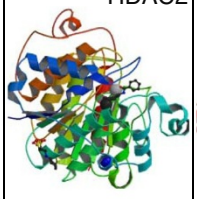
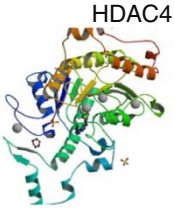
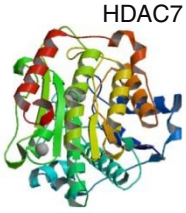
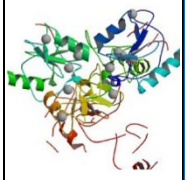
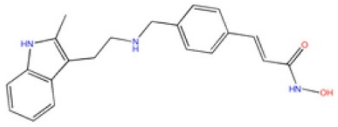
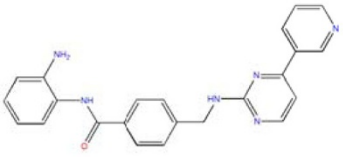
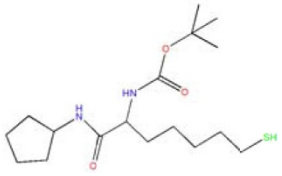
BMC Bioinformatics

HDACs / Ligands	Class I		Class IIa		Class IIb
	HDAC2 	HDAC8	HDAC4 	HDAC7 	HDAC6 
	0.742	0.391	0.524	0.347	0.996
	0.296	-0.946	-0.557	-1.018	-0.963
	-0.405	-0.731	0.137	0.159	0.010

Screening of selective histone deacetylase inhibitors by proteochemometric modeling

Wu *et al.*

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Screening of selective histone deacetylase inhibitors by proteochemometric modeling

Dingfeng Wu[1†], Qi Huang[1†], Yida Zhang[1], Qingchen Zhang[1], Qi Liu[1], Jun Gao[1,2], Zhiwei Cao[1*] and Ruixin Zhu[1,3,4*]

Abstract

Background: Histone deacetylase (HDAC) is a novel target for the treatment of cancer and it can be classified into three classes, i.e., classes I, II, and IV. The inhibitors selectively targeting individual HDAC have been proved to be the better candidate antitumor drugs. To screen selective HDAC inhibitors, several proteochemometric (PCM) models based on different combinations of three kinds of protein descriptors, two kinds of ligand descriptors and multiplication cross-terms were constructed in our study.

Results: The results show that structure similarity descriptors are better than sequence similarity descriptors and geometry descriptors in the characterization of HDACs. Furthermore, the predictive ability was not improved by introducing the cross-terms in our models. Finally, a best PCM model based on protein structure similarity descriptors and 32-dimensional general descriptors was derived (R^2 = 0.9897, Q^2_{test} = 0.7542), which shows a powerful ability to screen selective HDAC inhibitors.

Conclusions: Our best model not only predict the activities of inhibitors for each HDAC isoform, but also screen and distinguish class-selective inhibitors and even more isoform-selective inhibitors, thus it provides a potential way to discover or design novel candidate antitumor drugs with reduced side effect.

Keywords: Histone deacetylases inhibitors, Proteochemometric, Selective inhibitors

Background

All over the world, tumor is the second incurable disease only to cardiovascular disease. A wide range of proteins are found to be related to tumor formation and metastasis. However, only proteins with widespread biological significance for the tumor cells growth regulation are most possible to be the targets of broad-spectrum low-toxic antitumor drugs. In recent studies, histone deacetylases (HDACs) are proved to be novel epigenetic targets for the treatment of cancer [1-3]. Histone deacetylase inhibitors (HDACi) have extensively demonstrated the antitumor efficacy *in vitro* and *in vivo*. Therefore, the related study of HDACi has become one of the most important research fields of the antitumor drugs, especially during the coming area of epigenetics.

Histone deacetylases comprise a superfamily of 18 genes which is divided into two families and four classes in eukaryotic cells. Classes I, II, and IV consist of 11 family members, which are referred to as "classical" HDACs, whereas the 7 class III members are called "sirtuins" [4]. Classical HDACs which require Zn^{2+} as a cofactor for their deacetylase activity are a promising novel class of anticancer drug targets that can be inhibited by Zn^{2+} chelating compounds such as hydroxamic acids. In contrast, these compounds are not active against sirtuins as these class III enzymes have a different mechanism of action in requiring NAD^+ as an essential cofactor [5]. Recent researches indicate that sirtuins are linked to aging as well as metabolic and neurodegenerative diseases [6].

Classical HDACs are classified based on their homology to yeast proteins. HDACs 1, 2, 3, and 8 which belong to Class I have homologies to yeast RPD3, and they are located within the nucleus. HDACs 4, 5, 6, 7, 9, and 10 which belong to Class II have homologies to yeast HDA1 and located in both the nucleus and the cytoplasm. It should be noted that Class II HDACs can be further subdivided based on their sequence homolog and domain

* Correspondence: zwcao@tongji.edu.cn; rxzhu@tongji.edu.cn
†Equal contributors
[1]School of Life Sciences and Technology, Tongji University, Shanghai 200092, P.R. China
[3]Institute for Advanced Study of Translational Medicine, Tongji University, Shanghai 200092, P.R. China
Full list of author information is available at the end of the article

BioMed Central

© 2012 Wu et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

organization, *i.e.* Class IIa, which include HDACs 4, 5, 7, and 9 containing an N-terminal extension with regulatory function, and Class IIb, which include HDACs 6 and 10 containing two catalytic domains. HDAC 11 is categorized into class IV with conserved residues in its catalytic center that are shared by both classes I and II HDACs. The classification of classical HDACs is summarized in Table 1.

Histone deacetylase inhibitors (HDACi) that act on 11 zinc-dependent HDAC isozymes generally possess a zinc-binding group which coordinates the zinc ion in the active site, a cap substructure that interacts with amino acids at the entrance of the N-acetylated lysine binding channel, and a linker connecting the cap and the zinc-binding group at a proper distance [18]. HDACi can be categorized into four subtypes based on their chemical structures: (1) short chain fatty acid; (2) hydroxamic acid; (3) benzamides; and (4) cyclic peptides. Since HDACi do not inhibit all HDAC isoforms to the same extent, they can be categorized into pan-HDAC inhibitors and selective HDAC inhibitors including class I-specific inhibitors, class II-specific inhibitors, and class IV-specific inhibitors. Currently, many HDAC inhibitors have already been tested in clinical trials and shown certain antitumor or other biological activity. However, some HDAC inhibitors, especially pan-inhibitors, indicate serious side effects, such as fatigue, nausea, anorexia, diarrhea, thrombus formation, thrombocytopenia, neutropenia, anemia, myalgia, hypokalemia, hypophosphatemia, *etc.* [3]. Thus, HDAC inhibitors are possible to greatly improve the efficacy and reduce the certain toxicities only when they target the most relevant HDAC isoform rather than multiple ones. Consequently, it should be useful to discover or design novel antitumor drugs with fewer side effects when one method can analyze the interaction of inhibitors against multiple HDACs with further sorting out isoform- or class-specific inhibitors.

Table 1 "classical" HDACs

HDAC		Localization	Chromosomal site	References
Class I (RPD3 homologue)				
	HDAC1	Nucleus	1p34.1	[7]
	HDAC2	Nucleus	6p21	[8]
	HDAC3	Nucleus	5p31	[9]
	HDAC8	Nucleus	Xq13	[10]
Class II (HDA1 homologue)				
IIa	HDAC4	Nuc/Cyt	2q372	[11]
	HDAC5	Nuc/Cyt	17q21	[12]
	HDAC7	Nuc/Cyt	12q13	[13]
	HDAC9	Nuc/Cyt	7p21-p15	[14]
IIb	HDAC6	Mainly Cyt	Xp11.22-33	[15]
	HDAC10	Mainly Cyt	22q13.31-33	[16]
Class IV				
	HDAC11	Nuc/Cyt	3p25.2	[17]

As for *in silico* drug discovery, there are many methods available such as molecular docking [19,20], pharmacophore models, quantitative structure-activity relationship (QSAR) [21-23], protein-ligand interaction fingerprint-based screening [24,25] and others [26-29]. QSAR is a widely applied computational method for predicting chemical compounds' interactions with a single target protein. However, when thousands of chemical compounds interacted with 11 different HDAC isoforms, 11 separate QSAR models for each HDAC isoform are needed to create, which is quite complicated and time consuming. In addition, these separate models cannot extended to predict inhibitions of new HDACs [30]. Therefore, a new method should be proposed to predict cross-interactions of chemical compounds to multi-HDAC isoforms simultaneously.

More recently, proteochemometric (PCM) modeling has been widely used to study the cross-interactions between a series of compounds and a series of proteins. In this area Maris Lapinsh et.al studied melanocortin chimeric receptors using partial least-squares projections (PLS) to deduce PCM models [31,32]; Hanna Geppert et.al derived PCM models of eleven proteases from four different protease families by support vector machine [33]; Ilona Mandrika and Maris Lapinsh et.al applied PLS to model interactions of HIV mutants [30,34] and antibodies [35]. Contrary to traditional QSAR, PCM is based on the similarity of a group of ligands together with that of a group of targets [36]. Consequently, PCM can integrate several separate QSAR models into a global one. With the global PCM model in hand, we can study the cross interactions of all the ligands with all the targets in the data set or even outside the data set. By predicting the affinity for each ligand-target pair, PCM models can describe the specific interaction between a ligand and a target and discriminate the interaction strength between different ligand-target pairs. Therefore, in our study PCM models were applied to study the cross-interactions of a series of HDAC inhibitors to five HDAC isoforms, *i.e.*, HDAC2, HDAC4, HDAC6, HDAC7, and HDAC8.

Results and discussion

Proteochemometric modeling

In our study, 18 proteochemometric models were created from training set with combinations of different descriptors. As shown in Table 2, goodness-of-fits (R^2s) of all models were higher than 0.9619 and their cross validation coefficients Q^2_{cv} ranged from 0.5734 to 0.7162. The model derived based on P1 and GD showed to be the best model with the highest predictive ability (Q^2_{cv} = 0.7162 and Q^2_{test} = 0.7542). Accordingly P1-GD model was used in the subsequent analysis.

Table 2 Goodness-of-fit (R^2) and predictive ability (Q^2_{cv}, Q^2_{test}) of the obtained 18 models

Model	R^2	Q^2_{cv}	Q^2_{test}	Model	R^2	Q^2_{cv}	Q^2_{test}
P0-DLI[a]	0.9616	0.6564	0.7292	P1-GD-C[b]	0.9909	0.6732	0.7191
P0-GD	0.9895	0.6960	0.7331	P2-DLI-C	0.9883	0.6264	0.6519
P1-DLI	0.9619	0.6757	0.7427	P2-GD-C	0.9917	0.5734	0.6215
P1-GD	0.9897	0.7162	0.7542	C(P0,DLI)[c]	0.9860	0.6484	0.6941
P2-DLI	0.9614	0.6521	0.7272	C(P0,GD)	0.9914	0.6302	0.6772
P2-GD	0.9894	0.6858	0.7268	C(P1,DLI)	0.9811	0.6686	0.7227
P0-DLI-C	0.9871	0.6462	0.6944	C(P1,GD)	0.9904	0.6625	0.7190
P0-GD-C	0.9916	0.6319	0.6759	C(P2,DLI)	0.9898	0.6505	0.6067
P1-DLI-C	0.9846	0.6626	0.7251	C(P2,GD)	0.9941	0.5967	0.5597

[a] P0-DLI means this model is based on protein descriptor P0 and ligand descriptor DLI.
[b] P1-GD-C means this model is based on protein descriptor P1, ligand descriptor GD and cross-terms.
[c] C(P0,DLI) means this model is based on only cross-terms of P0 and DLI.

Table 3 R^2 and Q^2_{test} of 18 models grouped for comparing three protein descriptors ability

Model	R^2	Q^2_{test}	Model	R^2	Q^2_{test}
Group 1			Group 2		
P0-DLI	0.9616	0.7292	P0-GD-C	0.9916	0.6759
P1-DLI	0.9619	**0.7427**	P1-GD-C	0.9909	**0.7191**
P2-DLI	0.9614	0.7272	P2-GD-C	0.9917	0.6215
Group 3			Group 4		
P0-GD	0.9895	0.7331	C(P0,DLI)	0.9860	0.6941
P1-GD	0.9897	**0.7542**	C(P1,DLI)	0.9811	**0.7227**
P2-GD	0.9894	0.7268	C(P2,DLI)	0.9898	0.6067
Group 5			Group 6		
P0-DLI-C	0.9871	0.6944	C(P0,GD)	0.9914	0.6772
P1-DLI-C	0.9846	**0.7251**	C(P1,GD)	0.9904	**0.7190**
P2-DLI-C	0.9883	0.6519	C(P2,GD)	0.9941	0.5597

The highest Q^2_{test}s of every group are highlighted.

P0 vs P1 vs P2

Three protein descriptors, *i.e.*, sequence similarity descriptor (P0), structure similarity descriptor (P1) and geometry descriptor (P2), were used to describe HDACs in our study. Sequence similarity descriptor is based on the sequence identities of HDACs, while structure similarity descriptor and geometry descriptor characterize HDACs based on their 3D-structures. Protein descriptors are different from ligand descriptors since proteins have larger molecule structures to describe. If available, proteins are likely to be described on the basis of crystal structures. Protein structure similarity descriptor was calculated by protein 3D-struture alignment with more sufficient information considered. Contrary to P1, P0 only characterizes protein based on sequence alignment, and may lose certain 3D information of proteins. Not surprisingly, models derived from P1 showed a better predictive ability than those of P0 (Table 3). In addition, although P2 is also derived based on 3D-structure, it only measures bond length, bond angle and dihedral angle statistically without much of the detailed information of proteins, thus it is not sufficient to characterize proteins comprehensively. As a result, we also found that models based on geometry descriptor obtained the worst predictive ability (Q^2_{test} of models based on P2 in every group is the lowest) compared to the others.

GD vs DLI

Similar to protein descriptors, two typical kinds of ligand descriptors, *i.e.*, General Descriptor (GD) and Drug-Like Index (DLI) were applied. Our result indicates that there was no significant difference between Q^2 values of models based on GD and DLI (Table 4), with p-value bigger than 0.1 by paired *t*-test.

It should be noted that there are a large number of different descriptors available for ligands, and there is no optimal one suitable for all the data sets. Therefore, it is wise to try several different descriptors to identify the optimal one in a particular scenario [37]. In our study, we used two different ligand descriptors, GD and DLI to create PCM models. These two kinds of descriptors characterize physical properties and topological indices of ligands respectively. For our particular data set, there was no statistically difference in predictive ability between these two ligand descriptors.

Model performance with or without cross-terms

A multiplied cross-term was used in our models and it was shown to be helpless in the improvement of the predictive ability of PCM models. The Q^2_{test} of models with cross-terms is lower than that without cross-terms in every group (Table 5).

Although cross-terms are intended to describe the properties of the interface between ligand and protein, there is still no good descriptor for the representation of local receptor-ligand interfaces [37], which may possibly result in the worse performance of the multiplied cross-term in our PCM models. Recently, a new Protein-Ligand interaction fingerprint was derived for *in silico* screening [24,25]. This interaction fingerprint is a local descriptor to represent the interfaces of receptor-ligand and proved to be a good candidate cross-term in PCM. Theoretically, it should achieve better performance if the crystal complex structure exists. However, since there is no crystal structure available for most of the receptor-ligand pairs in our data set, thousands of complex structures have to be produced by molecular docking to apply interaction fingerprint, which may result in biases. Therefore, the interaction fingerprint was not adopted in our study.

Table 4 Q^2_{test} of 18 models grouped by ligand descriptors

X	P0-X	P1-X	P2-X	P0-X-C	P1-X-C	P2-X-C	C(P0,X)	C(P1,X)	P(P2,X)
GD	0.7331	0.7542	0.7268	0.6759	0.7191	0.6215	0.6772	0.7190	0.5597
DLI	0.7292	0.7427	0.7272	0.6944	0.7251	0.6519	0.6941	0.7227	0.6067

Paired t test: t = 1.746.

Selective ability of proteochemometric model

In our study, we aimed to exploit an effective method to screen selective HDAC inhibitors which has selective activity on a single or a specific class of HDAC isoforms. For this purpose, proteochemometrics was applied to analyze the interaction strength of inhibitors against multiple HDACs, and then select out isoform-specific, class-specific as well as pan inhibitors. To verify the performance of the derived PCM models, an external validation of ten inhibitors was carried out to predict affinity with the best model (P1-GD model). The predicted values are compared with the corresponding experimental ones as shown in Table 6.

Among the ten inhibitors for external validation, TSA, SAHA, LBH589 and PXD-101 are reported as pan-HDAC inhibitors and almost all their predicted affinity values are high for all the HDAC isoforms in our test (e.g. LBH HDAC2 0.742, HDAC8 0.391, HDAC4 0.524, HDAC7 0.347, HDAC6 0.996). In addition, MGCD0103, FK228 and Apicidin are reported as class I-specific inhibitors and our results also indicated that the predicted values for class I HDACs are higher than those of others (e.g. Apicidin HDAC2 0.238, HDAC8 0.096, HDAC4 -0.501, HDAC7 -0.176, HDAC6 -0.120). Finally APHA, Tubacin and NCT-10a are reported as class II-specific inhibitors and our results are consistent with the validation data that their predicted values are higher for class II HDACs (e.g. NCT-10a HDAC2 -0.405, HDAC8 -0.731, HDAC4 0.137, HDAC7 0.159, HDAC6 0.010).

Table 5 R^2 and Q^2_{test} of 12 models grouped by with- or without- cross-terms

Model	R^2	Q^2_{test}	Model	R^2	Q^2_{test}
Group 1			Group 2		
P0-DLI	0.9616	**0.7292**	P1-GD	0.9897	**0.7542**
P0-DLI-C	0.9871	0.6944	P1-GD-C	0.9909	0.7191
Group 3			Group 4		
P0-GD	0.9895	**0.7331**	P2-DLI	0.9614	**0.7272**
P0-GD-C	0.9916	0.6759	P2-DLI-C	0.9883	0.6519
Group 5		Group 6			
P1-DLI	0.9619	**0.7427**	P2-GD	0.9894	**0.7268**
P1-DLI-C	0.9846	0.7251	P2-GD-C	0.9917	0.6215

The highest Q^2_{test}s of every group are highlighted.

As a conclusion, our best PCM model performs well in screening selective HDAC inhibitors and distinguishing pan-HDAC inhibitors, class I-specific inhibitors and class II-specific inhibitors successfully. Therefore, this model can be further used to screen class-selective inhibitors as well as isoform-selective inhibitors of HDACs with fewer side effects.

Conclusion

Although more and more HDAC inhibitors have been identified to date, the number of class-selective inhibitors or isoform-selective inhibitors is insufficient. Thus, it is important to find these selective inhibitors which are candidate therapeutic agents for tumor with reduced side effects. In this study, proteochemometric models were derived to analyze the inhibitory activity of 1275 compounds with 5 HDAC isoforms simultaneously. Among these models, the best one, P1-GD model, was highly predictive (Q^2_{test} = 0.7542) and presented powerful ability to distinguish selective HDAC inhibitors from the pan ones. As a conclusion, proteochemometric modeling proves to be a suitable methodology for the prediction of inhibitor interactions with HDAC isoforms. Our study also indicates that the obtained optimal model can be potentially used for designing candidate antitumor drugs which can selectively target on a single HDAC or a specific class of HDAC isoforms.

Methods

Data set

To describe proteins more efficiently, five HDAC isoforms with known crystal structures were selected (Table 7). Among these isoforms, HDAC2 and HDAC8 are Class I HDACs; HDAC4, HDAC6, and HDAC7 are Class II HDACs, and more specifically, HDAC4 and HDAC7 belong to Class IIa; HDAC6 belongs to Class IIb.

The half maximal inhibitory concentration (IC50) values for 1443 chemical compounds (Additional file 1: Table S4) interacting with these HDAC isoforms were collected from the Binding Database (BindingDB, http://www.bindingdb.org). After filtration, the data set was reduced to 1275 compound-HDAC pairs with IC50 values, and it contained 215 pairs for HDAC2, 197 for HDAC4, 531 for HDAC6, 46 for HDAC7, and 286 for HDAC8 respectively (Table 8).

Table 6 The activity data and P0-GD model predict affinity data of ten HDAC inhibitors [a]

Class I		Class IIa		Class IIb
HDAC2	**HDAC8**	**HDAC4**	**HDAC7**	**HDAC6**
Pan-HDAC inhibitors				
TSA				
S	**W**	**S**	**S**	**S**
−0.720	***0.342***	***1.027***	***0.660***	***0.087***
SAHA				
S	**W**	**S**	**S**	**S**
−0.464	−1.092	***0.687***	***1.031***	−0.094
Panbinostat(LBH589)				
S	**W**	**S**	**S**	**W**
0.742	***0.391***	***0.524***	***0.347***	***0.996***
Belinostat(PXD-101)				
S	**W**	**S**	**S**	**S**
0.327	−0.330	***1.183***	***0.643***	***1.339***
Class I-specific inhibitors				
MGCD0103				
S	N	N	N	N
0.296	−0.946	−0.557	−1.018	−0.963
depsipeptide(FK228)				
S	nd	**W**	nd	N
0.954	−0.095	***0.687***	***0.438***	−0.167
Apicidin				
S	**W**	N	N	N
0.238	***0.096***	−0.501	−0.176	−0.120
Class II-specific inhibitors				
APHA				
nd	nd	**S**	nd	nd
−0.196	−0.089	−0.204	−0.194	***0.271***
Tubacin				
nd	nd	nd	nd	**S**
0.148	−0.687	−0.293	−0.301	***1.414***
NCT-10a				
nd	nd	**W**	nd	**S**
−0.405	−0.731	***0.137***	***0.159***	***0.010***

[a] S, W,N and nd is the experimental affinity of inhibitors and the numerical number is the predicted affinity data. Those predicted values larger than 0 are supposed to have inhibition with highlighted.
S strong inhibition.
W weak inhibition.
N no inhibition.
nd no data pubilshed.

The distribution of data set for every HDAC isoform is unbalanced. Therefore, we divided the data set into a training set (65%) and a test set (35%) by stratified sampling [38] (Additional file 2: Table S1, Additional file 3: Table S2).

Table 7 HDACs' sequences and 3D structures from NCBI and PDB

Protein	PDB entry	NCBI entry	Length(aa)	Class
HDAC2	3MAX	NP_001518	488	I
HDAC4	2VQJ	NP_006028	1084	IIa
HDAC6	3C5K	NP_006035	1215	IIb
HDAC7	3C0Z	NP_056216	991	IIa
HDAC8	1 T69	NP_060956	377	I
HDAC1	nd[a]	NP_004955	482	I
HDAC3	nd	NP_003874	428	I
HDAC5	nd	NP_005465	1122	IIa
HDAC9	nd	NP_478056	1011	IIa
HDAC10	nd	NP_114408	669	IIb
HDAC11	nd	NP_079103	347	IV

[a] nd means no data published.

Description of proteins

Three different sets of descriptors were used to characterize the five HDAC isoforms, *i.e.* sequence similarity descriptor (P0) [32], structure similarity descriptor (P1) and geometry descriptor (P2).

Sequence similarity descriptor

The amino acid sequences of all the HDACs were retrieved from NCBI (the entries are listed in Table 7). EMBOSS [39,40] was used to calculate sequence identities of the five selected HDAC isoforms with all the HDAC isoforms. Finally we obtained 11 sequence similarity descriptors (Table 9).

Structure similarity descriptor

This descriptor extends protein sequence alignment to structure alignment based on sequence similarity descriptor. By pairwise structure alignment using Protein Comparison Tool [41], we calculated pairwise structure identities of the five selected proteins and obtained five descriptors (Table 10).

Geometry descriptor

Protein contains various bonds like C-N, C-O, C-N-CA, and CA-C-O *etc*. By measuring the various bond

Table 8 The distribution of binding affinity IC50 data

Type	Total	Training set	Test set
HDAC2/ligands	215	139	76
HDAC4/ligands	197	128	69
HDAC6/ligands	531	345	186
HDAC7/ligands	46	29	17
HDAC8/ligands	286	186	100
Total	1275	827	448

Table 9 11 sequence similarity descriptors of HDAC2, 4, 6, 7 and 8

	Class I				Class IIa				Class IIb		Class IV
	HDAC1	HDAC2	HDAC3	HDAC8	HDAC4	HDAC5	HDAC7	HDAC9	HDAC6	HDAC10	HDAC11
Class I											
HDAC2	85.1	100.0	51.9	30.7	10.2	9.6	9.7	9.9	9.4	14.2	18.9
HDAC8	30.8	30.7	34.4	100.0	9.0	8.4	9.9	10.8	8.1	13.6	21.4
Class IIa											
HDAC4	11.1	10.2	9.5	9.0	100.0	58.4	46.9	54.3	20.6	11.1	8.6
HDAC7	10.0	9.7	10.5	9.9	46.9	40.5	100.0	39.7	19.1	13.0	9.6
Class IIb											
HDAC6	9.5	9.4	7.7	8.1	20.6	17.2	19.1	16.6	100.0	23.4	7.4

For all possible HDACs pairs, sequence identities (in %) are reported.

length, bond angle and dihedral angle [42], 30 protein Geometry descriptors were obtained for each HDAC protein (Additional file 4: Table S3).

Description of inhibitors

In our study, the HDAC inhibitors were represented by two kinds of feature space, *i.e.* 32-dimensional General Descriptors (GD) and 28-dimensional Drug-Like Index (DLI). These descriptors are widely applied to the construction of QSAR models. For general descriptors, they include atomic contributions to van der waals surface area, log P (octanol/water), molar refractivity, and partial charge. GD characterize physical properties and describe organic compounds in boiling point, vapor pressure, free energy of salvation in water, solubility in water, thrombin/trysin/factor Xa activity, blood–brain barrier permeability, and compound classification *etc.* [43]. On the other hand, DLI characterize simple topological indices of compounds and measure the hierarchy of drug structures in terms of rings, links, and molecular frameworks [44].

Protein-inhibitor cross-terms

Evidently, ligand-receptor recognition can only be partially explained by linear combinations of ligand and receptor descriptors. In reality, protein-ligand interactions are governed by complex processes that depend on the complementarity of the properties of the interacting entities. In PCM, this is accounted for by protein-inhibitor cross-terms [31,36], which in the simplest case is obtained by multiplication of mean centered descriptors of proteins and inhibitors. Therefore, we obtained 11 × 32 = 352, 5 × 32 = 160, 30 × 32 = 960, 11 × 28 = 308, 5 × 28 = 140, 30 × 28 = 840 cross-terms for P0-GD, P1-GD, P2-GD, P0-DLI, P1-DLI, and P2-DLI respectively.

Table 10 Five protein structure similarity descriptors of HDAC2, 4, 6, 7 and 8

	Class I		Class IIa		Class IIb
	HDAC2	HDAC8	HDAC4	HDAC7	HDAC6
Class I					
HDAC2	1.000	0.407	0.182	0.182	0.031
HDAC8	0.407	1.000	0.180	0.186	0.048
Class IIa					
HDAC4	0.182	0.180	1.000	0.706	0.027
HDAC7	0.182	0.186	0.706	1.000	0.036
Class IIb					
HDAC6	0.031	0.048	0.027	0.036	1.000

Preprocessing of data

To reduce the bias of the model, all descriptors were mean centered and scaled to unit variance prior to the calculation of protein-ligand cross-terms. Moreover, the binding affinities (IC50) were logarithmically transformed to pIC50 and then mean centered and scaled to unit variance.

Proteochemometric modeling

Support vector machine (SVM) is a non-linear modeling technique applied multiple times in PCM [33,45-50]. We created PCM models using support vector regression (SVR) built in Weka suit (Weka implementation "SMOreg"). Eighteen different combinations of descriptor blocks were constructed to derive PCM models, *i.e.*, six combinations of protein and ligand descriptors (P0-DLI, P1-DLI, P2-DLI, P0-GD, P1-GD, and P2-GD), six combinations of protein and ligand descriptors with cross-terms, and the only six kinds of cross-terms.

There are a lot of kernel functions used in SVM, such as Normalized Poly Kernel (normalized polynomial kernel), Poly Kernel (polynomial kernel), Precomputed Kernel Matrix Kernel, Puk (Pearson VII function-based universal kernel), RBF Kernel (Radial Basis Function kernel), and String Kernel. Although Poly Kernel and RBF Kernel are most commonly used kernel functions, Puk Kernel is considered as a

universal kernel that is capable of serving as a generic alternative to the common linear, polynomial and RBF kernel functions [51]. In fact, we also found that Puk kernel had a stronger mapping power than the other kernels for our data set. For this reason, all models were created using SVR with Puk kernel.

Validation of PCM models

For each combination of descriptors, 10-fold cross-validation was carried out for the model. The performance of the derived eighteen models was assessed by the goodness-of-fit (R^2) and predictive ability (Q^2_{cv}, Q^2_{test}).

Finally, ten known inhibitors [4] were selected as the external validation dataset to assess the specificity performance of the best model. These inhibitors are listed in Table 6 including four pan-HDAC inhibitors (TSA, SAHA, panbinostat, and belinostat), three class I-specific inhibitors (MGCD0103, depsipeptide, and apicidin), and three class II-specific inhibitors (APHA, Tubacin, and NCT-10a). We predicted all the affinity values of the ten inhibitors against all the HDACs with the best model. According to the predicted results, we analyzed the interaction strength of the inhibitors with multiple HDACs and then select out isoform-specific, class-specific as well as pan inhibitors.

The framework of this work is presented in Figure 1.

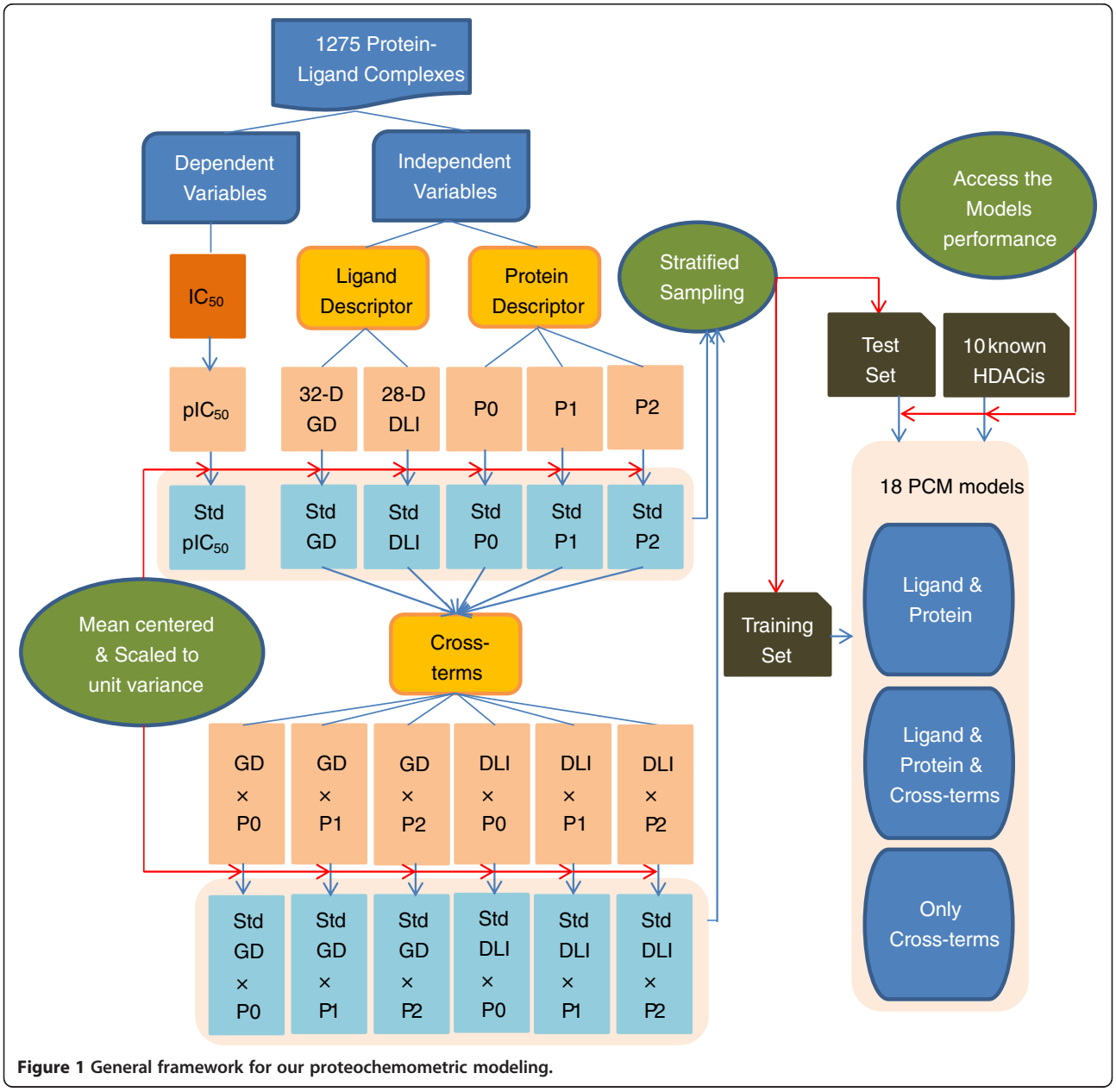


Figure 1 General framework for our proteochemometric modeling.

Additional files

Additional file 1: Table S4. Structures of 1443 chemical compounds (in SMILE format).

Additional file 2: Table S1. Train set used for construction of the proteochemometric models.

Additional file 3: Table S2. Test set used for assessment of the proteochemometric models.

Additional file 4: Table S3. Protein geometry descriptors.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Conceived and designed the experiments: RZ ZC. Performed the experiments: DW QH YZ QZ. Analyzed the data: DW QH YZ QZ QL JG ZC RZ. Wrote the paper: DW QH YZ QZ QL JG ZC RZ. All authors read and approved the final manuscript.

Acknowledgements
This work was supported in part by grants from National Natural Science Foundation of China (30976611, 31100956 and 61173117), Research Fund for the Doctoral Program of Higher Education of China (20100072120050, 20110072120048), Project Shanghai Pujiang Talents Funding (11PJ1407400), TCM modernization of Shanghai (09dZ1972800), and Students Innovation Training Program (VI) of Tongji University.

Author details
[1]School of Life Sciences and Technology, Tongji University, Shanghai 200092, P.R. China. [2]School of Information Engineering, Shanghai Maritime University, Shanghai 201306, P.R. China. [3]Institute for Advanced Study of Translational Medicine, Tongji University, Shanghai 200092, P.R. China. [4]School of Pharmacy, Liaoning University of Traditional Chinese Medicine, Dalian, Liaoning 116600, P.R. China.

Received: 22 March 2012 Accepted: 16 August 2012
Published: 22 August 2012

References
1. Park H, Kim S, Kim YE, Lim SJ: **A structure-based virtual screening approach toward the discovery of histone deacetylase inhibitors: identification of promising zinc-chelating groups.** *Chem Med Chem* 2010, **5**(4):591–597.
2. Bertrand P: **Inside HDAC with HDAC inhibitors.** *Eur J Med Chem* 2010, **45**(6):2095–2116.
3. Rikiishi H: **Autophagic and apoptotic effects of HDAC inhibitors on cancer cells.** *J Biomed Biotechnol* 2011, **2011**:830260.
4. Witt O, Deubzer HE, Milde T, Oehme I: **HDAC family: What are the cancer relevant targets?** *Cancer Lett* 2009, **277**(1):8–21.
5. De Ruijter AJM, Van Gennip AH, Caron HN, Kemp S, Van Kuilenburg ABP: **Histone deacetylases (HDACs): characterization of the classical HDAC family.** *Biochem J* 2003, **370**:737–749.
6. Jungel A, Distler JHW, Gay S, Distler O: **Epigenetic modifications: novel therapeutic strategies for systemic sclerosis?** *Expert Rev Clin Immu* 2011, **7**(4):475–480.
7. Ito A, Kawaguchi Y, Lai CH, Kovacs JJ, Higashimoto Y, Appella E, Yao TP: **MDM2-HDAC1-mediated deacetylation of p53 is required for its degradation.** *EMBO J* 2002, **21**(22):6236–6245.
8. Huang BH, Laban M, Leung CH, Lee L, Lee CK, Salto-Tellez M, Raju GC, Hooi SC: **Inhibition of histone deacetylase 2 increases apoptosis and p21Cip1/WAF1 expression, independent of histone deacetylase 1.** *Cell Death Differ* 2005, **12**(4):395–404.
9. Juan LJ, Shia WJ, Chen MH, Yang WM, Seto E, Lin YS, Wu CW: **Histone deacetylases specifically down-regulate p53-dependent gene activation.** *J Biol Chem* 2000, **275**(27):20436–20443.
10. Waltregny D, Glenisson W, Tran SL, North BJ, Verdin E, Colige A, Castronovo V: **Histone deacetylase HDAC8 associates with smooth muscle alpha-actin and is essential for smooth muscle cell contractility.** *FASEB J* 2005, **19**(3):966.
11. Chen B, Cepko CL: **HDAC4 regulates neuronal survival in normal and diseased retinas.** *Science* 2009, **323**(5911):256–259.
12. Spiegelberg BD, Hamm HE: **G beta gamma binds histone deacetylase 5 (HDAC5) and inhibits its transcriptional co-repression activity.** *J Biol Chem* 2005, **280**(50):41769–41776.
13. Kasler HG, Verdin E: **Histone deacetylase 7 functions as a key regulator of genes involved in both positive and negative selection of thymocytes.** *Mol Cell Biol* 2007, **27**(14):5184–5200.
14. Mejat A, Ramond F, Bassel-Duby R, Khochbin S, Olson EN, Schaeffer L: **Histone deacetylase 9 couples neuronal activity to muscle chromatin acetylation and gene expression.** *Nat Neurosci* 2005, **8**(3):313–321.
15. Bali P, Pranpat M, Bradner J, Balasis M, Fiskus W, Guo F, Rocha K, Kumaraswamy S, Boyapalle S, Atadja P, *et al*: **Inhibition of histone deacetylase 6 acetylates and disrupts the chaperone function of heat shock protein 90 - a novel basis for antileukemia activity of histone deacetylase inhibitors.** *J Biol Chem* 2005, **280**(29):26729–26734.
16. Tong JJ, Liu J, Bertos NR, Yang XJ: **Identification of HDAC10, a novel class II human histone deacetylase containing a leucine-rich domain.** *Nucleic Acids Res* 2002, **30**(5):1114–1123.
17. Feng W, Lu Z, Luo RZ, Zhang X, Seto E, Liao WS, Yu Y: **Multiple histone deacetylases repress tumor suppressor gene ARHI in breast cancer.** *Int J Cancer* 2007, **120**(8):1664–1668.
18. Suzuki T: **Explorative study on isoform-selective histone deacetylase inhibitors.** *Chem Pharm Bull(Tokyo)* 2009, **57**(9):897–906.
19. Seal A, Aykkal R, Babu RO, Ghosh M: **Docking study of HIV-1 reverse transcriptase with phytochemicals.** *Bioinformation* 2011, **5**(10):430–439.
20. Ma C, Kang H, Liu Q, Zhu RX, Cao ZW: **Insight into potential toxicity mechanisms of melamine: an in silico study.** *Toxicology* 2011, **283**(2–3):96–100.
21. Zhu RX, Liu Q, Tang J, Li HL, Cao ZW: **Investigations on inhibitors of hedgehog signal pathway: a quantitative structure-activity relationship study.** *Int J Mol Sci* 2011, **12**(5):3018–3033.
22. Liu Q, Zhou H, Liu L, Chen X, Zhu RX, Cao ZW: **Multi-target QSAR modelling in the analysis and design of HIV-HCV co-inhibitors: an in-silico study.** *BMC Bioinforma* 2011, **12**:294.
23. Liu Q, Che DS, Huang Q, Cao ZW, Zhu RX: **Multi-target QSAR study in the analysis and design of HIV-1 inhibitors.** *Chinese J Chem* 2010, **28**(9):1587–1592.
24. Huang D, Kang H, Zhang DF, Sheng Z, Liu Q, Zhu RX, Cao ZW: **Comparison of ligand-, target structure-, and protein-ligand interaction fingerprint-based virtual screening methods.** *Acta Chim Sinica* 2011, **69**(5):515–522.
25. Kang H, Sheng Z, Zhu R, Huang Q, Liu Q, Cao Z: **Virtual drug screen schema based on multiview similarity integration and ranking aggregation.** *J Chem Inf Model* 2012, **52**(3):834–843.
26. Dai T, Liu Q, Gao J, Cao ZW, Zhu RX: **A new protein-ligand binding sites prediction method based on the integration of protein sequence conservation information.** *BMC Bioinforma* 2011, **12**(Suppl 14):S9.
27. Sheng Z, Huang Q, Kang H, Liu Q, Cao ZW, Zhu RX: **A new fingerprint of chemical compounds and its application to drugs virtual screening.** *Acta Chim Sinica* 2011, **69**(16):1845–1850.
28. Sheng Z, Kang H, Dai TL, Liu Q, Zhu RX: **Complementary study of structure features and gene profile features for chemical compounds.** *Acta Chim Sinica* 2010, **68**(23):2395–2400.
29. Zhu RX, Hu LW, Li HY, Su J, Cao ZW, Zhang WD: **Novel natural inhibitors of CYP1A2 identified by in silico and in vitro screening.** *Int J Mol Sci* 2011, **12**(5):3250–3262.
30. Lapins M, Wikberg JES: **Proteochemometric modeling of drug resistance over the mutational space for multiple HIV protease variants and multiple protease inhibitors.** *J Chem Inf Model* 2009, **49**(5):1202–1210.
31. Lapinsh M, Veiksina S, Uhlen S, Petrovska R, Mutule I, Mutulis F, Yahorava S, Prusis P, Wikberg JES: **Proteochemometric mapping of the interaction of organic compounds with melanocortin receptor subtypes.** *Mol Pharmacol* 2005, **67**(1):50–59.
32. Lapinsh M, Prusis P, Mutule I, Mutulis F, Wikberg JES: **QSAR and proteo-chemometric analysis of the interaction of a series of organic compounds with melanocortin receptor subtypes.** *J Med Chem* 2003, **46**(13):2572–2579.
33. Geppert H, Humrich J, Stumpfe D, Gartner T, Bajorath J: **Ligand prediction from protein sequence and small molecule information using support vector machines and fingerprint descriptors.** *J Chem Inf Model* 2009, **49**(4):767–779.

34. Lapins M, Eklund M, Spjuth O, Prusis P, Wikberg JES: **Proteochemometric modeling of HIV protease susceptibility.** *BMC Bioinforma* 2008, **9:**181.
35. Mandrika I, Prusis P, Yahorava S, Shikhagaie M, Wikberg JES: **Proteochemometric modelling of antibody-antigen interactions using SPOT synthesised peptide arrays.** *Protein Eng Des Sel* 2007, **20**(6):301–307.
36. Lapinsh M, Prusis P, Uhlen S, Wikberg JES: **Improved approach for proteochemometrics modeling: application to organic compound - amine G protein-coupled receptor interactions.** *Bioinformatics* 2005, **21**(23):4289–4296.
37. van Westen GJP, Wegner JK, IJzerman AP, van Vlijmen HWT, Bender A: **Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets.** *Med Chem Comm* 2011, **2**(1):16–30.
38. Lutu PEN, Engelbrecht AP: **Using OVA modeling to improve classification performance for large datasets.** *Expert Syst Appl* 2012, **39**(4):4358–4376.
39. Rice P, Longden I, Bleasby A: **EMBOSS: the European molecular biology open software suite.** *Trends Genet* 2000, **16**(6):276–277.
40. *EMBOSS*; http://www.ebi.ac.uk/Tools/emboss/align/index.html.
41. *RCSB PDB Protein Comparison Tool*; http://www.rcsb.org/pdb/workbench/workbench.do.
42. *RCSB PDB Geometry: Structure Variance Analysis (find it in Geometry of a protein)*; http://www.rcsb.org/pdb/home/home.do.
43. Labute P: **A widely applicable set of descriptors.** *J Mol Graph Model* 2000, **18**(4–5):464–477.
44. Xu J, Stevenson J: **Drug-like index: a new approach to measure drug-like compounds and their diversity.** *J Chem Inf Comp Sci* 2000, **40**(5):1177–1187.
45. Bock JR, Gough DA: **A new method to estimate ligand-receptor energetics.** *Mol Cell Proteomics* 2002, **1**(11):904–910.
46. Bock JR, Gough DA: **Virtual screen for ligands of orphan G protein-coupled receptors.** *J Chem Inf Model* 2005, **45**(5):1402–1414.
47. Stroembergsson H, Daniluk P, Kryshtafovych A, Fidelis K, Wikberg JES, Kleywegt GJ, Hvidsten TR: **Interaction model based on local protein substructures generalizes to the entire structural enzyme-ligand space.** *J Chem Inf Model* 2008, **48**(11):2278–2288.
48. Jacob L, Hoffmann B, Stoven V, Vert JP: **Virtual screening of GPCRs: an in silico chemogenomics approach.** *BMC Bioinforma* 2008, **9:**363.
49. Ning X, Rangwala H, Karypis G: **Multi-assay-based structure-activity relationship models: improving structure-activity relationship models by incorporating activity information from related targets.** *J Chem Inf Model* 2009, **49**(11):2444–2456.
50. Fernandez M, Ahmad S, Sarai A: **Proteochemometric recognition of stable kinase inhibition complexes using topological autocorrelation and Support Vector Machines.** *J Chem Inf Model* 2010, **50**(6):1179–1188.
51. Ustun B, Melssen WJ, Buydens LMC: **Facilitating the application of support vector regression by using a universal Pearson VII function based kernel.** *Chemometr Intell Lab* 2006, **81**(1):29–40.

doi:10.1186/1471-2105-13-212
Cite this article as: Wu *et al.*: Screening of selective histone deacetylase inhibitors by proteochemometric modeling. *BMC Bioinformatics* 2012 **13**:212.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central