biological activity of 35 active and 7 inactive compounds.[a]

Compd. #	CID[b]	Mol. formula	MW	Compound name	IC_{50} (μM)	log IC_{50}	Set
1	286532	$C_{18}H_{14}N_2O_6$	354.09	[4-(4-Methoxybenzoyl)-2-oxido-1,2,5-oxadiazol-2-ium-3-yl]-(4-methoxyphenyl)methanone	11.458	−4.941	Training
2	573353	$C_{16}H_{12}FN_5O$	309.10	4-[1-[(4-Fluorophenyl)methyl]benzimidazol-2-yl]-1,2,5-oxadiazol-3-amine	33.855	−4.470	Training
3	646525	$C_{15}H_{13}N_3O_4S_2$	363.04	[5-Amino-1-(4-methylphenyl)sulfonylpyrazol-3-yl] thiophene-2-carboxylate	1.990	−5.701	Training
4	646749	$C_{18}H_{17}N_3O_5S$	387.09	[5-Amino-1-(4-methoxyphenyl)sulfonylpyrazol-3-yl] 4-methylbenzoate	12.265	−4.911	Test
5	647599	$C_{14}H_{10}FN_3O_5S$	351.03	[5-Amino-1-(4-fluorophenyl)sulfonylpyrazol-3-yl] furan-2-carboxylate	1.261	−5.899	Training
6	648315	$C_{17}H_{14}N_2O_3S_2$	358.05	[2-Oxo-1-pyridin-2-yl-2-(thiophen-2-ylmethylamino)ethyl] thiophene-2-carboxylate	6.356	−5.197	Training
7	651936	$C_{15}H_{13}N_3O_5S$	347.06	[5-Amino-1-(4-methylphenyl)sulfonylpyrazol-3-yl] furan-2-carboxylate	1.749	−5.757	Training
8	653316	$C_{16}H_{18}N_6O_2$	326.15	2-[2-(4-Amino-1,2,5-oxadiazol-3-yl)benzimidazol-1-yl]-1-piperidin-1-ylethanone	44.577	−4.351	Training
9	653862	$C_{15}H_{13}N_3O_6S$	363.05	[5-Amino-1-(4-methoxyphenyl)sulfonylpyrazol-3-yl] furan-2-carboxylate	0.924	−6.035	Test
10	654815	$C_7H_6Cl_2N_2O_4$	251.97	N-[(3,4-dichloro-5-oxo-2H-furan-2-yl)carbamoyl]acetamide	2.120	−5.674	Training
11	655490	$C_{17}H_{15}N_3O_5S$	373.07	[5-Amino-1-(4-methoxyphenyl)sulfonylpyrazol-3-yl] benzoate	9.563	−5.019	Training
12	658111	$C_{17}H_{21}N_3O_4S$	363.13	Methyl(5-cyano-3,3-dimethyl-8-morpholin-4-yl-1,4-dihydropyrano[4,5-d]pyridin-6-yl)sulfanylformate	6.715	−5.173	Training
13	658152	$C_{14}H_{20}N_6O_4S$	368.13	Diethyl 2-[cyano-(4-dimethylamino-6-methylsulfanyl-1,3,5-triazin-2-yl)amino]propanedioate	19.686	−4.706	Training
14	658724	$C_{19}H_{16}N_2O_4$	336.11	[4-[(2-Methoxyphenyl)iminomethyl]-2-phenyl-1,3-oxazol-5-yl] acetate	8.927	−5.049	Test
15	658964	$C_{20}H_{18}N_2O_4$	350.13	[4-[(2-Methoxyphenyl)iminomethyl]-2-phenyl-1,3-oxazol-5-yl] propanoate	39.987	−4.398	Training
16	660829	$C_{19}H_{12}N_2O_5$	348.08	[2-Furan-2-yl-4-(phenyliminomethyl)-1,3-oxazol-5-yl] furan-2-carboxylate	38.471	−4.415	Training
17	665480	$C_{20}H_{29}N_3O_5S$	423.18	tert-Butyl N-[(1R)-2-methyl-1-[5-[(3-methylphenyl)methylsulfonyl]-1,3,4-oxadiazol-2-yl]butyl]carbamate	2.087	−5.680	Training
18	714967	$C_{11}H_{15}N_7O_2$	277.13	2-[Cyano-(4-methoxy-6-pyrrolidin-1-yl-1,3,5-triazin-2-yl)amino]acetamide	14.196	−4.848	Training
19	794694	$C_{12}H_{14}ClNO_3S_2$	319.01	2-(4-Chlorophenyl)sulfonyl-4,5-dimethyl-3,6-dihydrothiazine 1-oxide	4.170	−5.380	Test
20	971438	$C_{18}H_{17}N_5O_2$	335.14	4-[1-[(5-Methoxy-2-methylphenyl)methyl]benzimidazol-2-yl]-1,2,5-oxadiazol-3-amine	37.190	−4.430	Training
21	1506381	$C_{17}H_{15}N_5O_2$	321.12	4-[1-[(3-Methoxyphenyl)methyl]benzimidazol-2-yl]-1,2,5-oxadiazol-3-amine	45.971	−4.338	Training
22	2212050	$C_{15}H_{13}N_3OS$	283.08	Benzotriazol-1-yl-(2-ethylsulfanylphenyl)methanone	7.114	−5.148	Training
23	2998380	$C_{16}H_{16}N_2O_4S_3$	396.03	N-(2-phenylsulfonyl-3,6-dihydrothiazin-1-ylidene)benzenesulfonamide	9.389	−5.027	Test
24	3236798	$C_{16}H_{14}N_6O_2$	322.12	1,3-Dimethyl-5-phenyl-6-(1,2,4-triazol-4-yl)pyrrolo[3,4-e]pyrimidine-2,4-dione	1.185	−5.926	Training
25	3240114	$C_{15}H_{13}N_3O_5S_2$	379.03	[5-Amino-1-(4-methoxyphenyl)sulfonylpyrazol-3-yl] thiophene-2-carboxylate	0.692	−6.160	Training
26	3241895	$C_{14}H_{10}FN_3O_4S_2$	367.01	[5-Amino-1-(4-fluorophenyl)sulfonylpyrazol-3-yl] thiophene-2-carboxylate	0.435	−6.362	Training
27	3243025	$C_{11}H_8F_3NO_4$	275.04	[3-Oxo-2-(trifluoromethyl)-4H-1,4-benzoxazin-2-yl] acetate	0.845	−6.073	Test

Table 1 (*Continued*)

Compd. #	CID[b]	Mol. formula	MW	Compound name	IC_{50} (μM)	log IC_{50}	Set
28	3243128	$C_{14}H_{11}N_3O_4S_2$	349.02	(5-Amino-1-phenylsulfonylpyrazol-3-yl) thiophene-2-carboxylate	0.247	−6.608	Training
29	3243168	$C_{16}H_{12}N_2O_6$	328.07	(1,3-Dioxoisoindol-2-yl)methyl 2-(furan-2-carbonylamino)acetate	8.563	−5.067	Training
30	3250046	$C_{22}H_{19}NO_6$	393.12	[2-(4-Methoxyphenyl)-2-oxo-1-phenylethyl] 2-(furan-2-carbonylamino)acetate	18.349	−4.736	Training
31	5293426	$C_{19}H_{14}N_4O_3S$	378.08	2-[[5,6-Di(furan-2-yl)-1,2,4-triazin-3-yl]sulfanyl]-N-phenylacetamide	2.247	−5.648	Test
32	11834381	$C_{15}H_{13}N_3O_4S_2$	363.41	[5-Amino-1-(benzenesulfonyl)pyrazol-3-yl]4-methylthiophene-2-carboxylate	2.82	−5.550	Training
33	11834392	$C_{20}H_{15}N_3O_6S_3$	489.54	[1-(4-Methoxyphenyl)sulfonyl-5-(thiophene-2-carbonylamino)pyrazol-3-yl] thiophene-2-carboxylate	3.23	−5.490	Training
34	3685806	$C_9H_9N_3O_4S_2$	287.32	(5-Amino-1-methylsulfonylpyrazol-3-yl) thiophene-2-carboxylate	22.28	−4.652	Training
35	11834389	$C_{15}H_{12}N_2O_4S_2$	348.40	[1-(4-Methylphenyl)sulfonylpyrazol-3-yl] thiophene-2-carboxylate	33.10	−4.480	Training
36	1205147	$C_{23}H_{26}N_6O_4$	450.20	2-[2-(4-Amino-1,2,5-oxadiazol-3-yl)benzimidazol-1-yl]-N-[2-(3,4-diethoxyphenyl)ethyl]acetamide	500	−3.300	Test
37	1306035	$C_{26}H_{20}N_2O_4$	424.14	N-(3-cyano-4,5-diphenylfuran-2-yl)-2-(2-methoxyphenoxy)acetamide	500	−3.300	Training
38	3236935	$C_{19}H_{25}N_3O_6S$	423.15	2-[N-[(3,5-dimethyl-1,2-oxazol-4-yl) sulfonyl]-4-ethoxyanilino]-N-(oxolan-2-ylmethyl)acetamide	500	−3.300	Training
39	3240711	$C_{22}H_{26}N_4O_4S$	442.17	N-(2,3-dimethylphenyl)-2-[N-[(3,5-dimethyl-1H-pyrazol-4-yl)sulfonyl]-4-methoxyanilino]acetamide	500	−3.300	Training
40	3244032	$C_{14}H_{12}N_2S_2$	272.04	6-Cyclopropyl-2-methylsulfanyl-4-thiophen-2-ylpyridine-3-carbonitrile	500	−3.300	Training
41	3333	$C_{18}H_{19}Cl_2NO_4$	383.07	3-O-ethyl 5-O-methyl 4-(2,3-dichlorophenyl)-2,6-dimethyl-1,4-dihydropyridine-3,	500	−3.300	Training
42	380199	$C_{11}H_{13}N_3OS$	235.08	N′-(6-methoxy-1,3-benzothiazol-2-yl)-N,N-dimethylmethanimidamide	500	−3.300	Training

[a] The information was obtained from CID: PubChem database (BioAssay #: 820).
[b] CID: compound PubChem identification number; Mol. formula: molecular formula; MW: molecular weight.

compounds, a process that involving multiple-steps of iterations. In each step, the conformation with the lowest correlation coefficient with all other compounds was eliminated from the list. Through the iteration, the conformation with the highest correlation coefficient was retained at the end, one for each compound. The last one conformation of each compound was then used to build the final QSAR model.

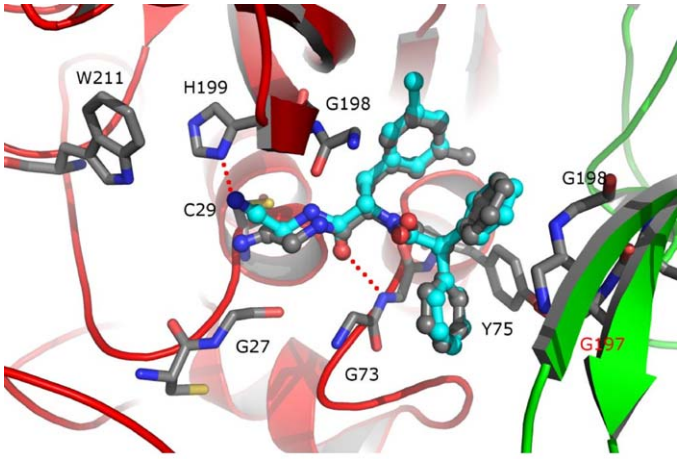

Fig. 1. The docked pose (stick-ball mode in cyan) of DPN compared with the experimental structure of the inhibitor (stick-ball mode in grey) in the crystal complex (PDB code: 1GMY). The key interacting residues are shown in stick mode. The red dot lines indicate the hydrogen bonds formed between the docked pose of DPN and protein amino acids. The dimer proteins are shown in ribbon-cartoon mode with the primary protein in red and the second protein in green.

2.4. QSAR modeling by statistical analysis

Tripos Sybyl 7.0 program was used to conduct CoMFA and CoMSIA QSAR modeling. The Partial Least-Squares (PLS) analysis method was used to conduct statistical analysis and to derive a QSAR model based on the field values of the CoMFA and CoMSIA descriptors. The CoMFA standard scaling and column filtering of 5.0 were used in PLS analysis.

Leave-one-out (LOO) cross-validation in PLS was used to obtain the optimum number of components (ONC) for building the regression models, where ONC is defined as the number of components used to construct model with the smallest cross-validated standard error (or the number giving the largest value of q^2, as they are consistent most of the time). The LOO technique predicts the activity of each compound using a QSAR model built by excluding that particular compound, providing a good way to quantitatively evaluate the internal predictive ability of a model. By literarily calculating the activity for all compounds in a data set, the LOO attempts to evaluate the robustness of a model and lower the over-fitting effect that could exist in conventional regression analysis due to the inclusion of the compound information in its prediction. The quality of a model is expressed as the cross-validated correlation coefficient q^2.

The optimum number of components obtained by cross-validated regression was then used in all conventional fitting regression to derive the final QSAR models including all of the compounds (without cross-validation). Four methods including LOO, cross-validation, Bootstrapping validation, and training-test sets were used for model cross-validations. The Bootstrapping method performs a statistical simulation by randomly generating many new data sets from the original one and calculating statistical parameters for the original set and each of the new sets (see Tri-

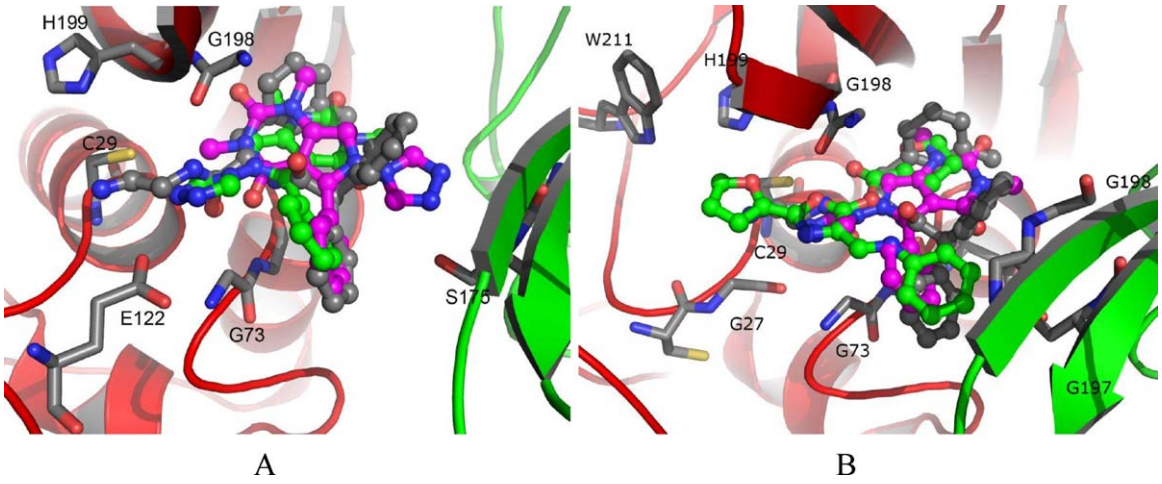

Fig. 2. The superposition of the docked poses of the compounds with the experimental structure of the DPN inhibitor (in grey) in the crystal complex (PDB code: 1GMY). The ligands are depicted in stick-ball mode and key interaction residues are in stick mode. The dimer proteins are shown with ribbon-cartoon mode with the primary protein in red and the secondary protein in green. (A) The docked poses (in green and magenta) of compound 24 (CID: 3236798). The major difference of the two docked poses for compound 24 is the triazole ring which flips over from one side to the other. (B) The comparison of the docked pose of compound 16 (in green, CID: 660829) and 24, and the experimental structure of DPN (in grey). It indicates that the two compound can adapt a similar binding structures in the binding site. (For interpretation of the references to color in the figure caption, the reader is referred to the web version of the article.)

pos Sybyl manual on CD for details). The cross-validated correlation coefficient (q^2) calculated from these cross-validations were used to evaluate the predictability and robustness of a model and the conventional correlation coefficient (r^2) is used to measure the quality of the model.

3. Results and discussion

3.1. Docking validation

The docking protocol was first validated by reproducing the binding structure of the DPN ligand to active site of cathepsin B protein in the experimental crystal complex (PDB code: 1GMY [23]). The dimer coordinates were extracted from the PDB file and used for the docking simulations. In the original crystal complex structure, the DPN ligand is covalently linked to the thiol group of amino acid Cys30 of cathepsin B [23]. Prior to the docking grid preparation, the S-C bond between Cys30 and DPN was first broken to separate the ligand from the protein, and the N-C group of the ligand was modified into linear nitrile, a pre-reaction format. A docking grid was then generated based on the protein dimer using GLIDE after DPN ligands were removed from the active sites. To validate the docking protocol, the DPN ligand was then docked back into the active site of the first protein of the dimer. As depicted in Fig. 1, the docked pose from the dominant cluster [40] superimposes very well with the original ligand except that the methyl-phenyl ring close to G198 rotated for 180°, and as a result, the methyl group turns to the opposite direction. It seems that it is difficult for GLIDE to prioritize the original conformation, either due to lack of a structural element to distinguish the two, or because the experimental conformer does not provide enough binding elements to be recognized by the force field employed in GLIDE. Nevertheless, the overlap RMS deviations between the docked poses and the original pose are ~1.2 Å. If the 6 carbons of the methyl-phenyl ring are not distinguished, the RMS deviations will be as low as 0.4 Å. In the docked structure, it is noticed that the nitrile group has potential to form hydrogen bond with H199 amino acid. The hydrogen bonding helps to stabilize the position of the nitrile group that C29 amino acid attacks to form covalent bond. The results show that the docking simulation protocol is able to reproduce the ligand binding structure as observed in the experimental crystal complex. The same docking protocol was applied to determine the "bound" conformations for the cathepsin B inhibitors, as well as, to align these compounds together according to the receptor features for the QSAR study.

3.2. Conformation determination and alignment construction

When applying the described docking protocol to this set of compounds, binding structures were obtained from the docking simulation for 43 of them. The docking simulation did not output any pose for one compound. The results of these 42 compounds were included in the QSAR study. The molecular formula, name, PubChem compound ID (CID) and the activity (IC_{50}) of these compounds as obtained from PubChem BioAssay database (AID: 820 and 523), are listed in Table 1.

It was noted that multiple docking poses were obtained for all these compounds especially for smaller compounds that can be docked into the binding pocket with different modes and only slight differences on docking score (Gscore). As shown in Fig. 2A, two poses of compound 24 (CID: 326798) obtained from the docking simulation have different binding structures, where the whole molecule flip over with the triazole ring directs to opposite directions, but with the two different orientations, the molecule fits well into the binding site and align well with the experimental bound structure of DPN. b. Interestingly, with the respective binding modes reflected by the two docking poses, the triazole ring either had direct hydrogen bond/ionic interaction with amino acid E122 of the first protein or with amino acid S175 of the second protein. Since both binding modes have strong interactions with one part or the other of the protein dimer, it was difficult to prioritize one over the other merely based on docking score or general knowledge. The similar phenomena have been observed in other docked compounds. This uncertainty is sometimes common in docking simulations, and selection among the multi docked poses for one compound is the key to the success of QSAR and other computational studies. In the docking simulations, similar binding positions and orientations have been observed among the poses of the similar compounds. As exampled in Fig. 2B, the docked poses of compound 16 (CID: 660829) and 24 were superposed very well. Also their docked structures are well aligned with DPN. How to determine which theoretically derived binding mode is the closest to the status in the biological environment is a challenge. 4D QSAR was suggested to tackle the complication/difficulty on the conformer

selection for each compound by including multi-conformations of each small molecule into QSAR model construction. In the light of this method, a multi-conformation 3D QSAR approach is introduced in the work to model the 3D QSAR models of cathepsin inhibitors. The conformers of each compound were generated from docking simulations and multi-conformers were included for each compound in QSAR constructions. Multi-step iteration was adopted to eliminate the abundant conformers to get the final models. To balance the accuracy, complexity, and computer time consumption, we included five top poses for each compound based on the Gscore in this work to evaluate the multi-conformation QSAR concept.

3.3. QSAR model construction

The five selected poses of each compound were initially included in model constructions. CoMFA and CoMSIA descriptors were calculated separately to build models. Based on the calculated field values, the correlation of each sample (incident, row) with all the rest were analyzed and evaluated using PLS statistical methods, and the least correlated pose for each compound was eliminated from the dataset. The PLS method has been used in numerous applications in correlating the activity with various physicochemical properties, especially when the number of descriptors is larger than the number of samples (incidents, row) in which a traditional multi-regression is not suitable. The PLS regression tries to build a correlation between a dependent variable (normally an activity) and several independent variables (property descriptors).

The same statistical analysis process was iterated to eliminate extra conformations for each compound step by step. In each step, the docked conformation with least correlation to other compounds was removed for each compound until a single pose was left for each of the compounds in the optimized dataset. The final optimized set with single conformer for each compound was used to construct the final models. The superposition of the selected docking pose for the 36 active compounds is shown in Fig. 3. It was seen that all these compounds aligned well against each other based on their molecular similarity and fit properly into the binding sites considering the structural feature (shape) and the physicochemical properties (pharmacophore) of the binding site. In this approach, the binding site information of the receptor was integrated into molecular alignment of small molecule ligands and was

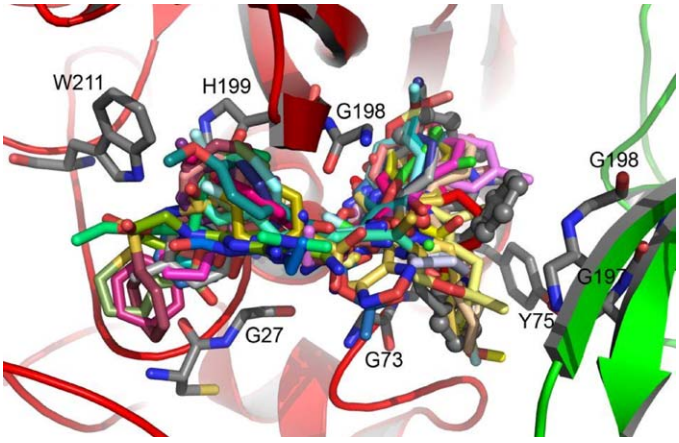

Fig. 3. The cluster of all compounds docked in the binding site of cathepsin B. The inhibitors are represented as stick mode in different colors. The original DPN ligand is shown in stick-ball mode with grey color. The proteins of the dimer are depicted as ribbon-cartoon mode, the primary protein is in red and the second is in green. (For interpretation of the references to color in the figure caption, the reader is referred to the web version of the article.)

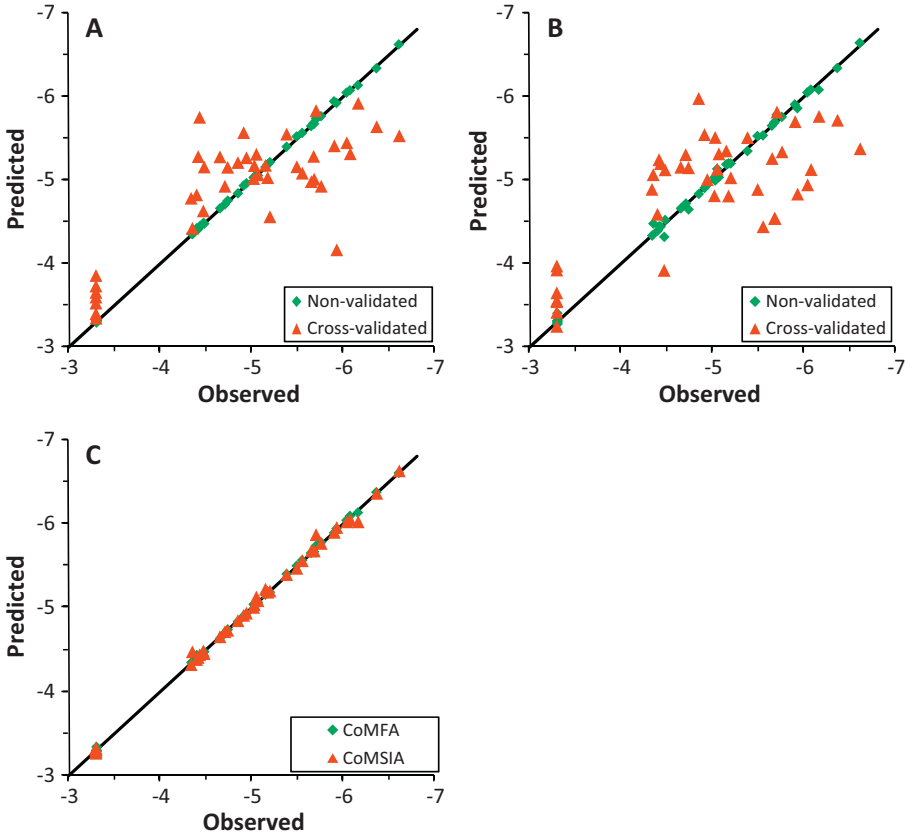

Fig. 4. The plot of predicted vs. observed bioactivity ($\log IC_{50}$) for compounds predicted from the models. (A) Predicted activities by CoMFA modes. (B) Predicted activities by CoMSIA models. (C) Predicted activities using Bootstrapping methods based on CoMFA and CoMSIA modes.

further integrated into 3D QSAR construction. The alignment based on the binding site focuses on the overall molecular fit and pharmacophore overlay, which resulted in a different image of compound superposition comparing to that derived by atom fitting as shown in Fig. 3, in which certain common molecular fragments used for alignment were over emphasized in fitting. Studies have showed that the docking alignment based on this approach resulted in better results than atom fit alignment method [42].

The statistical analyses for QSAR model construction were carried out based on the clusters of molecules selected as described earlier. An optimum numbers of components were first determined separately by cross-validation regression for both the CoMFA and CoMSIA models, which were then applied in the final analysis in the subsequent study. Both the CoMFA and CoMSIA models demonstrated good statistical results as shown in Table 2 (whole set). The regression coefficient (r^2) and the cross-validated coefficient (q^2) of the QSAR model constructed by CoMFA are 0.999 and 0.605, respectively. The corresponding coefficients for the CoMSIA model are 0.997 and 0.605, respectively. There is no apparent difference in the quality of the two models, although results from the CoMSIA model show slightly higher cross-validated correlation coefficients than those from the CoMFA model, which indicates moderately enhanced predictability of the CoMSIA model. The standard errors of estimate for the two models are comparable as 0.012 and 0.054, respectively. In the CoMFA model, the steric descriptors make marginally larger contribution to the model than the electrostatic descriptors (0.568 vs. 0.432). In the CoMSIA model, the hydrophobic descriptors are the largest contributor and the electrostatic

Table 2
Summary of statistics and field contributions calculated using different methods for the best CoMFA mode.

Parameter[a]	Whole set		Training set	
	CoMFA	CoMSIA	CoMFA	CoMSIA
No. of components	6	8	6	8
r^2	0.999	0.997	0.999	0.998
q^2	0.605	0.605	0.562	0.540
SEE	0.016	0.054	0.017	0.045
F-value	18052.7	1523.9	12359.8	1778.7
Steric	0.568	0.154	0.543	0.140
Electrostatic	0.432	0.153	0.457	0.156
Hydrophobic	N/A[b]	0.225	N/A	0.241
Donor	N/A	0.224	N/A	0.222
Acceptor	N/A	0.245	N/A	0.242

[a] r^2: regression (noncross-validated) correlation coefficient. q^2: cross-validated correlation coefficient. SEE: standard error of estimate. *F*-value: *F* test value. No. of components: the optimum number of components obtained from cross-validated PLS analysis which was used in final regression analysis.
[b] N/A: not applicable.

Table 3
The predicted bioactivity by CoMFA models vs. the observed bioactivity.

Compd. #	CID[a]	log IC_{50}	Non-validated		Cross-validated		Bootstrapping	
			Activity	Residue	Activity	Residue	Activity	Residue
1	286532	−4.941	−4.952	0.011	−5.264	0.323	−4.930	−0.011
2	573353	−4.470	−4.476	0.006	−4.624	0.154	−4.471	0.001
3	646525	−5.701	−5.729	0.028	−5.831	0.130	−5.736	0.035
4	646749	−4.911	−4.924	0.013	−5.566	0.655	−4.901	−0.010
5	647599	−5.899	−5.935	0.036	−5.408	−0.491	−5.890	−0.009
6	648315	−5.197	−5.205	0.008	−4.554	−0.643	−5.184	−0.013
7	651936	−5.757	−5.757	0.000	−4.921	−0.836	−5.776	0.019
8	653316	−4.351	−4.341	−0.010	−4.415	0.064	−4.356	0.005
9	653862	−6.035	−6.039	0.004	−5.444	−0.591	−6.039	0.004
10	654815	−5.674	−5.674	0.000	−5.281	−0.393	−5.668	−0.006
11	655490	−5.019	−5.022	0.003	−5.013	−0.006	−5.039	0.020
12	658111	−5.173	−5.179	0.006	−5.025	−0.148	−5.187	0.014
13	658152	−4.706	−4.700	−0.006	−4.921	0.215	−4.701	−0.005
14	658724	−5.049	−5.053	0.004	−5.306	0.257	−5.060	0.011
15	658964	−4.398	−4.412	0.014	−4.818	0.420	−4.436	0.038
16	660829	−4.415	−4.404	−0.011	−5.278	0.863	−4.393	−0.022
17	665480	−5.680	−5.665	−0.015	−5.006	−0.674	−5.665	−0.015
18	714967	−4.848	−4.831	−0.017	−5.206	0.358	−4.836	−0.012
19	794694	−5.380	−5.390	0.010	−5.548	0.168	−5.400	0.020
20	971438	−4.430	−4.435	0.005	−5.750	1.320	−4.423	−0.007
21	1506381	−4.338	−4.348	0.010	−4.777	0.439	−4.346	0.008
22	2212050	−5.148	−5.146	−0.002	−5.172	0.024	−5.155	0.007
23	2998380	−5.027	−5.016	−0.011	−5.170	0.143	−5.042	0.015
24	3236798	−5.926	−5.916	−0.010	−4.157	−1.769	−5.940	0.014
25	3240114	−6.160	−6.128	−0.032	−5.919	−0.241	−6.131	−0.029
26	3241895	−6.362	−6.332	−0.030	−5.637	−0.725	−6.373	0.011
27	3243025	−6.073	−6.068	−0.005	−5.311	−0.762	−6.091	0.018
28	3243128	−6.608	−6.618	0.010	−5.528	−1.080	−6.602	−0.006
29	3243168	−5.067	−5.045	−0.022	−5.059	−0.008	−5.062	−0.005
30	3250046	−4.736	−4.743	0.007	−5.149	0.413	−4.740	0.004
31	5293426	−5.648	−5.638	−0.010	−4.976	−0.672	−5.651	0.003
32	11834381	−5.550	−5.556	0.006	−5.079	−0.471	−5.564	0.014
33	11834392	−5.490	−5.513	0.023	−5.156	−0.334	−5.497	0.007
34	3685806	−4.652	−4.651	−0.001	−5.275	0.623	−4.636	−0.016
35	11834389	−4.480	−4.466	−0.014	−5.153	0.673	−4.476	−0.004
36	1205147	−3.300	−3.306	0.006	−3.333	0.033	−3.300	0.000
37	1306035	−3.300	−3.303	0.003	−3.848	0.548	−3.301	0.001
38	3236935	−3.300	−3.279	−0.021	−3.390	0.090	−3.294	−0.006
39	3240711	−3.300	−3.314	0.014	−3.516	0.216	−3.307	0.007
40	3244032	−3.300	−3.298	−0.002	−3.588	0.288	−3.344	0.044
41	3333	−3.300	−3.296	−0.004	−3.721	0.421	−3.296	−0.004
42	380199	−3.300	−3.299	−0.001	−3.645	0.345	−3.296	−0.004

[a] CID: compound PubChem identification number.

Table 4
The predicted bioactivity by CoMSIA models vs. the observed bioactivity.

Compd. #	CID	log IC_{50}	Non-validated		Cross-validated		Bootstrapping	
			Activity	Residue	Activity	Residue	Activity	Residue
1	286532	−4.941	−4.94	−0.001	−5.264	0.323	−4.931	−0.01
2	573353	−4.470	−4.314	−0.156	−4.624	0.154	−4.479	0.009
3	646525	−5.701	−5.782	0.081	−5.831	0.13	−5.866	0.165
4	646749	−4.911	−4.901	−0.01	−5.566	0.655	−4.905	−0.006
5	647599	−5.899	−5.904	0.005	−5.408	−0.491	−5.894	−0.005
6	648315	−5.197	−5.195	−0.002	−4.554	−0.643	−5.194	−0.003
7	651936	−5.757	−5.754	−0.003	−4.921	−0.836	−5.762	0.005
8	653316	−4.351	−4.474	0.123	−4.415	0.064	−4.47	0.119
9	653862	−6.035	−6.047	0.012	−5.444	−0.591	−6.02	−0.015
10	654815	−5.674	−5.684	0.01	−5.281	−0.393	−5.675	0.001
11	655490	−5.019	−5.027	0.008	−5.013	−0.006	−5	−0.019
12	658111	−5.173	−5.197	0.024	−5.025	−0.148	−5.18	0.007
13	658152	−4.706	−4.714	0.008	−4.921	0.215	−4.709	0.003
14	658724	−5.049	−5.13	0.081	−5.306	0.257	−5.122	0.073
15	658964	−4.398	−4.388	−0.01	−4.818	0.42	−4.379	−0.019
16	660829	−4.415	−4.442	0.027	−5.278	0.863	−4.403	−0.012
17	665480	−5.680	−5.695	0.015	−5.006	−0.674	−5.673	−0.007
18	714967	−4.848	−4.83	−0.018	−5.206	0.358	−4.841	−0.007
19	794694	−5.380	−5.344	−0.036	−5.548	0.168	−5.391	0.011
20	971438	−4.430	−4.436	0.006	−5.75	1.32	−4.435	0.005
21	1506381	−4.338	−4.33	−0.008	−4.777	0.439	−4.318	−0.02
22	2212050	−5.148	−5.185	0.037	−5.172	0.024	−5.218	0.07
23	2998380	−5.027	−4.991	−0.036	−5.17	0.143	−5.034	0.007
24	3236798	−5.926	−5.858	−0.068	−4.157	−1.769	−5.953	0.027
25	3240114	−6.160	−6.08	−0.08	−5.919	−0.241	−6.019	−0.141
26	3241895	−6.362	−6.34	−0.022	−5.637	−0.725	−6.359	−0.003
27	3243025	−6.073	−6.082	0.009	−5.311	−0.762	−6.046	−0.027
28	3243128	−6.608	−6.643	0.035	−5.528	−1.08	−6.628	0.02
29	3243168	−5.067	−5.03	−0.037	−5.059	−0.008	−5.08	0.013
30	3250046	−4.736	−4.644	−0.092	−5.149	0.413	−4.727	−0.009
31	5293426	−5.648	−5.649	0.001	−4.976	−0.672	−5.674	0.026
32	11834381	−5.550	−5.532	−0.018	−5.079	−0.471	−5.556	0.006
33	11834392	−5.490	−5.525	0.035	−5.156	−0.334	−5.465	−0.025
34	3685806	−4.652	−4.656	0.004	−5.275	0.623	−4.65	−0.002
35	11834389	−4.480	−4.517	0.037	−5.153	0.673	−4.449	−0.031
36	1205147	−3.300	−3.305	0.005	−3.333	0.033	−3.305	0.005
37	1306035	−3.300	−3.278	−0.022	−3.848	0.548	−3.325	0.025
38	3236935	−3.300	−3.309	0.009	−3.39	0.09	−3.285	−0.015
39	3240711	−3.300	−3.266	−0.034	−3.516	0.216	−3.275	−0.025
40	3244032	−3.300	−3.286	−0.014	−3.588	0.288	−3.261	−0.039
41	3333	−3.300	−3.299	−0.001	−3.721	0.421	−3.309	0.009
42	380199	−3.300	−3.397	0.097	−3.645	0.345	−3.305	0.005

descriptors are the second followed by hydrogen acceptor and steric descriptors. No apparent correlation was noticed between these descriptors and experimental activity (log IC_{50}), so hydrogen donor descriptors were not included in the model construction. The result that the hydrophobic descriptor is the largest contributor for the CoMFA QSAR models is consistent with the observation that more than half of the binding site area is occupied by hydrophobic groups (will be discussed in following section as shown in Fig. 8).

The predicted activities and residues (the difference between the predicted and the observed activities) by different statistical methods are summarized in Tables 3 and 4. Four statistical methods, fitting regression (non-validated), LOO validation, cross-validation, and Bootstrapping validation were applied based on both CoMFA and CoMSIA models.

The results for the CoMFA models (Table 3) show that the calculated descriptors were fitted precisely with the experimental activity (log IC_{50}) with the residues between the experimental data and the predicted value less than 0.05 for all compounds (Fig. 5). Most compounds (39/42, 93%) had residue values less than 0.03. The prediction errors of the cross-validation method are slightly higher than those of the fitting regression method. Except three outliers, all other 39 compounds have predicted errors less than 1.0 in the cross-validated results. The Bootstrapping validation produced similar results as the fitting regression. The overall low prediction error rates from the four methods indicate that the models are robust and reliable with strong predictability when applied to this set of compounds. The scattered plots (Fig. 4A) of the predicted activities vs. the observed activities (log IC_{50}) based on the CoMFA models show that the fitting regression produced nearly perfect results as the data dots for the cross-validated results are somewhat scatted away from linear trend line (perfect case).

The predicted results for the CoMSIA models are listed in Table 4. The calculated CoMSIA descriptors were fitted well with the observed activity (log IC_{50}) in the fitting regression model. The overall prediction errors for all compounds are less than 0.2 (Fig. 5). Except for two compounds, the prediction errors for all other compounds are less than 0.1. The Bootstrapping validation also produced competent results with error less than 0.2 for all compounds, while the cross-validated results produced errors less than half (0.5) unit for 30 (71%) compounds. The plots of the predicted vs. observed activities for the CoMFA and CoMSIA models are depicted in Fig. 4A and B, which show that the predicted results of the CoMSIA models are similar to those of the CoMFA models. The predictions of the Bootstrapping validation for the respective CoMFA and CoMSIA models are compared and described in Fig. 4C. It is seen that the Bootstrapping regression produced a very similar results as conventional regression that yielded very accurate predicted activity compared with experimental results in the work. All these validations demonstrated that both the CoMFA and CoMSIA

model have strong predictability for the experimental activity of these 36 compounds.

To further evaluate and validate the models, a training-test set approach was selected to construct QSAR models. The data set was randomly split into training set and test set. The list of compounds in the training set (28 compounds) and the test set (8 compounds) is marked in Table 1. Regression model were constructed using the training set, which was then applied to activity prediction for the compounds contained in the test set. The statistical results for the training set are listed in Table 2, which agree well with the corresponding results of the whole set for both CoMFA and CoMSIA models, which indicates that the training set has very similar statistical properties to the original whole set of compounds.

The predicted activity ($\log IC_{50}$) for the compounds in training set and test set for the CoMFA models are listed in Table 5. It is seen that the calculated data for the training set fits very well to the experimental results with errors less than 0.05 for all compounds. The model constructed from training set of compounds also provide accurate predictions for the biological activities of the test set with the errors between the predicted values and the experimental data less than 1.0 for all of the 8 compounds in that set and less than 0.5 for half of the compounds.

Similar cross-validation analysis was conducted for the CoMSIA models, which again demonstrates strong prediction power of these models. The calculated data (Table 6) for compounds in training set fit well with experimental results with errors less than 0.1 for all compounds and the prediction errors for the all but one tested compounds are less than 1.0.

All results from the three validation experiments and the training-test experiments suggest that the CoMFA and CoMSIA models provide a reliable computational approach for the study of binding activities of the small molecules as cathepsin B inhibitors. The consistent performances and suggestions of these different validation experiments prove the robustness and the prediction power of the QSAR models, thus allow one to utilize the information gained from the models and look into to the interaction mechanisms between the identified inhibitors and their protein receptor.

3.4. Analysis of ligand receptor interaction

Based on the docked structures (Figs. 1–3), apparently several amino acid residues including G27, C29, Y75, G198, H199, and W211 interact directly with the ligands. These interactions (polar and non-polar) between these residues and ligand presumably play important roles in the ligand binding affinity and binding mode recognitions. To further explore the hypothetical interaction feature of a ligand with its receptor, the contributions and distributions of the steric, electrostatic, and hydrophobic factors are analyzed.

Table 5
The predicted bioactivity of 8 randomly selected compounds in test set based on CoMFA model that was constructed with the rest 34 compounds in training set.

Compd. #	CID	$\log IC_{50}$	Predicted	
			Activity	Residue
Training set				
1	286532	−4.941	−4.957	0.016
2	573353	−4.470	−4.471	0.001
3	646525	−5.701	−5.746	0.045
5	647599	−5.899	−5.911	0.012
6	648315	−5.197	−5.187	−0.010
7	651936	−5.757	−5.767	0.010
8	653316	−4.351	−4.344	−0.007
10	654815	−5.674	−5.687	0.013
11	655490	−5.019	−5.018	−0.001
12	658111	−5.173	−5.188	0.015
13	658152	−4.706	−4.693	−0.013
15	658964	−4.398	−4.412	0.014
16	660829	−4.415	−4.415	0.000
17	665480	−5.680	−5.661	−0.019
18	714967	−4.848	−4.826	−0.022
20	971438	−4.430	−4.433	0.003
21	1506381	−4.338	−4.342	0.004
22	2212050	−5.148	−5.133	−0.015
24	3236798	−5.926	−5.920	−0.006
25	3240114	−6.160	−6.131	−0.029
26	3241895	−6.362	−6.349	−0.013
28	3243128	−6.608	−6.605	−0.003
29	3243168	−5.067	−5.066	−0.001
30	3250046	−4.736	−4.747	0.011
32	11834381	−5.550	−5.555	0.005
33	11834392	−5.490	−5.502	0.012
34	3685806	−4.652	−4.647	−0.005
35	11834389	−4.480	−4.460	−0.020
37	1306035	−3.300	−3.315	0.015
38	3236935	−3.300	−3.293	−0.007
39	3240711	−3.300	−3.305	0.005
40	3244032	−3.300	−3.277	−0.023
41	3333	−3.300	−3.295	−0.005
42	380199	−3.300	−3.317	0.017
Test set				
4	646749	−4.911	−5.651	0.740
9	653862	−6.035	−5.487	−0.548
14	658724	−5.049	−5.000	−0.049
19	794694	−5.380	−4.951	−0.429
23	2998380	−5.027	−5.152	0.125
27	3243025	−6.073	−5.182	−0.891
31	5293426	−5.648	−4.896	−0.752
36	1205147	−3.300	−3.457	0.157

Steric and electrostatic contour maps of the CoMFA models are shown in Fig. 6, and the steric, electrostatic, and hydrophobic contour maps of the CoMSIA models are shown in Fig. 7A–C. Compound 26 (CID: 3241895, $\log IC_{50} = -6.362$) is shown along with the contour map to facilitate the analysis. Considering the steric contours of the CoMFA (Fig. 6A) and CoMSIA (Fig. 7A) models first, it is seen

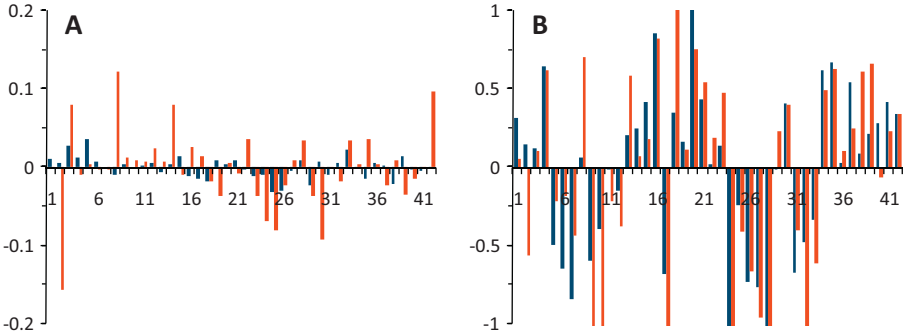

Fig. 5. The plot of residues of the predicted activity compared to observed activity ($\log IC_{50}$) for compounds based on the models. The blue bars are the results of CoMFA results and the red bars are results of CoMSIA modes. (A) Regression (fitting) mode. (B) Cross-validated. (For interpretation of the references to color in the figure caption, the reader is referred to the web version of the article.)

Table 6
The predicted bioactivity of 8 randomly selected compounds in test set based on CoMSIA model that was constructed with the rest 34 compounds in training set.

Compd. #	CID	log IC_{50}	Predicted	
			Activity	Residue
Training set				
1	286532	−4.941	−4.960	0.019
2	573353	−4.470	−4.406	−0.064
3	646525	−5.701	−5.769	0.068
5	647599	−5.899	−5.904	0.005
6	648315	−5.197	−5.202	0.005
7	651936	−5.757	−5.796	0.039
8	653316	−4.351	−4.411	0.060
10	654815	−5.674	−5.729	0.055
11	655490	−5.019	−5.015	−0.004
12	658111	−5.173	−5.188	0.015
13	658152	−4.706	−4.729	0.023
15	658964	−4.398	−4.386	−0.012
16	660829	−4.415	−4.419	0.004
17	665480	−5.680	−5.665	−0.015
18	714967	−4.848	−4.865	0.017
20	971438	−4.430	−4.409	−0.021
21	1506381	−4.338	−4.379	0.041
22	2212050	−5.148	−5.152	0.004
24	3236798	−5.926	−5.936	0.010
25	3240114	−6.160	−6.085	−0.075
26	3241895	−6.362	−6.346	−0.016
28	3243128	−6.608	−6.561	−0.047
29	3243168	−5.067	−5.006	−0.061
30	3250046	−4.736	−4.675	−0.061
32	11834381	−5.550	−5.552	0.002
33	11834392	−5.490	−5.514	0.024
34	3685806	−4.652	−4.636	−0.016
35	11834389	−4.480	−4.485	0.005
37	1306035	−3.300	−3.222	−0.078
38	3236935	−3.300	−3.337	0.037
39	3240711	−3.300	−3.286	−0.014
40	3244032	−3.300	−3.276	−0.024
41	3333	−3.300	−3.296	−0.004
42	380199	−3.300	−3.378	0.078
Test set				
4	646749	−4.911	−5.432	0.521
9	653862	−6.035	−4.828	−1.207
14	658724	−5.049	−4.924	−0.125
19	794694	−5.380	−4.704	−0.676
23	2998380	−5.027	−5.513	0.486
27	3243025	−6.073	−5.141	−0.932
31	5293426	−5.648	−5.123	−0.525
36	1205147	−3.300	−3.428	0.128

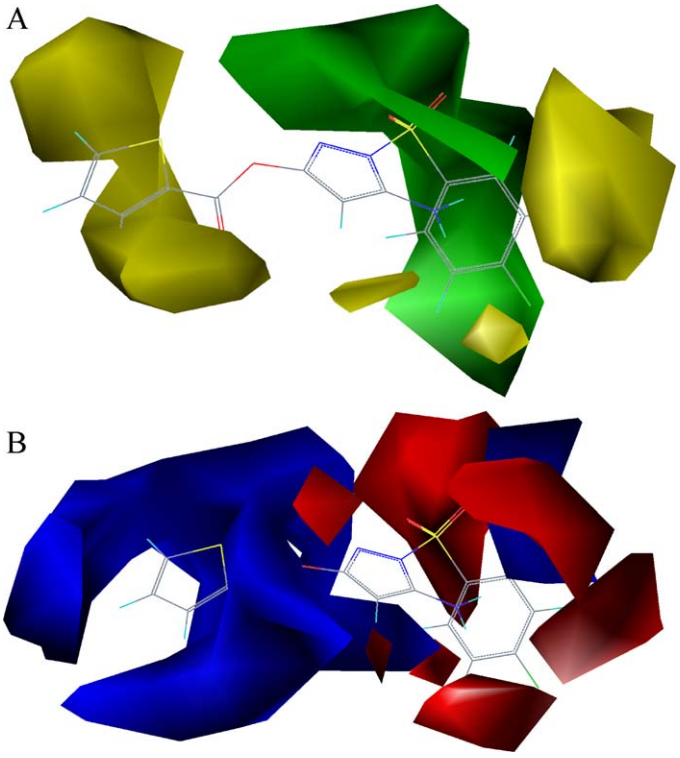

Fig. 6. The contour maps of CoMFA results. (A) The steric map, the favored regions of binding are colored in green and the disfavored regions are colored in yellow. (B) The electrostatic map, the favored regions are colored in blue and the disfavored regions are colored in red. The compound 26 (CID: 3241895, log IC_{50} = −6.362) is shown along with the map. (For interpretation of the references to color in the figure caption, the reader is referred to the web version of the article.)

that two models gave a very similar predictions for the steric interactions. Both models suggest that the areas around thiofuran ring were disfavored regions (yellow). In addition, the areas on phenyl ring facing out of paper were also predicted as disfavored regions (yellow) by both models, which indicate that addition of bulky groups in these regions would not favor activity increase. The areas on the other side of the phenyl ring were predicted as favored regions (green), however, where an addition of bulky groups may increase the compound's activity. By checking the docked structure (Fig. 8) of the compound in cathepsin B protein, it was noted that the favored region predicted from QSAR around the phenyl ring has space for extra substituent to fit in. On the other hand, the disfavored regions on the other side of the phenyl ring are already occupied by amino acid residues of the receptor, and there is no room for extra chemical group. Thus the suggestions from QSAR analysis of these two regions are consistent with those based on the docked structure. On the other hand, while the area around the thiofuran ring was predicted as disfavored regions, the docked structure shows open space in the surrounding area.

When checking the electrostatic maps of the CoMFA (Fig. 6B) and CoMSIA models (Fig. 7B), it is noticed that both models provided similar predictions for two favored regions – one large area located around thiofuran ring, and the other area close to the sulfone group and phenyl ring. The CoMFA model predicted a disfavored area near the sulfone group, which has a small portion of overlap with the favored areas predicted by CoMSIA model. It is reasonable to have disfavored areas around phenyl ring, while the prediction of disfavored areas around sulfone group seems contradictory to common understanding that the interaction of the polar group (sulfone) with polar amino acids increases ligand binding affinity. The hydrophobic property map (Fig. 7C) from the CoMSIA model suggested two favored regions (cyan)-one area around the oxygen atom of ester and the other area located on one side of the phenyl ring nearby the triazole ring. Four disfavored regions (magenta) were identified as well by this analysis, which include areas located on the thiofuran ring, triazole ring, and the two sides of phenyl ring. Comparing the protein surface properties at the binding site (Fig. 8), the locations of three predicted hydrophobic-favored areas match the hydrophobic areas on the protein surface. Also, some hydrophilic areas are found on the protein surface map close to the corresponding locations of the disfavored hydrophobic areas on contour map of the CoMSIA model. Generally speaking, suggestions from the contour map agree with the general feature of the protein surface properties around the protein binding site. All these predictions demonstrated that both the CoMFA and CoMSIA model provide valuable insights into the factors critical for determining the ligand binding and binding affinity against cathepsin B.

In the docking simulation, shape fitting and electrostatic matching are the two most important forces to determine a preferable binding structure of a ligand. If studied ligands have similar electrostatic patterns and are able to fit into the binding site, they will bind in a similar position with the same orientation. If one ligand has an opposite electrostatic pattern comparing to the other ligands, it could ideally be docked in with an orientation which is turned

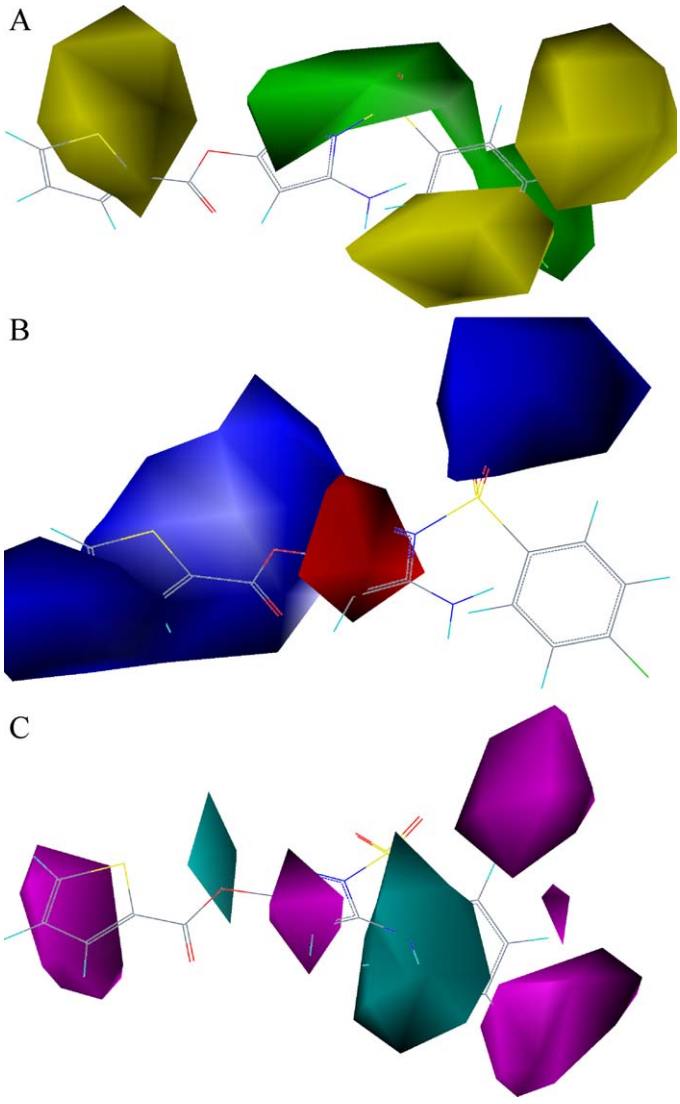

Fig. 7. The contour maps of CoMSIA results. (A) The steric map, the favored regions are colored in green and the disfavored regions are colored in yellow. (B) The electrostatic map, the favored regions are colored in blue and the disfavored regions are colored in red. (C) The hydrophobic map, the favored regions are colored in cyan and the disfavored regions are colored in magenta. The compound 26 is shown along with the map. (For interpretation of the references to color in the figure caption, the reader is referred to the web version of the article.)

180°. However, a ligand in such a case may be aligned with the same orientation as the template ligand with the template based atom fitting method. The alignment from docking simulation, however, seems to have an edge over the template-based atom-fitting alignment and produced nearly ideal results for such a situation. However, the effect of such ligands on the entire 3D QSAR model needs to be checked manually in case they may need to be treated as outliers.

To build CoMFA-like 3D QSAR, determination of "active" conformation and alignment are the two most important elements to determine the quality and success of a QSAR model. When applied to determining the conformation and alignment, docking simulations normally produces multi-conformations for a compound. A preferable conformation for a compound could be determined based on the knowledge of known bound structures, docking score, etc. in some case. But in other cases, it may be difficult to select a preferable conformation. In general, how to select the docked structure for a compound is tricky, challenging, and difficult, it sometimes means whether a successful model can be constructed.

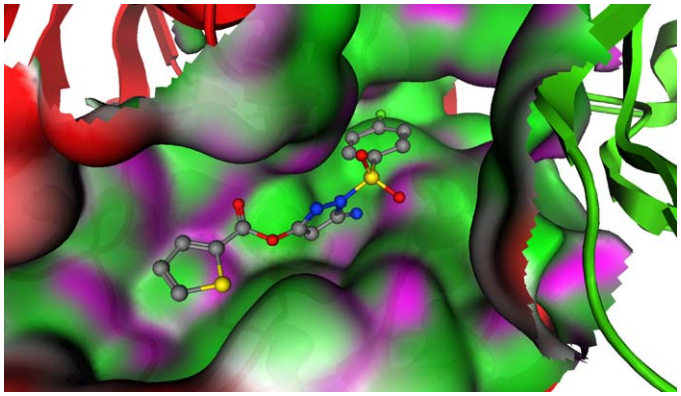

Fig. 8. The surface properties of the binding site represented by colors: magenta, hydrophilic; green, hydrophobic; red, exposed. The proteins are depicted with cartoon mode, the primary protein of the dimer on left is colored in red and the second one on right is colored in green. The compound 26 is shown along with the map. (For interpretation of the references to color in the figure caption, the reader is referred to the web version of the article.)

In this work, we introduced multi-conformation QSAR approach that utilizes multi-poses from docking simulation in QSAR model construction. The current approach allows gradual and step by step elimination of docking poses during the QSAR building process. It combines receptor-based docking simulation with ligand-based 3D QSAR (CoMFA and CoMSIA) by introducing the binding site properties into the ligand-based models, thus enhancing the prediction power of conventional QSAR models. The results of the statistical models demonstrated the power of using the current approach to determine conformation and alignment. These results provided insights into crucial structure features affecting ligand–receptor interactions and their binding affinities, thus demonstrating the advantage of the proposed approach for 3D QSAR modeling in drug design.

4. Conclusion

By combining docking simulations with CoMFA and CoMSIA analysis, the binding structures and quantitative structure–activity relationship (QSAR) of cathepsin B inhibitors have been studied. Robust 3D QSAR models were constructed with cross-validated correlation coefficients (q^2) and regression correlation coefficients (r^2) of 0.605 and 0.999 for CoMFA models, and 0.605 and 0.997 for CoMSIA models, respectively. Various methods were used to validate these studies, including LOO, cross-validation, Bootstrapping, and training-test set methods. In the last experiment, the activity of the tested compounds predicted by the models built from the training compounds was shown to be well in line with their experimental activity with a maximum error of 1 for CoMFA and CoMSIA models. All validations show that the models exhibited strong prediction for evaluating the inhibitory activity ($\log IC_{50}$) of the compounds as cathepsin B inhibitors. The strong correlation between the calculated activity and the experimental activity demonstrated the robustness of the models, and the feasibility and advantage of the computational approach in this study.

The contour maps predicted from CoMFA and CoMSIA analyses were discussed and compared with the surface property maps of the protein binding site. The areas on phenyl ring facing out of paper (Figs. 6A and 7A) are disfavored regions for bulky groups, while the areas on the other side of the phenyl ring are favored regions for bulky groups. It is also predicted that the area located around thiofuran ring and the area close to the sulfone group are favored electrostatic regions. The two analyses provided similar suggestions for favored and disfavored regions around the ligands

when considering steric factors. These suggestions generally match the amino acids allocation and distribution around docked ligand at the binding site. The major features of the electrostatic maps from the two analyses are also consistent. The hydrophobic map for favored and unfavored areas derived from the CoMSIA analysis roughly matches the hydrophobic and hydrophilic areas of the protein surface of the binding site. Such analysis based on the CoMFA and CoMSIA results may provide valuable information for understanding the interactions between small molecules and cathepsin B, and thus can provide guidance for modifying the identified small molecules, or in designing novel ones for optimized potency and target specificity.

One remaining challenge for the traditional 3D QSAR study is binding conformation determination. When experimental structure data is unavailable, docking simulation is often chosen as an alternative approach to obtain "active" conformation for studied compounds. However, many studies have shown that selecting a conformation for a compound from multi-docked poses is tricky and difficult, and is often the leading cause of failure to achieve good QSAR models. In this work, we introduced a multi-conformation QSAR approach in an attempt to tackle this challenge. Multi-docked poses were introduced into statistical analysis for QSAR model construction and the final conformation for a compound is identified by an iterative procedure that optimizes the conformation selection for the best correlation. The final model is constructed based on the optimized conformer data set. This approach for conformation selection contributed substantially to the successful construction of the suggested models, which may significantly increase the chance to derive a model with satisfactory results in general.

In a traditional QSAR model study, no receptor information is included, therefore suggestions from such model cannot guarantee to agree with the features of receptor binding site, which eventually determine the possibility and strength of a ligand binding and its biological consequence (activity). This work utilized receptor structure-based docking simulation to determine the "active" conformations of each ligand compound and the compound alignments. The success of the QSAR models demonstrated that the approach, by combining both receptor and ligand features, is superior to the conventional template atom-fitting QSAR approach. The models built in this study provide detailed and deep insights for understanding the individual physical chemical elements affecting ligand–receptor interactions and provide valuable suggestions for novel inhibitor design and further chemical optimization. This would help to generate a library of cathepsin inhibitors and identify lead compounds with improved inhibition potency, target selectivity and specificity. This work will doubtlessly help the chemical optimization for this set of compounds, as well as facilitate the design and development of *novo* cathepsin B inhibitors as drugs to cure human diseases where the members of the papain superfamilies are known to play important roles.

Acknowledgements

This research was supported by the Intramural Research Program of the NIH, NLM. The authors thank NIH Fells Editorial Board (FEB) for assistance on the manuscript reviewing and the developers of Pymol software for sharing the program to prepare the molecular figures used in the paper.

References

[1] A.J. Barrett, H. Kirschke, B. Cathepsin, H. Cathepsin, L. cathepsin, Methods Enzymol. 80 (Pt C) (1981) 535–561.
[2] H.A. Chapman Jr., J.S. Munger, G.P. Shi, The role of thiol proteases in tissue injury and remodeling , Am. J. Respir. Crit. Care Med. 150 (1994) S155–S159.
[3] H.K. Rooprai, D. McCormick, Proteases and their inhibitors in human brain tumours: a review , Anticancer Res. 17 (1997) 4151–4162.
[4] L.R. Roberts, P.N. Adjei, G.J. Gores, Cathepsins as effector proteases in hepatocyte apoptosis , Cell Biochem. Biophys. 30 (1999) 71–88.
[5] I. Giusti, S. D'Ascenzo, D. Millimaggi, G. Taraboletti, G. Carta, N. Franceschini, et al., Cathepsin B mediates the pH-dependent proinvasive activity of tumor-shed microvesicles , Neoplasia 10 (2008) 481–488.
[6] E. Gounaris, C.H. Tung, C. Restaino, R. Maehr, R. Kohler, J.A. Joyce, et al., Live imaging of cysteine-cathepsin activity reveals dynamics of focal inflammation, angiogenesis, and polyp growth , PLoS One 3 (2008) e2916.
[7] S.D. Ha, A. Martins, K. Khazaie, J. Han, B.M. Chan, S.O. Kim, Cathepsin B is involved in the trafficking of TNF-alpha-containing vesicles to the plasma membrane in macrophages , J. Immunol. 181 (2008) 690–697.
[8] A. Haque, N.L. Banik, S.K. Ray, New insights into the roles of endolysosomal cathepsins in the pathogenesis of Alzheimer's disease: cathepsin inhibitors as potential therapeutics , CNS Neurol. Disord. Drug Targets 7 (2008) 270–277.
[9] E. Sandes, C. Lodillinsky, R. Cwirenbaum, C. Arguelles, A. Casabe, A.M. Eijan, Cathepsin B is involved in the apoptosis intrinsic pathway induced by Bacillus Calmette-Guerin in transitional cancer cell lines , Int. J. Mol. Med. 20 (2007) 823–828.
[10] S.P. Lutgens, K.B. Cleutjens, M.J. Daemen, S. Heeneman, Cathepsin cysteine proteases in cardiovascular disease , FASEB J. 21 (2007) 3029–3041.
[11] V. Hook, M. Kindy, G. Hook, Cysteine protease inhibitors effectively reduce in vivo levels of brain beta-amyloid related to Alzheimer's disease , Biol. Chem. 388 (2007) 247–252.
[12] L.S. Downs Jr., P.H. Lima, R.L. Bliss, C.H. Blomquist, Cathepsins B and D activity and activity ratios in normal ovaries, benign ovarian neoplasms, and epithelial ovarian cancer , J. Soc. Gynecol. Invest. 12 (2005) 539–544.
[13] O. Vasiljeva, M. Korovin, M. Gajda, H. Brodoefel, L. Bojic, A. Kruger, et al., Reduced tumour cell proliferation and delayed development of high-grade mammary carcinomas in cathepsin B-deficient mice , Oncogene 27 (2008) 4191–4199.
[14] O. Vasiljeva, A. Papazoglou, A. Kruger, H. Brodoefel, M. Korovin, J. Deussing, et al., Tumor cell-derived and macrophage-derived cathepsin B promotes progression and lung metastasis of mammary cancer , Cancer Res. 66 (2006) 5242–5250.
[15] D.T. Jane, L. Morvay, L. Dasilva, D. Cavallo-Medved, B.F. Sloane, M.J. Dufresne, Cathepsin, B localizes to plasma membrane caveolae of differentiating myoblasts and is secreted in an active form at physiological pH , Biol. Chem. 387 (2006) 223–234.
[16] B. Akache, D. Grimm, X. Shen, S. Fuess, S.R. Yant, D.S. Glazer, et al., A two-hybrid screen identifies cathepsins B and L as uncoating factors for adeno-associated virus 2 and 8 , Mol. Ther. 15 (2007) 330–339.
[17] K. Chandran, N.J. Sullivan, U. Felbor, S.P. Whelan, J.M. Cunningham, Endosomal proteolysis of the ebola virus glycoprotein is necessary for infection science , Science 308 (2005) 1643–1645.
[18] G. Simmons, D.N. Gosalia, A.J. Rennekamp, J.D. Reeves, S.L. Diamond, P. Bates, Inhibitors of cathepsin L prevent severe acute respiratory syndrome coronavirus entry , Proc. Natl. Acad. Sci. U.S.A. 102 (2005) 11876–11881.
[19] V.Y. Hook, M. Kindy, G. Hook, Inhibitors of cathepsin B improve memory and reduce beta-amyloid in transgenic Alzheimer disease mice expressing the wild-type, but not the Swedish mutant, beta-secretase site of the amyloid precursor protein , J. Biol. Chem. 283 (2008) 7745–7753.
[20] M. Hosokawa, K. Kashiwaya, H. Eguchi, H. Ohigashi, O. Ishikawa, M. Furihata, et al., Over-expression of cysteine proteinase inhibitor cystatin 6 promotes pancreatic cancer growth , Cancer Sci. 99 (2008) 1626–1632.
[21] B.S. Parker, D.R. Ciocca, B.N. Bidwell, F.E. Gago, M.A. Fanelli, J. George, et al., Primary tumour expression of the cysteine cathepsin inhibitor Stefin A inhibits distant metastasis in breast cancer , J. Pathol. 214 (2008) 337–346.
[22] A. Yamamoto, K. Tomoo, T. Hara, M. Murata, K. Kitamura, T. Ishida, Substrate specificity of bovine cathepsin B and its inhibition by CA074, based on crystal structure refinement of the complex , J. Biochem. 127 (2000) 635–643.
[23] P.D. Greenspan, K.L. Clark, R.A. Tommasi, S.D. Cowen, L.W. McQuire, D.L. Farley, et al., Identification of dipeptidyl nitriles as potent and selective inhibitors of cathepsin B through structure-based drug design , J. Med. Chem. 44 (2001) 4524–4534.
[24] H.H. Otto, T. Schirmeister, Cysteine Proteases and their Inhibitors, vol. 97, 1997, pp. 133–172.
[25] M. Mladenovic, K. Ansorg, R.F. Fink, W. Thiel, T. Schirmeister, B. Engels, Atomistic insights into the inhibition of cysteine proteases: first QM/MM calculations clarifying the stereoselectivity of epoxide-based inhibitors , J. Phys. Chem. B 112 (2008) 11798–11808.
[26] I. Redzynia, A. Ljunggren, M. Abrahamson, J.S. Mort, J.C. Krupa, M. Jaskolski, et al., Displacement of the occluding loop by the parasite protein, chagasin, results in efficient inhibition of human cathepsin B , J. Biol. Chem. 283 (2008) 22815–22825.
[27] D. Watanabe, A. Yamamoto, K. Tomoo, K. Matsumoto, M. Murata, K. Kitamura, et al., Quantitative evaluation of each catalytic subsite of cathepsin B for inhibitory activity based on inhibitory activity-binding mode relationship of epoxysuccinyl inhibitors by X-ray crystal structure analyses of complexes , J. Mol. Biol. 362 (2006) 979–993.
[28] P. Markt, C. McGoohan, B. Walker, J. Kirchmair, C. Feldmann, G.D. Martino, et al., Discovery of novel cathepsin S inhibitors by pharmacophore-based virtual high-throughput screening , J. Chem. Inf. Model. 48 (2008) 1693–1705.
[29] M.P. Beavers, M.C. Myers, P.P. Shah, J.E. Purvis, S.L. Diamond, B.S. Cooperman, et al., Molecular docking of cathepsin L inhibitors in the binding site of papain , J. Chem. Inf. Model. 48 (2008) 1464–1472.

[30] Z. Zhou, Y. Wang, S.H. Bryant, Computational analysis of the cathepsin B inhibitors activities through LR-MMPBSA binding affinity calculation based on docked complex , J. Comput. Chem. 30 (2009) 2165–2175.
[31] Z. Zhou, Y. Wang, S.H. Bryant, QSAR models for predicting cathepsin B inhibition by small molecules—continuous and binary QSAR models to classify cathepsin B inhibition activities of small molecules , J. Mol. Graph. Model. 28 (2010) 4.
[32] A.J. Hopfinger, S. Wang, J.S. Tokarski, B. Jin, M. Albuquerque, P.J. Madhav, et al., Construction of 3D-QSAR models using the 4D-QSAR analysis formalism , J. Am. Chem. Soc. 119 (1997) 10509–10524.
[33] M.C. Myers, A.D. Napper, N. Motlekar, P.P. Shah, C.-H. Chiu, M.P. Beavers, et al., Identification and characterization of 3-substituted pyrazolyl esters as alternate substrates for cathepsin B: the confounding effects of DTT and cysteine in biological assays , Bioorg. Med. Chem. Lett. 17 (2007) 4761–4766.
[34] W.L. Jorgensen, J. Tirado-Rives, The OPLS potential function for proteins energy minimizations for crystals of cyclic peptides and crambin , J. Am. Chem. Soc. 110 (1988) 1657–1666.
[35] W. Damm, A. Frontera, J. Tirado-Rives, W.L. Jorgensen, OPLS all-atom force field for carbohydrates , J. Comput. Chem. 18 (1997) 1955–1970.
[36] R.C. Rizzo, W.L. Jorgensen, OPLS all-atom model for amines: resolution of the amine hydration problem , J. Am. Chem. Soc. 121 (1999) 4827–4836.
[37] W.L. Jorgensen, D.S. Maxwell, J. Tirado-Rives, Development, Testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids , J. Am. Chem. Soc. 118 (1996) 11225–11236.
[38] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, et al., Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy , J. Med. Chem. 47 (2004) 1739–1749.
[39] T.A. Halgren, R.B. Murphy, R.A. Friesner, H.S. Beard, L.L. Frye, W.T. Pollard, et al., Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening , J. Med. Chem. 47 (2004) 1750–1759.
[40] Z. Zhou, M. Bates, J.D. Madura, Structure modeling, ligand binding, and binding affinity calculation (LR-MM-PBSA) of human heparanase for inhibition and drug design , Proteins: Struct. Funct. Bioinf. 65 (2006) 580–592.
[41] Z. Zhou, M. Khaliq, J.-E. Suk, C. Patkar, L. Li, R.J. Kuhn, et al., Antiviral compounds discovered by virtual screening of small-molecule libraries against dengue virus E protein , ACS Chem. Biol. 3 (2008) 765–775.
[42] Z. Zhou, J.D. Madura, CoMFA 3D-QSAR analysis of HIV-1 RT nonnucleoside inhibitors, TIBO derivatives based on docking conformation and alignment , J. Chem. Inf. Comput. Sci. 44 (2004) 2167–2178.