A pipeline for improved QSAR analysis of peptides: physiochemical property parameter selection via BMSF, near-neighbor sample selection via semivariogram, and weighted SVR regression and prediction

Zhijun Dai, Lifeng Wang, Yuan Chen, Haiyan Wang, Lianyang Bai, Zheming Yuan

Published Version Information

Citation: Dai, Z., Wang, L., Chen, Y., Wang, H., Bai, L., & Yuan, Z. (2014). A pipeline for improved QSAR analysis of peptides: Physiochemical property parameter selection via BMSF, near-neighbor sample selection via semivariogram, and weighted SVR regression and prediction. Amino Acids, 46(4), 1105-1119.

Digital Object Identifier (DOI): doi:10.1007/s00726-014-1667-5

Publisher's Link: http://link.springer.com/article/10.1007%2Fs00726-014-1667-5

Original Article

A pipeline for improved QSAR analysis of peptides: physiochemical property parameter selection via BMSF, near-neighbor sample selection via semivariogram, and weighted SVR regression and prediction

Zhijun Dai [1,2,†], Lifeng Wang[1,2,†], Yuan Chen[1,2], Haiyan Wang[3], Lianyang Bai [2,4] and Zheming Yuan [1,2,*]

[1] Hunan Provincial Key Laboratory of Crop Germplasm Innovation and Utilization, Hunan Agricultural University, Changsha, China; E-Mail: daizhijun@foxmail.com (Z.J.D.)

[2] Hunan Provincial Key Laboratory for Biology and Control of Plant Diseases and Insect Pests, Hunan Agricultural University, Changsha, China; E-Mail: wanglifeng112358@163.com (L.F.W.), chenyuan0510@126.com (Y.C.)

[3] Department of Statistics, Kansas State University, Manhattan, Kansas, USA; E-Mail: hwang@ksu.edu (H.Y.W.)

[4] Hunan Academy of Agricultural Sciences, Changsha, China; E-Mail: bailianyang2005@aliyun.com (L.Y.B.)

[†] These authors contributed equally to this work.

[*] Author to whom correspondence should be addressed; E-Mail: zhmyuan@sina.com; Tel.: +86-731-84613956; Fax: +86-731-84673775.

Received: / Accepted: / Published:

Abstract: In this paper, we present a pipeline to perform improved QSAR analysis of peptides. The modeling involves a double selection procedure that first performs feature selection and then conducts sample selection before the final regression analysis. Five hundred and thirty-one physicochemical property parameters of amino acids were used as descriptors to characterize the structure of peptides. These high-dimensional descriptors then go through a feature selection process given by the Binary Matrix Shuffling Filter (BMSF) to obtain a set of important low dimensional features. Each descriptor that passed the BMSF filtering also receives a weight defined through its contribution to reduce the estimation error. These selected features were served as the predictors for subsequent sample selection and modeling. Based on the weighted Euclidean distances between samples, a common range was determined with high-dimensional semivariogram and then used as a threshold to select the near-neighbor samples from the training set. For each sample to be predicted, the QSAR model was established using SVR with the weighted, selected features based on the exclusive set of near-neighbor training samples. Prediction was conducted for each test sample accordingly. The performances of this pipeline are tested with the QSAR analysis of angiotensin-converting enzyme (ACE) inhibitors and HLA-A*0201 data sets. Improved prediction accuracy was obtained in both applications. This pipeline can optimize the QSAR modeling from both the feature selection and sample selection perspectives. This leads to improved accuracy over single selection methods. We expect this pipeline to have extensive application prospect in the field of regression prediction.

Keywords: Peptides; Quantitative structure-activity regression; Feature selection; Semivariogram; Support vector regression

1. Introduction

Known as the important elements in biological world especially in human life, peptides have attracted considerable interest from biochemist and pharmacologist (Sewald and Jakubke 2002). With the development of peptide library, thousands of different peptides have been designed, synthesized, and then subjected to experimental screening procedures and biological assays. To be effectively used, the peptides data have been analyzed more and more using quantitative structure-activity regression (QSAR) method in recent years (Liang et al. 2006; Liang et al. 2009; Zhou et al. 2010; Hemmateenejad et al. 2012). The weak and transient interactions between peptides and modular domains often mediate protein-protein interactions. So characterizing the interaction interface of domain–peptide complexes and predicting binding specificity for modular domains are critical for deciphering protein–protein interaction networks. The most abundant peptide recognition domain in the human proteome is the Src homology 3 (SH3) domain. Based on homology modeling, molecular dynamics and molecular field analysis, Hou et al. (2006) have constructed a complex structure of the amphiphysin-1 SH3 domain and a high-affinity peptide ligand and then performed three-dimensional QSAR analyses on the 200 peptides with known binding affinities to the amphiphysin-1 SH3 domain. A proof of concept study based on the molecular interaction energy components (MIECs) was conducted for predicting binding affinities of amphiphysin-1 SH3 domain interacting with its peptide ligands and for classifying peptides into binder and non-binder categories (Hou et al. 2008). A generic structure-based model was proposed to decipher the binding specificity of SH3 domains (Hou et al. 2009) and then it was used for predicting SH3 domain-mediated protein–protein interaction network in Yeast (Hou et al. 2012). Integrated computational prediction method and peptide microarray were used for detecting Abl1 SH3-binding peptides on proteome-wide, in which a comprehensive list of candidate interacting partners were provided for the Abl1 protein (Xu et al. 2012).

In QSAR modeling, how to characterize the properties of peptides is an important task. Since Kidera et al. (1985) first coded 10 orthogonal factors from 188 reported physicochemical properties through factor analysis, a series of inductive descriptors have been constructed and applied in peptide computational study. Some examples are Z-scales (Hellberg et al. 1991; Sandberg et al. 1998), ISA-ECI (Collantes and Dunn 1995), SZOTT (Liang et al. 2006), T-scales (Tian et al. 2007), ATS-QTMS (Yousefinejad et al. 2012), etc. However, these inductive descriptors are the linear combinations of the multiple physicochemical property parameters selected for the amino acids and hence, the QSAR models established using these descriptors could not clearly elucidate the correlation between the initial physicochemical properties and the bioactivity of peptides. As a relatively comprehensive summary of amino acid physicochemical properties, the 531 features derived from AAindex database (Kawashima and Kanehisa 2008) can be used as descriptors to characterize the primary structure of peptide and protein.

Even though using AA531 features as descriptors provides rich information, the feature dimensions also sharply increase. High dimensionality has adverse impact on QSAR modeling. In our previous study with gene expression data, a novel method called Binary Matrix Shuffling Filter (BMSF) has been proposed to select informative genes from high-dimension feature set for classification problems (Zhang et al. 2012). BMSF first conducts multi-round of filtering to reduce the dimensions of the features to a manageable low dimension and then performs a backward elimination to refine the selected feature set. Matthews Correlation Coefficient was the criterion in BMSF of Zhang et al. (2012) for feature selection and comparison of pattern classification accuracy among different models. In this study, the activity of pipetides is a continuous variable. We apply a modified BMSF method to select important descriptors for QSAR regression analysis by changing the selection criterion to mean squared prediction error. Importance ranking of the selected features can be given afterward.

Beyond the feature selection problem, another concern in QSAR modeling is how much weight each feature should be given to better describe its contribution to the model. Without assigning different weights, each feature was assumed to play a uniform role. In this case, the features with large values can mask the contribution of all the other features with small values. Consequently, the modeling is largely focused on only those features with large values, neglecting the information provided by other features. Li et al. (2005) considered the particular contributions of each feature and proposed a novel feature weighted fuzzy clustering algorithm ReliefF to assign weight for every feature. Wölfel et al. (2005) proposed to weight the different features in the Mahalanobis distance according to their distances after the variance normalization. Vivencio et al. (2007) constructed a feature weighting method based on a χ^2 statistical test to be used in conjunction with the k-NN classifier. To consider the different contributions of different features, noisy features should receive less weight and noise free features deserve more weight since they are more reliable. In this work, we assign a weight to each descriptor that passed the BMSF filtering based on how much contribution the selected descriptor helps to reduce the estimation error.

In addition to the aforementioned feature selection and weight assignment consideration in QSAR analysis of peptides, sample selection within a training set was also found to be helpful to improve prediction accuracy. It is important that the molecules in the training data are representative of the samples to be predicted. The prediction of a molecule's bioactivity is generally more accurate if the QSAR model was built with similar molecules. However, how to measure the similarity between two molecules and in particular how close a query molecule is to the training sample can be defined very differently. Some examples are multidimensional rectangle or ellipsoid containing a given fraction of the training set, each dimension corresponding to a chemical descriptor. Eriksson et al. (2000) have proposed a procedure for training set selection recognizing clustering and then, further tested and elaborated it by applying it to a series of 351 chemical substances. Furusjö et al. (2006) have demonstrated the importance of appropriate training set selection for QSAR analysis and how the reliability of QSAR predictions can be increased by outlier diagnostics. Sheridan et al. (2004) have proposed a way to estimate the reliability of the prediction of an arbitrary chemical structure on a given QSAR model, and given the training set from which the model was derived. By studying several non-locally fitted QSAR methods (random forest, ensemble artificial neural network, k-NN, Support Vector Machine with the linear kernel) without feature selection, Sheridan et al. found that the models constructed using the k nearest neighbor (KNN) samples often obtain better performance in prediction accuracy and time saving ($k \leqslant n$) compared to the model containing all the samples (global prediction). However, how to determine the optimal k value is still an unresolved issue at present. In addition, with the AA531 features as descriptors, the feature space for most dataset is very sparse due to high dimensionality. A direct consequence of the sparseness is that it requires to use majority of the feature space in order to find a certain proportion of nearest neighbors out of a given sample size. Therefore the k nearest neighbors are not local any more in that they are not close to the query molecule.

In this work, we define the closeness of peptide molecules based on the AA531 features. Peptides with similar physicochemical property parameters of amino acids are expected to have similar bioactivity levels. However, not all physicochemical property parameters are related to the bioactivity. Instead, we believe it is helpful to first perform descriptor selection and then define the closeness of a query peptide to a training set using the selected features. Through the high-dimensional semivariogram on weighted features, we give a common range that defines a multidimensional ellipsoid within the training data for each test sample. The training samples with selected features falling inside of the ellipsoid were selected as the near-neighbor samples for further QSAR modeling. The final QSAR model and prediction for this sample was established using SVR with the weighted, selected features and the near-neighbor training samples in the ellipsoid.

2. Results and Analysis

2.1. QSAR analysis on angiotensin-converting enzyme (ACE) inhibitors

To assess the performance of the pipeline proposed in this study, a set of 55 tripeptides as inhibitors of the angiotensin-converting enzyme (ACE) (Lin et al. 2008) was first analyzed. We conducted 5 different random partitions with the same ratio of 45/10 as Lin et al (2008) to form the training and test sets. The bioactivity of ACE inhibitors was expressed as the log values of $1/IC_{50}$ (pIC_{50}). The sequences of those tripeptides and their corresponding experimental data were presented in **Table 1**.

Table 1 Sequences and bioactivities of ACE inhibitors

In test set[a]	Peptide	Bioactivity	In test sample	Peptide	Bioactivity	In test sample	Peptide	Bioactivity
1	VVV	1.63	0	YGY	1.82	0	PGG	3.14
0	RPG	3.09	0	GYY	1.07	2	PGP	1.82
1	GRP	0.48	0	YYY	1.54	0	GPG	2.65
2,3	LLL	1.35	0	FIV	2.04	1	GGP	1.28
5	GLG	2.45	1	FPP	1.50	1,2	PGI	2.23
4	LGL	1.52	0	FPK	2.45	0	KPK	2.63
2	FGG	2.79	4,5	PFP	1.74	1	ADA	2.17
3	GFG	2.53	2	RRR	1.77	3,5	GEG	2.28
0	GGF	1.11	0	PPP	1.86	3,5	LEL	1.19
4,5	FFG	2.71	0	FFF	1.20	2,4	RGP	1.73
1,2	FGF	1.29	3,5	RGP	1.73	0	PIP	1.69
4	GFF	1.02	1	PGR	2.67	0	FPF	1.32
3,5	GGG	2.61	0	GGV	1.99	3	KPF	1.51
3	GYG	2.33	4	GVV	1.82	4	VYP	0.82
2	GGY	1.35	3,5	PPG	3.18	1	YPF	1.60
0	LGG	2.49	2,5	YGG	3.07	4	RPF	1.59
1	GGL	1.63	4	YYG	2.79	4,5	PPF	1.68
0	LLG	2.33	3	LDL	1.42			
0	GLL	1.47	2	VIF	0.78			

[a] tells in which partition, the peptide was used as a test sample; 0 means the peptide was not used for test sample in any of the partitions.

Using the AA531 descriptors to represent the structural information of peptides yielded a total of 1593 features for each tripeptide. In each of partitions, the high-dimensional features were screened firstly by BMSF performing multi-rounds of selection on the training set. At this stage, a large number of redundant features were rejected and the feature dimension was reduced to an acceptable lower degree in all the 5 partitions. Further fine evaluation step of BMSF on those retained descriptors would gave an optimal subset consists of the final reserved descriptors. A summary of number of features selected by BMSF including the initial filtering and fine evaluation steps from 5 partitions were shown in **Table 2**. We could see that the number of rounds in initial filtering of BMSF ranged from 5 to 9, and the average number of final reserved features is 15.4 from the 5 different partitions. The residue position information and detailed description of retained features from 5 partitions for ACE inhibitors were shown in the first sheet of "**Supplementary Table 1.xlsx**".

Table 2 Number of retained features obtained by BMSF from 5 partitions for ACE inhibitors

Repetition	Rounds of initial filtering	Number of retained features[a]	Number of retained features[b]

1	6	24	17
2	9	22	18
3	5	24	14
4	6	21	12
5	6	24	16

[a] Number of retained features after the initial filtering of BMSF; [b] Number of retained features after the fine evaluation of BMSF

Based on the DMSE values of retained descriptors, the weight was calculated for each descriptor. Then, a new descriptor set was obtained through multiplying the values of the original descriptors with their corresponding weights. The weighted descriptors were used for further QSAR modeling as the independent variables.

With the weighted descriptors, the mean pairwise Euclidean distances were calculated and the semivariogram curve based on high-dimensional geostatistics (GS) was then plotted. Even though there are some variations among the semivariograms corresponding to the 5 different partitions in the left panel of **Figure 1**, there is a clear overall increasing pattern. This show that this dataset has strong structural property in that the correlation of the observed bioactivity level is strong among molecules of similar structure and the correlation is weak between molecules of different structures. The class mid-value corresponding to the largest semivariogram value was set as the range *a*. The values of the range *a* in these 5 partitions are 0.2458, 0.2499, 0.2468, 0.2501 and 0.2499, respectively. For each test sample in each partition, a multi-dimensional ball with radius *a* was formed with center given by the values of the weighted descriptors for the test sample. The training samples inside of the ball were the near-neighbors in that their mean Euclidean distances to the center calculated with the weighted descriptors were less than *a*. For the 5 partitions in ACE inhibitors dataset, the average values of the selected near-neighbor samples are 9.9, 11.1, 10.9, 11.2 and 11.3, respectively, which imply that heterogeneity exists in the dataset.

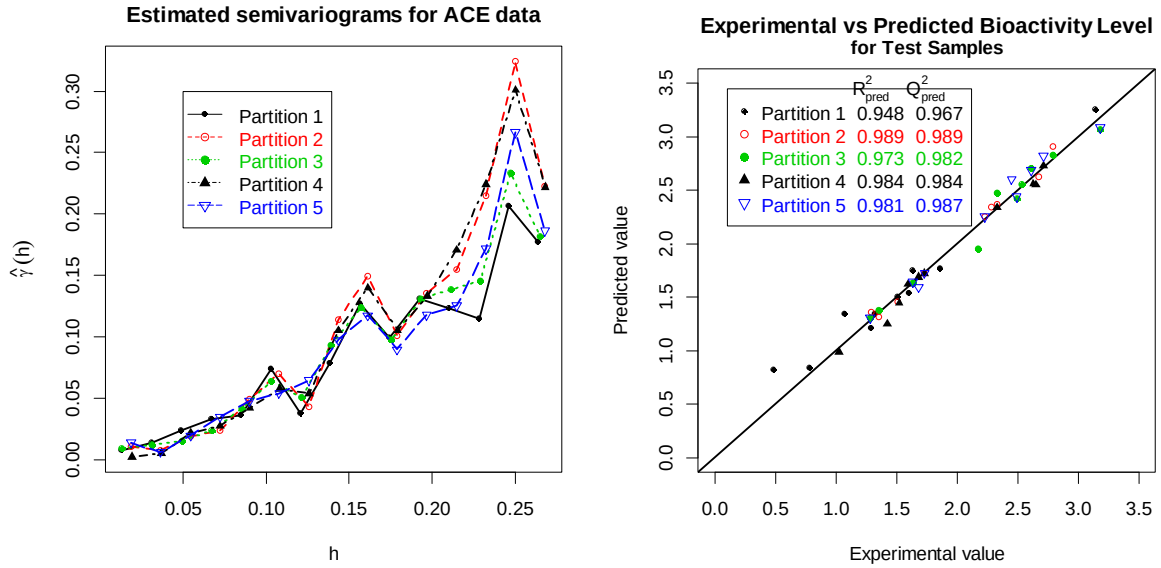

Figure 1 Estimated semivariogram (left panel) and experimental versus predicted bioactivity level plot (right panel) from the 5 partitions of the ACE inhibitors data. The predicted values in the right panel are from the full pipeline model.

Based on the weighted, selected descriptors, the QSAR model was constructed using SVR for each test sample with their exclusive set of selected near-neighbor samples in each of partition. The right panel of **Figure 1** shows the plot of the predicted values versus their experimental data on the test set from the 5 partitions for ACE inhibitors. In general, all the samples in the test set were uniformly scattered around the diagonal, and all values of the R^2_{pred} are over 0.940 and Q^2_{pred} are over 0.960. Thus the models could be regarded as valid for these test samples.

Table 3 lists the mean and standard deviation (in parenthesis) of the performance statistics from the 5 different partitions for ACE inhibitors established with different QSAR models. The results obtained by single SVR are relatively poor in predictive ability in terms of both the Q^2_{LOO} on the training set and the Q^2_{pred} on the test set. With feature selection, the method FS-SVR first performs QSAR modeling after BMSF screening of the high-dimensional features and then uses SVR for subsequent modeling. It is clear that after the process of feature selection, the performance of the constructed model was significantly improved as the Q^2_{pred} changed from 0.460 to 0.918. This verifies the efficiency of the feature selection on removing the redundancy among features or descriptors. The FS-Weight-SVR model not only consults with feature selection step but also adds the weights from BMSF for each descriptor. It can be seen that FS-Weight-SVR has some improvement over the FS-SVR model in predictive performance, which demonstrates the contribution of the weighting process. The full pipeline FS-Weight-GS-SVR uses the selected, weighted descriptors coupling with near-neighbor samples. In this case, the near-neighbor sample selection used the range of the semivariogram and the weights obtained in the end of feature selection step. It gave the best performance for predication of the test samples with average R^2_{pred} and Q^2_{pred} being 0.975 and 0.982 respectively. The two reference models Multiple Linear Regression (MLR) and Genetic Algorithm-Partial Least Squares (GA-PLS) were reported in Lin et al. (2008) and Hemmateenejad et al. (2011). The MLR does not involve feature selection procedure. The variable selection procedure in GA-PLS was conducted 5 times. The average number of selected variables is 10.2 and the average number of PLS latent variables used is 3. Our pipeline FS-Weight-GS-SVR has better prediction results than these two modes as can be seen in R^2_{pred} and Q^2_{pred}.

Table 3 Average performance statistics and their standard deviations from 5 partitions of ACE inhibitors using different QSAR models

Model	Descriptors	n_{UV}[a]	RMSEE	R^2	Q^2_{LOO}	RMSEP	R^2_{pred}	Q^2_{pred}
MLR	Lin scale	9	0.112 (0.004)	0.968 (0.002)	0.936 (0.007)	0.165 (0.015)	0.946 (0.012)	0.949 (0.009)
GA-PLS	QTMS-CUFQ	10.2(3)[b]	0.232 (0.016)	0.864 (0.020)	0.782 (0.030)	0.309 (0.065)	0.808 (0.079)	0.816 (0.077)
SVR	AA531	1593	0.009 (0.006)	1.000 (0.000)	0.407 (0.030)	0.540 (0.070)	0.437 (0.107)	0.460 (0.083)
FS-SVR	AA531	15.4	0.040 (0.004)	0.986 (0.002)	0.872 (0.027)	0.180 (0.069)	0.914 (0.063)	0.918 (0.057)
FS-Weight-SVR	AA531	15.4	0.078 (0.054)	0.975 (0.024)	0.892 (0.028)	0.171 (0.049)	0.943 (0.032)	0.945 (0.031)
FS-Weight-GS-SVR	AA531	15.4	**0.047** (0.006)	**0.995** (0.001)	**0.931** (0.021)	**0.092** (0.038)	**0.975** (0.016)	**0.982** (0.009)

[a] Average number of variables used, except for MLR and SVR; [b] The value in the parentheses is the average number of PLS latent variables

We summarized the union of the optimal feature sets from 5 partitions and obtained 72 features. The statistical results were shown in the second sheet of the "**Supplementary Table 1.xlsx**". In terms of the number of reserved descriptors for each residue, there were 19 and 24 descriptors reserved for the first and second residues, respectively, but 29 descriptors reserved for the third residue, which imply that the third residue site has more important effect on bioactivity of ACE inhibitor. The amino acid indices we used have been grouped into 6 groups with the result of clustering through analyzing amino acid indices and mutation matrices for sequence comparison (Tomii and Kanehisa, 1996). The frequency distributions of grouped descriptors in each residue position were shown in **Table 4**. We can see that the composition of amino acids in the first residue plays an important role in the activity of peptides, and the hydrophobicity for the second and the third residue has a significant association with the activities of ACE inhibitors.

Table 4 Frequency distributions of grouped descriptors in each residue position for ACE inhibitors

	No.1 Residue	No.2 Residue	No.3 Residue
A	4	5	4
B	2	3	1
C	5	1	6
H	3	9	10
O	4	5	3
P	1	1	5

A. alpha and turn propensities; B. beta propensity; C. Composition; H. Hydrophobicity; P. Physicochemical properties; O. Other properties

2.2. QSAR analysis on HLA data set

The binding of 177 nonameric peptides to the HLA-A*0201 molecule were used as the second peptide panel (Doytchinova et al. 2005) (Shown in **Table 5**). The binding activity was expressed as the logarithm half-maximal binding level (BL_{50}) which is the peptide concentration yielding the half-maximal fluorescence index of the reference peptide FLPSDFFPSV (IC_{50}=2.6nM). We conducted 5 different random partitions with the same ratio of 131/46 as Doytchinova et al. (2005) in training and test set.

Table 5 Sequences and bioactivities of HLA peptides

In test set[a]	Peptide	Exp.	In test set	Peptide	Exp.	In test set	Peptide	Exp.
1	ALCRWGLLL	4.91	3	ILDPFPVTN	5.29	1,3	RLWPFYHNV	5.72
4	ALIHHNTHL	4.30	0	ILDPFPVTP	5.82	2	RLWPIYHNV	5.77
3	ALPYWNFAT	4.66	2	ILDPFPVTQ	5.28	3,5	RLWPLYPNV	5.57
2,3	CLTSTVQLV	4.93	2	ILDPFPVTS	4.78	1,4,5	SIISAVVGI	4.47
1,3	FLCKQYLNL	5.21	5	ILDPFPVTT	5.54	3,5	SLHVGTQCA	3.79
1,2	FLDQVPFSV	5.98	0	ILDPFPVTV	8.65	5	SLNFMGYVI	4.00
2,4	FLLSLGIHL	5.17	5	ILDPFPVTW	4.71	3,5	SLYADSPSV	5.24
1,5	FLLTRILTI	4.95	1	ILDPFPVTY	3.19	0	TLGIVCPIC	4.68
4	FLNPFYPNV	6.16	3	ILDPIPPTV	7.30	4	TLHEYMLDL	4.94
1,2	FLWPFYHNV	5.99	0	ILDQVPFSV	6.09	3,5	TTAEEAAGI	3.39
1	FLWPFYPNV	5.89	3,4	ILFPGPVTA	6.23	1,4	VCMTVDSLV	4.20
0	FLWPIYHDV	6.16	1,2,4	ILKEPVHGV	5.59	0	VLHSFTDAI	4.54
2,4	FLWPIYHNV	6.37	3	ILWPIYHNV	6.24	3,4	VLIQRNPQL	5.06
2	FLWPLYPNV	6.14	1,5	ILWQVPFSV	5.91	4	VLLDYQGML	4.52
3,5	FVTWHRYHL	4.21	0	IMDPFPVTV	7.21	4	VTWHRYHLL	4.38
2	GLLGWSPQA	5.13	0	IMDQVPFSV	5.71	0	WILRGTSFV	4.06
0	GLSRYVARL	4.78	1,2	INDPFPVTV	4.78	3	WLDQVPFSV	5.23
3	GLYSSTVPV	5.15	1,4	IPDPFPVTV	5.10	0	YAIDLPVSV	5.63
0	HLESLFTAV	3.79	3	IQDPFPVTV	6.05	1,2	YLAPGPVTA	5.74
1,2,4,5	HLLVGSSGL	3.91	3,5	ISDPFPVTV	5.50	0	YLAPGPVTV	6.00
5	HLYSHPIIL	5.41	0	ITAQVPFSV	4.43	2,4	YLCPGPVTA	6.18
4,5	IADPFPVTV	5.76	0	ITDPFPVTV	6.08	1	YLEPGPVTL	5.41

3,5	ICDPFPVTV	5.45	1,2	ITDQVPFSV	4.48	0	YLFDGPVTA	5.50
2,3	IDDPFPVTV	4.36	1,2,4,5	ITFQVPFSV	4.42	3,5	YLFNGPVTA	5.80
0	IFDPFPVTV	4.89	0	ITWQVPFSV	5.01	3,5	YLFNGPVTV	5.65
2,3	IGDPFPVTV	3.92	5	IVDPFPVTV	6.21	3,5	YLFPCPVTA	6.63
2,4,5	IHDPFPVTV	4.96	1,5	IWDPFPVTV	5.13	1,2,4,5	YLFPDPVTA	6.09
0	IIDPFPVTV	6.31	3	IYDPFPVTV	5.41	1	YLFPGPETA	5.81
0	IISCTCPTV	5.17	0	KIFGSLAFL	4.40	2,3,4	YLFPGPFTV	5.81
0	ILDDFPVTV	7.16	1,2,3,4	KLHLYSHPI	4.77	0	YLFPGPMTA	5.98
1,2,5	ILDDLPPTV	7.14	5	KLPQLCTEL	4.50	1,5	YLFPGPMTV	5.85
1,3	ILDPFPPEV	7.68	0	KTWGQYWQV	4.43	0	YLFPGPSTA	5.69
4,5	ILDPFPPTV	8.17	2,4	LLFGYPVYV	5.45	1,4	YLFPGPVQA	6.14
1	ILDPFPVTA	6.32	3,4	LLMGTLGIV	4.21	0	YLFPGPVTA	6.31
3,4	ILDPFPVTC	5.65	2,5	LLWFHISCL	4.13	2	YLFPGPVTG	5.22
5	ILDPFPVTD	2.94	4,5	LQTTIHDII	3.90	1,4	YLFPPPVTA	5.75
1,2,4	ILDPFPVTE	3.13	1,2	MLDLQPETT	4.36	4	YLFPPPVTV	6.19
2	ILDPFPVTF	5.67	3	MLGTHTMEV	5.37	1	YLNPGPVTA	5.53
1,3	ILDPFPVTG	6.66	2,3,4	NLQSLTNLL	3.96	3	YLSPGPVTA	5.44
0	ILDPFPVTH	3.60	1,2	NLSWLSLDV	4.75	1,2,4	YLWQYIPSV	5.17
2,5	ILDPFPVTI	6.69	4	NMVPFFPPV	5.60	5	YLYPGPVTA	5.77
4	ILDPFPVTK	4.59	0	PLLPIFFCL	5.32	5	YMNGTMSQV	4.67
2,4	ILDPFPVTL	7.03	2,4	RLLQETELV	4.83	0	YTDQVPFSV	4.80
1	ILDPFPVTM	6.13	4	RLMKQDFSV	4.97	0	YLFDGPVTV	4.96
1	ALMPLYACI	5.08	5	ILKPLYHNV	5.25	2,3	YLFPFPITV	6.68
3	FLDDHFCTV	6.68	2,3	ILNPFYHNV	6.16	1,2	YLFPGPFTA	5.65
5	FLFPGPVTA	6.18	2	ILNPFYPDV	6.11	3	YLFPGPVWA	5.59
0	FLFPLPPEV	6.53	2	ILWPLFHEV	6.03	3,4	YLFPGTVTA	6.16
4	FLKPFYHNV	5.73	1,4,5	ILWPLYPNV	6.06	3,5	YLFPGVVTA	6.17
0	FLNPIYHDV	6.16	0	ILYQVPFSV	5.06	0	YLFQGPVTA	5.21
0	FTDQVPFSV	4.76	1,3	ITSQVPFSV	4.06	4	YLKPGPVTA	5.26
2	GILTVILGV	4.57	2,3	LLAQFTSAI	4.51	3	YLMPGPVTA	5.27
4	GLGQVPLIV	4.76	1,2	LMAVVLASL	3.99	5	YLWDHFIEV	6.36
1	GTLGIVCPI	5.23	4	RLNPFYHDV	4.24	5	YLWPGPVTV	5.70
2,5	ILDDFPPTV	7.08	1,2	RLNPLYPNV	5.37	5	YLWQYIFSV	4.94
3,5	ILDPFPITV	8.14	1,5	RLWPFYPNV	5.24	0	YMLDLQPET	5.28
4,5	ILDPFPPPV	7.44	1,4	RLWPIYHDV	5.55	4	YVITTQHWL	4.39
5	ILDPLPPTV	7.15	3	SLDDYNHLV	5.27			
0	ILFPFPVEV	6.80	0	SVYDFFVWL	5.12			
0	ILFPFVHSV	6.58	1,3	VMGTLVALV	5.03			

[a] tells in which partition the peptide was used as a test sample ; 0 means the peptide was not used for test sample in any of the partitions.

The AA531 was used to represent the structural information of peptide giving 4779 descriptors for each nonapeptide. In each of partitions, the high-dimensional features were screened firstly by BMSF performing over 10 rounds of initial filtering on the training set due to the high dimensional features. At this stage, over 4700 features were screened and there

were about an average of 20 features retained from these 5 partitions. Further feature selection was then conducted for those reserved features with the fine evaluation of BMSF. Finally, an average of 15.8 features was reserved from 5 partitions. The summary of the number of selected features by BMSF for HLA peptides was shown in **Table 6**. The residue position information and detailed description of retained features from these 5 partitions for HLA peptides were shown in the first sheet of "**Supplementary Table 2.xlsx**".

Table 6 Number of retained features obtained by BMSF from 5 partitions for HLA peptides

Repetition	Rounds of initial filtering	Number of retained features[a]	Number of retained features[b]
1	13	25	19
2	12	20	17
3	11	21	16
4	10	17	13
5	12	18	14

[a] Number of retained features after the initial filtering of BMSF; [b] Number of retained features after the fine evaluation of BMSF

The weight factor for each reserved descriptor was calculated based on its corresponding DMSE value. By multiplying the values of the original descriptors with their corresponding weight factors, the final weighted feature set was obtained and then used for subsequent QSAR analysis.

The variation of the semivariogram values $\hat{\gamma}(h)$ in each of the 5 partitions with class mid-values h was depicted in the left panel of **Figure 2**. There are a lot more variations in these semivariograms than those for the previous data sets. The mid-values, i.e., the range a, in 5 different partitions were 0.1751, 0.1751, 0.1644, 0.1745 and 0.1745, respectively, at which the semivariogram achieved the largest value. These values were used as the threshold to select the exclusive near-neighbor training samples for each test sample in each partition. For the 5 partitions in HLA peptides dataset, the average values of the selected near-neighbor samples are 18.6, 19.2, 21.3, 19.3 and 20.0, respectively, which also imply that heterogeneity exists in the dataset.

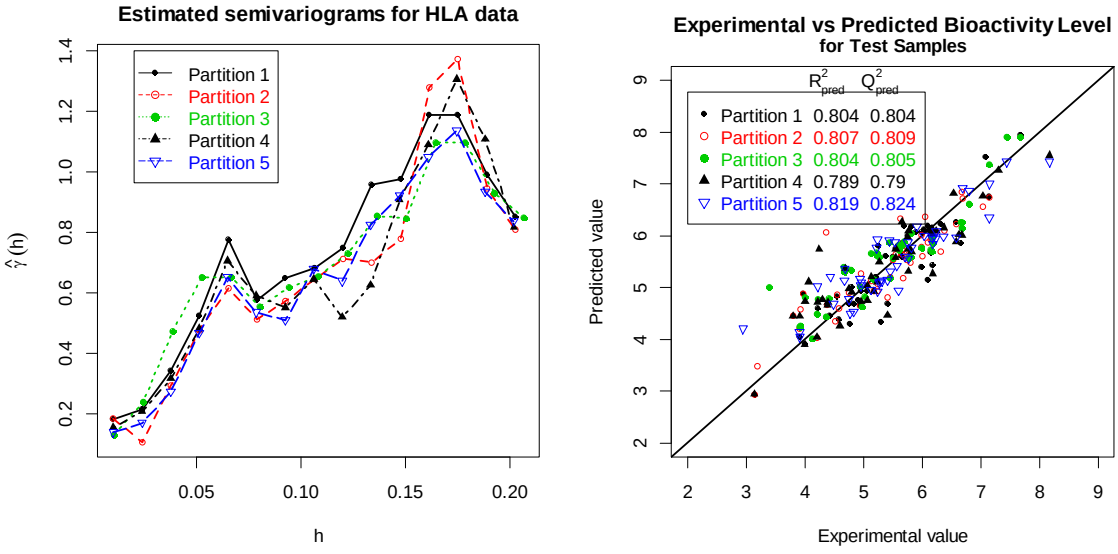

Figure 2 Estimated semivariogram (left panel) and experimental versus predicted bioactivity level plot (right panel) from the 5 partitions of the HLA peptides data. The predicted values in the right panel are from the full pipeline model.

Based on the final selected and weighted descriptors, the QSAR model was constructed using SVR for each test sample with their exclusive set of selected near-neighbor samples in each of repetition. The right panel of **Figure 2** shows the plot of the predicted values versus their experimental data on the test set in these 5 partitions for HLA peptides, the samples were scattered around the diagonal but with some deviations. This tells that the established QSAR model has satisfactory

extrapolating ability but large variations exist among the data, which can also be seen from the variations in semivariogram plots.

Table 7 lists the mean and standard deviation of the performance statistics of our proposed pipeline and other reference methods from these 5 partitions for HLA peptides dataset. The models with the feature selection process using BMSF are significantly superior to the single SVR model. This reveals that significant redundancy exists in the original feature set. For the SVR model, the prediction performance on both leave-one-out cross validation and external prediction are poor. Even though the fitting measure R^2 on the training set almost reached 1, its external prediction measure R^2_{pred} and Q^2_{pred} on test sets are still very low. This shows that there is extreme over-fitting in the single SVR model. For the three SVR models processed with feature selection, the predictive R^2_{pred} on test sets are all over 0.770, and the predictive Q^2_{pred} are in the range of 0.778-0.806. In addition, these three models all performed better than existing PLS (Doytchinova et al. 2005) and GA-PLS models (Hemmateenejad et al. 2011). The model that used the range of semivariogram to select near-neighbor samples is has better average prediction measure R^2_{pred} and Q^2_{pred} than those without sample selection. The best average predictive result was obtained by the full pipeline FS-Weight-GS-SVR model, with statistics R^2_{pred} and Q^2_{pred} of 0.805 and 0.806, respectively. This demonstrates that using the exclusive selection of training samples for each test sample, the prediction performance of the QSAR model can be improved. However, due to the large standard deviations shown in **Table 7**, the improvement of prediction accuracy appears to be marginal.

Table 7 Average and standard deviation (in parenthesis) of the performance statistics from the 5 partitions of HLA peptides using different QSAR models

Model	Descriptors	n_{UV}[a]	RMSEE	R^2	Q^2_{LOO}	RMSEP	R^2_{pred}	Q^2_{pred}
PLS	Additive	161(3.4)[b]	0.377 (0.031)	0.854 (0.025)	0.291 (0.027)	0.561 (0.017)	0.603 (0.058)	0.627 (0.038)
GA-PLS	QTMS-ADFQ	19(3.6)	0.697 (0.071)	0.494 (0.081)	0.309 (0.072)	0.669 (0.067)	0.436 (0.096)	0.469 (0.081)
SVR	AA531	4779	0.024 (0.023)	0.999 (0.002)	-0.015 (0.000)	0.704 (0.064)	0.411 (0.108)	0.436 (0.063)
FS-SVR	AA531	15.8	0.241 (0.021)	0.938 (0.014)	0.641 (0.038)	0.425 (0.057)	0.771 (0.029)	0.778 (0.035)
FS-Weight-SVR	AA531	15.8	0.235 (0.024)	0.947 (0.022)	0.656 (0.053)	0.421 (0.062)	0.785 (0.018)	0.790 (0.027)
FS-Weight-GS-SVR	AA531	15.8	**0.275** (0.029)	**0.919** (0.015)	**0.666** (0.041)	**0.430** (0.032)	**0.805** (0.011)	**0.806** (0.012)

[a] Number of variables used; [b] The value in the parentheses is the number of PLS latent variables

The union of the final selected feature sets from the 5 partitions was given in the second sheet of the "**Supplementary Table 2.xlsx**". The numbers of selected descriptors are 17, 13 and 20 for the second, fourth and ninth residues, which indicates that these three residuals play important roles for the bioactivity of HLA peptides. The numbers of retained features in others positions are all less than 7. The frequency distributions of grouped descriptors on these 3 important sites are shown in **Table 8**. We can see that the alpha and turn propensities in the second residue might be correlated with the activity of HLA peptides. The hydrophobicity and the alpha and turn propensities in the fourth residue have apparent relevance with the bioactivity, and the hydrophobicity for the ninth residue plays an important role in the activity of HLA peptides. However, the beta propensity in all of the 3 sites is almost absent, which indicates that the variety of its values has little influence on the bioactivities of HLA peptides.

Table 8 Frequency distributions of grouped descriptors in each residue position for HLA peptides

	No.2 Residue	No.4 Residue	No.9 Residue
A	6	4	3

B	0	2	0
C	1	1	3
H	5	5	11
O	2	1	2
P	3	0	1

A. alpha and turn propensities; B. beta propensity; C. Composition; H. Hydrophobicity; P. Physicochemical properties; O. Other properties

3. Principles and Methodologies

3.1 Data set

Two peptide data sets with known bioactivity were used to test the performance of the QSAR models constructed with our pipeline. They were a set of 55 angiotensin-converting enzyme (ACE) inhibitors and a set of 177 nonameric peptides binding to the HLA-A*0201 molecule. The data sets were reported in Lin et al. (2008) and Doytchinova et al. (2005).

3.2 Structural description of peptide ligand

To characterize the peptide or protein structures, we consider the physicochemical properties of 20 amino acids derived from the AAindex database (Kawashima and Kanehisa 2008). There were 531 physicochemical properties that can be used as descriptors. Since these descriptors belong to different property groups such as electrical property, hydrophobicity, hydrophilicity and so on, there is quite a lot of heterogeneity across these descriptors. To eliminate their heterogeneity and prepare for subsequent descriptor weighting, each physicochemical property of 20 amino acids was normalized to a unified scale with mean zero and standard deviation one. In this work, the normalized 531 physicochemical descriptors (AA531) for each amino acid residue were first arranged in tandem order and then subject to feature selection before modeling.

3.3 Feature (descriptor) selection and reserved descriptor weighting

By using AA531 features to characterize peptide sequences, the tripeptides ACE inhibitors and nonameric peptides contain 1593 and 4779 descriptors, respectively. The sharp increase in feature dimensions is adverse for accurate modeling. For rapid and efficient selection of high dimensional features, we have reported a novel method named Binary Matrix Shuffling Filter (BMSF) based on Support Vector Classification (SVC). The method was successfully applied to classification of 9 cancer datasets and obtained excellent results (Zhang et al. 2012).

In this method, a controlled matrix with the same number of columns as the number of features was generated. The matrix consists of equal number of randomly positioned 0s and 1s per column. Then a 10-fold cross validation with SVM on the training set was conducted for each row of the matrix using only the features corresponding to the 1s in the row. The importance of a feature was judged based on contrasting the prediction accuracy of two sets of models, one set with the feature included and the other with the feature excluded. The accuracy was defined as the Matthews correlation coefficient (MCC) of the constructed model. The detailed procedures can be found in the reference (Zhang et al. 2012). This method is able to find a parsimonious set of features which has high joint prediction power.

To apply the idea of BMSF to QSAR analysis, we replace the support vector classification in the algorithm by support vector regression and use the mean squared prediction error (MSE) as the criterion instead of MCC to determine the contribution of each feature set. For each round of initial filtering, we use the features and the bioactivity on the training set. Starting with the generated matrix with binary values as mentioned above, each row defines a subset of features by including only features such that the value 1 appears in the row for those features. For example, if the first five elements of

a row in the binary matrix are 0, 1, 0, 0, 1, then 2nd and 5th features are in the subset while 1st, 3rd, and 4th features are excluded. With the subset of features and the bioactivity level of the training set as the response variable, conduct a 10-fold cross validation using SVM regression and obtain the cross-validated MSE. This MSE gave us an initial idea of the contribution from this subset of features. Such MSE was obtained for every row of the binary matrix. Since the binary matrix contains equal number of 1s and 0s in each column, every feature was included in about half of the computed MSEs. However, as many features are in each subset, the obtained MSEs contain the mutual contribution of many features.

To consider the contribution of the i^{th} feature alone, we obtain a new matrix by changing all the 1s in i^{th} column of the binary matrix to 0 and all the 0s in that column to 1 while keeping the remaining columns of the matrix unchanged. This switch between 1 and 0 in the ith column alone gives us a contrasting feature set compared to the feature set defined by the same row of the original binary matrix. A SVM regression model was then trained using the MSE values as the response variable and the original binary matrix as the independent variables. Then the model was used to predict the value of the response variable (i.e. MSE) for each row of the new binary matrix. All the differences between the cross-validated MSE and the predicted MSE values reflect the change due to the switch between exclusion (0) and inclusion (1) for the i^{th} feature while holding all other features in each subset unchanged. Next form two vectors Z_1 and Z_0 by first initializing them equal to the cross-validated MSE and predicted MSE, respectively. Then switch the j^{th} element of Z_1 with the corresponding element of Z_0 if the value in j^{th} row and i^{th} column of the original binary matrix is 0, where j is an integer between 1 and the number of rows in the binary matrix. After the switch, Z_1 and Z_0 give us a paired data set, one with i^{th} feature included and the other with i^{th} feature excluded from the model. Conduct a paired comparison with Z_1 and Z_0. If the mean of Z_0 is greater than the mean of Z_1, then including the i^{th} feature tends to give better prediction performance measured by MSE. Such paired comparison gives us an idea of how significant the i^{th} feature contributes to explain the variations in the bioactivity level conditional on various combinations of other features included in the model. Repeat the process for all the features and discard those features with higher mean of Z_0. This finishes one round of initial filtering.

Generally, the feature selection goes through several rounds of initial filtering and would be stopped when the MSE for the variable subset from cross-validation starts to increase. After the initial filtering with BMSF, the dimensions of the selected feature set often declines quickly to an acceptable lower degree. To further improve the model robustness in QSAR analysis, a backward elimination was performed to refine the selected feature set. Starting with all the features retained after the initial filtering, a MSE_0 can be obtained during the 10-fold cross-validation on the training set. Then a cross-validation MSE vector (MSE_1, …, MSE_j, …, MSE_{q1}) is obtained by eliminating the j^{th} descriptor one by one, where $q1$ is the number of initially selected descriptors. If all the elements of the MSE vector are greater than the MSE_0, the backward elimination will stop since by deleting any of the descriptors will cause the model accuracy decrease. Otherwise, the descriptor that corresponds to the minimum in the MSE vector is deleted since the precision of the model improves the most when the descriptor is absence. Further backward elimination continues in the same way repeatedly. Every time either the least significant descriptor is eliminated from the pool of selected descriptors or the process stops when no further improvement can be made.

In the last round of backward elimination, suppose the final set of selected descriptors is $\{X_1, \ldots, X_q\}$. Let $MSE_{X_1,\ldots,X_{j-1},X_{j+1},\ldots,X_q}$ and $MSE_{X_1,\ldots,X_j,\ldots,X_q}$ be the mean squared error from 10-fold cross validation without and with the j^{th} selected descriptor, respectively. The difference between the two tells the relative importance of the j^{th} descriptor when compared to other selected features:

$$DMSE_j = MSE_{X_1,\ldots,X_{j-1},X_{j+1},\ldots,X_q} - MSE_{X_1,\ldots,X_j,\ldots,X_q} \quad (1)$$

To differentiate the contribution of different features, the normalized weight w_j for descriptor x_j can be calculated through the equation below:

$$w_j = \frac{DMSE_j}{\sum_{j=1}^{q} DMSE_j} \tag{2}$$

3.4 Near-neighbor sample selection and individual prediction

The bioactivity of a peptide is a stochastic process which depends on the structure of the peptide and the physicochemical properties of the amino acids that formed the peptide. It has been a common belief to majority that peptides with similar structure and physicochemical property parameters tend to have close activities. To improve the bioactivity prediction accuracy for a new peptide not in the training sample, it is important that the model is built with only peptides that are close to the new sample. Models using the entire set of descriptors selected by a feature selection procedure (such as the BMSF and backward elimination described above) can not guarantee that only closely related peptides be used in the modeling. When unrelated peptides are used in the model, the prediction becomes extrapolation which could lead to erroneous results.

Here we give a peptide selection method based on semivariogram of geostatistics extended to high-dimensional descriptor space. In geostatistics, the modeling of a process (such as the precipitation) typically considers the spatial correlation into account. Nearby locations with close latitude and longitude tend to be more correlated than locations that are far apart. Here in the bioactivity modeling, the physicochemical properties of amino acids serve a similar role as the longitude and latitude. The difference is that the location data (latitude and longitude) are in low dimensional space while the many physicochemical properties are in high dimensional space.

The traditional semivariogram in geostatistics depicts the spatial dependence of a stochastic process at two different locations. Specifically, for locations l_1 and l_2, the semivariogram is defined as $\gamma(l_1, l_2) = 0.5Var(Z(l_1) - Z(l_2))$, where $Z(l_i)$ represents the stochastic process at location l_i, i = 1, 2. It is a decreasing function of the covariance between the process values at the two locations. When the process is stationary, the semivariogram can be expressed as a function of the separating distance between the two locations. In general, the correlation decreases as the distance increases. Hence the semivariogram as a function of the distance h typically increases and then approaches a constant as the distance increases to infinity. At a certain distance a, the value of the semivariogram can be so close to the limit of the semivariogram such that the difference is negligible. This distance is referred as the range in geostatistics. It is the maximum separating distance such that the correlation between two locations needs to be considered.

In QSAR analysis, the bioactivity as a stochastic process also resembles the correlation pattern of the spatial process in that as the structure and physicochemical properties between two peptides differ more and more, their bioactivities are expected to be less and less correlated. Therefore, we extend the definition of semivariogram from the domain of 2-dimensional spatial locations to high dimensional descriptors. Specifically, we consider the weighted, selected descriptors $\{w_j \cdot X_j, j = 1, \ldots, q\}$ from the BMSF and backward elimination procedure. These weighted descriptors can be calculated for both the training and test samples. We will use the bioactivity levels of the training sample along with the values of their weighted descriptors to obtain the empirical semivariogram as a function of the separating distance.

Working with high dimensional descriptors, we need a measure of distance between molecules. Here we use the mean weighted Euclidean distance between the selected descriptors. Specifically, the distance between peptides A and B with selected descriptors $x_A = (x_{A1}, x_{A2}, \ldots, x_{Aq})$ and $x_B = (x_{B1}, x_{B2}, \ldots, x_{Bq})$ is defined as follows:

$$d_{AB} = q^{-1}\left(\sum_{j=1}^{q} w_j^2 \left(x_{A_j} - x_{B_j}\right)^2\right)^{1/2}, \tag{3}$$

where the w_j are the weights defined in equation (2).

To estimate the semivariogram function at different distances, we collect all the pairwise distances between the training samples and their corresponding observed bioactivity levels $\{(d_{AB}, y_A, y_B) : 1 \le A < B \le n_{tr}\}$. It is not possible to estimate the semivariogram at all values of distance due to limited samples. Instead, the whole interval between the minimum distance $min_{1\le A\le B\le n_{tr}} d_{AB}$ and maximum distance $max_{1\le A\le B\le n_{tr}} d_{AB}$ is further partitioned into overlapping intervals of equal width and the semivariogram is estimated for the representative value of each interval. More specifically, let M be the number of intervals to be partitioned and calculate $L = M^{-1}(max_{1\le A\le B\le n_{tr}} d_{AB} - min_{1\le A\le B\le n_{tr}} d_{AB})$. Then the lower and upper bounds of the intervals are given by $C_i = min_{1\le A\le B\le n_{tr}} d_{AB} + (i-1)L$ and $U_i = C_i + 1.5L$, $i = 1, \ldots, M$. The centers of the intervals $h_i = (U_i + C_i) / 2$ are the representative distances, for which the semivariogram function will be estimated. For reliable estimate, the number of intervals M and the maximum distance h for estimating the semivariogram need to follow the rule (Journel and Huijbregts 1978) that:

$$h \le 0.5(max_{1\le A\le B\le n_{tr}} d_{AB} - min_{1\le A\le B\le n_{tr}} d_{AB}) \tag{4}$$

For each interval i, $i = 1, \ldots, M$, let $Q_i = \{(A,B) : 1 \le A \le B \le n_{tr}, C_i \le d_{AB} \le U_i\}$ be the collection of indices for sample pairs whose distance is in the i^{th} interval. Let $N(h_i)$ be the number of sample pairs in Q_i. Then the empirical semivariogram at distance h_i is estimated as follows:

$$\hat{\gamma}(h_i) = \frac{1}{2N(h_i)} \sum_{(A,B)\in Q_i}^{N(h_i)} (y_A - y_B)^2 = \frac{1}{2N(h_i)} \sum_{1\le A\le B\le n_{tr}}^{N(h_i)} (y_A - y_B)^2 I(C_i \le d_{AB} \le U_i), \tag{5}$$

where $I(.)$ is the indicator function.

Once the semivariogram values are estimated for all $i = 1, \ldots, M$, we can find the range a of the semivariogram by examining the plot of h_i versus $\hat{\gamma}(h_i)$. Even though the theoretical semivariogram typically increases as the distances increases, the empirical values may fluctuate showing variations. We report the estimated range as the smallest h value among $h_1, \ldots, h_M$ that satistifies inequality (4) such that $\hat{\gamma}(h_i)$ is maximized.

In spatial analysis, the semivariogram is often fitted with a parametric model such as the exponential, spherical or Gaussian model. Such models involve additional parameters including the nugget and sill parameters whose estimation increases the computational burden. In addition, the choice of the model may not correctly describe the relationship between the semivariogram and distance. Hence we do not recommend to fit a particular parametric variogram model for estimating the range.

With the estimated range of the high dimensional semivariogram, we can decide which molecules in the training sample are close to the query molecule. To be more specific, for a query peptide with values for the selected descriptors given by ($x_{01}, \ldots, x_{0q}$), the set of training samples to be used for modeling of the bioactivity of the query peptide is $\left\{all(x_{A1},\ldots,x_{Aq},y_A) : q^{-1}\left(\sum_{j=1}^{q} w_j^2 (x_{Aj} - x_{0j})^2\right)^{1/2} \le a\right\}$, for $A = 1, \ldots, n_{tr}$. The range a essentially defines a high dimensional ellipsoid that are centered at the query peptide. Instead of a multidimensional ball, the ellipsoid differentiates the contribution of different selected descriptors by the different lengths of semi-principle axes of the ellipsoid through the weights. In terms of the weighted descriptors $z_{Aj} = w_j x_{Aj}$ and $z_{0j} = w_j x_{0j}$, the near-neighbor samples are located inside of a multidimensional ball $all(z_{A1},\ldots,z_{Aq},y_A)$ such that $\sum_{j=1}^{q}(z_{Aj} - z_{0j})^2 \le a^2 q^2$, for $A = 1, \ldots, n_{tr}$.

The prediction of a query peptide is based on the QSAR model using SVR fitted with the selected descriptors using the selected training samples from above peptide selection method.

3.5 QSAR Modeling and Validation

For a given training data set and any query sample, the QSAR model was fitted with the SVR using the selected samples and selected features. The test set was not used in the feature selection step since the bioactivities of peptides in the test set are unknown for the actual prediction and another important reason is that this would generate overly optimistic estimation for the external prediction if the test set were involved in variable selection. The features of the test samples were used in sample selection procedure but the bioactivity levels of the test samples were never used in any part of the model selection. We refer the entire pipeline as FS-Weight-GS-SVR in the reported result. The SVR model fit calls the software LIBSVM developed by Chang and Lin (2011). The epsilon-SVR was used and the nonlinear radial basis function (RBF) was used as the SVM kernel since the RBF kernel has shown much better generalization capability in most dataset. The tuning parameters of LIBSVM including cost *c*, RBF kernel parameter gamma *g* and epsilon *p* in loss function were optimized by a grid search strategy in cross validation on the training set. The range of *c* is the base 2 power of the elements of -1 to 6, the range of *g* is the base 2 power of the elements of -8 to 0 and the range of *p* is base 2 power of the elements of -1 to -8. The subroutines for doing FS-Weight-GS-SVR were written in MATLAB (Mathwork Inc., version 7.12.0.635 (R2011a)).

The reference models include PLS and GA-PLS. The PLS was conducted using the 'plsregress.m' program in the statistic toolbox of MATLAB. As an accepted procedure of refinement process in selecting the optimum number of PLS latent variables, minimum estimate of MSE was used from leave-one-out cross validation (LOO-CV). The GA-PLS algorithm toolbox developed by (Leardi 2000) was used to select the most suitable set of input variables for GA-PLS model We used the default values for the set of tuning parameters in this toolbox: i.e., the population size is 30 chromosomes on average 5 variables per chromosome in the original population; the deletion groups is 5; the maximum number of variables selected in the same chromosome is 30; the probability of mutation is 1%; the probability of cross-over is 50%; the maximum number of components is 15; the number of runs is 100; backward elimination was conducted after 100 evaluations; the window size for smoothing is 3.

For performance evaluation on the training set, we report the root mean squared error of estimation (RMSEE) for all peptides in the entire training data and the coefficient of determination R^2. Additionally, leave-one-out (LOO) cross-validation was also conducted for the training data. During the LOO cross-validation, the entire set of features first went through BMSF and backward elimination to perform feature selection. Then the selected features were used with all peptides in the training dataset to determine the common range *a* of the semivariogram. Afterward, each peptide was left out and the remaining $n_{tr} - 1$ peptides went through the sample selection procedure. All training samples that were in the ellipsoid centered at the left out sample were used to build the QSAR model with the SVR using the selected features. The prediction for the left out peptide was then conducted. For each of the training sample with the observed bioactivity level y_i, let $\hat{y}_i$ be the predicted level and $\bar{y}$ be the mean of the observed values of all samples in the training set. The following LOO coefficient of determination Q^2_{LOO} was reported by Golbraikh and Tropsha (2002):

$$Q^2_{LOO} = 1 - \frac{\sum_{i=1}^{n_{tr}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{tr}} (y_i - \bar{y})^2} \tag{6}$$

Since the feature selection and range determination were based on the entire training data, the Q^2_{LOO} is for reference only. It does not represent the extra sample performance.

For each peptide in the test data set, we used the selected descriptors and the common range *a* of the semivariogram determined from the training data to define an ellipsoid centered at the test sample. The training samples that fall inside of the ellipsoid were used to build the SVR model using the selected features based on the training data. To evaluate the performance of the proposed pipeline on all test data, we report the root mean squared error of prediction (RMSEP), the

coefficient of determination R^2_{pred} (Gedeck et al. 2006), and the external predictive evaluation index Q^2_{pred} proposed by Tropsha et al. (2003):

$$Q^2_{pred} = 1 - \frac{\sum_{i=1}^{n_{te}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{te}} (y_i - \bar{y}_{tr})^2} \quad (7)$$

where y_i is the experimentally observed activity level of the i^{th} sample in the test set, $\hat{y}_i$ is the predicted value, $i = 1, \ldots, n_{te}$. The $\bar{y}_{tr}$ is the mean of the experimentally observed values of all samples in the training set. We use Q^2_{pred} and R^2_{pred} as our major criteria to compare different models or methods. The only difference between R^2_{pred} and Q^2_{pred} is the mean value that is used in the denominator (Q^2_{pred}: mean of training data, R^2_{pred}: mean of test data). Since the numerator is equal in both cases and since $\bar{y}_{te}$ minimizes the sum of squares in the denominator, R^2_{pred} is always smaller than Q^2_{pred}.

In order to assess the risk of chance correlation for the pipeline proposed in this study, we report results from 5 different random partitions with the same ratio of references in training and test set for each data set. Based on the variability of the estimates for the different partitions, a rough estimate of the significance of feature selection and sample selection can be gained.

4. Discussion and Conclusions

Impact of feature selection and sample selection:

Based on the analyses of the two data sets, we could see that the precision of external prediction in SVR series models improved significantly after features selection was carried out (Tables 3 and 7). The evidence is obvious as the R^2_{pred} and Q^2_{pred} using SVR only are less than 0.500 for both datasets, but these measures in FS related methods are all over 0.918 for the ACE inhibitors and over 0.778 for the HLA peptides data set respectively. This clearly demonstrates that our feature selection method can significantly improve the SVR prediction accuracy. The models with extra feature weighting and sample selection appear to have even higher average prediction measures for both data sets. However, the numerical values of the model quality changes only slightly. Although our full pipeline gave the highest Q^2_{pred} value for both datasets (over 0.980 and 0.800 respectively), we could only conclude significant improvement of the full pipeline over FS-SVR and FS-Weight-SVR models in one data set. Specifically, for the ACE inhibitors dataset, the one-sided upper 95% confidence bound for Q^2_{pred} for FS-Weight-SVR and FS-SVR are 0.975 and 0.972, respectively (these were obtained based on t-distribution with 4 degrees of freedom for the standardized averages). This suggests that adding feature weighting does not make a significant improvement of accuracy at 0.05 level if the prediction does not involve sample selection (p-value from one-sided t-test =). Both upper confidence bounds are less than the Q^2_{pred} 0.982 for the full pipeline model, which suggests that results from the full pipeline with sample selection are significantly better than those of FS-SVR or FS-Weight-SVR at 0.05 level for this data set (p-value from one-sided test = and , respectively). For the HLA peptides dataset, the one-sided upper confidence bounds for Q^2_{pred} are 0.816 and 0.811 for the FS-Weight-SVR and FS-SVR methods respectively. Both upper bounds are greater than the Q^2_{pred} 0.806 of full pipeline performance, which tells that the impact of sample selection on prediction accuracy in the HLA peptides data set is not as significant as in the previous dataset.

There were several reasons that might have resulted in the small impact of sample selection relative to feature selection. Firstly, due to the small sample size in the training set for both datasets, the near-neighbor samples for modeling were insufficient as the range estimation of the semivariogram is restricted by the training sample size. When this happens, we can not expect big gains from sample selection. For the ACE inhibitors, the average number of near-neighbors was about 10 for the test set from the 5 partitions. There were only three or four near-samples for some samples in the test set, whose information for regression modeling was inadequate and lack of statistical significance. For the HLA peptides dataset, the

average number of near-neighbor sample was about 20 from 5 partitions, which accounted for a small part of the training set. Although the relationship between near-neighbors and the query sample was extremely relevant, there were still some related samples exist in the training set that were not included in the near neighbors. These samples could be included in the near-neighbors as the ranges of the semivariogram are extended. However, the near-neighbor samples for modeling would be close to all of the samples in the training set once the estimated range is too large if the sample size stays unchanged, and thus would lose the meaning of sample selection.

Secondly, due to the nature of high-dimensionality on the feature space and low-dimensionality on the samples space for both datasets, there were a large number of redundant features but far fewer uncorrelated samples to a query sample. Significant improvement in the precision of prediction was able to be achieved by feature selection because plenty of irrelevant features were eliminated. If the sample size is also large (for example, in thousands), the information in the samples is sufficient and there must exists a large number of irrelevant samples to a query sample in the test set. In this case, the impact of sample selection may be more prevalent and the impact of sample selection on model quality might be improved significantly.

Thirdly, the marginal effect of the sample selection might be resulted from the experimental flow. In our pipeline, the feature selection was first conducted followed by the sample selection in the end. However, we can not recommend a reverse experimental flow for these two datasets since the impact of feature selection is expected to be more important than sample selection due to high-dimensional features and small sample size in these two datasets. On the other hand, the experimental process can be reversed if the dataset has the characteristic of large number of samples and low-dimensional features. In this case, the impact of sample selection may become dominant. So the order of selection on features and samples should be considered on different datasets.

Conclusions:

Redundancy often exists among high-dimensional features, which are adverse for QSAR analysis. Feature selection is one of the critical steps to improve QSAR prediction accuracy. Recent studies also suggested that prediction of the bioactivity of a new molecule using a given training sample could give poor result if the training samples are very different from the new molecule. It has been reported in the literature that QSAR analysis based on the near-neighbors samples are often better than that with all the samples.

In this work, we presented a pipeline FS-Weight-GS-SVR for QSAR analysis that performs selection of important descriptors using the training data, conducts sample selection for each query molecule to reduce modeling error due to heterogeneity, and builds a QSAR prediction model with the selected samples based on the selected descriptors. For feature selection, this work extends our feature selection method BMSF previously given for classification problem to the current setting of regression problem. Through the selection with BMSF, the dimensions of the retained physicochemical properties decline to an easy-to-handle lower degree. In our application to the two datasets, 15.4 and 15.8 features on average from 5 partitions were retained from the original set with 1593 and 4779 features, respectively. To consider the different contributions of different features, each selected descriptor that passed the BMSF filtering also received a weight defined through its contribution to reduce the 10-fold cross validation estimation error.

For sample selection, we give a procedure to define a multi-dimensional ellipsoid centered around the query sample. The dimension of the ellipsoid is equal to the number of selected features. The lengths of the semi-principle axes of the ellipsoid depend on the weights of the selected descriptors and the range of the semivariogram function estimated with the training data. For each test sample, training samples that fall into the ellipsoid define an exclusive set of near-neighbor samples that can be used for further modeling. With the weighted selected features as the predictors and the exclusive set of near-eighbor samples as training data, the QSAR analysis was conducted for each test sample. The performance of our

proposed QSAR analysis pipeline was evaluated with the QSAR modeling of ACE inhibitors and HLA data sets. Satisfactory results confirm the validity and reliability of this method.

Overall, the method can optimize the QSAR model from both the feature selection and sample selection perspectives. This leads to improved accuracy over single selection methods, which has an extensive application prospect in the field of regression prediction for bioactivity of molecules.

Acknowledgments

This work was supported by the Doctoral Foundation of Ministry of Education of China (No. 20124320110002), the Scientific Research Fund of the Hunan Provincial Financial Department (No. 62020411074). The work of H. Wang was partially supported by a grant from the Simons Foundation (#246077).

Conflict of Interest

The authors declare no conflict of interest.

References

Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM T Intelli Syst Techn (TIST) 2:27

Collantes ER, Dunn WJ (1995) Amino acid side chain descriptors for quantitative structure-activity relationship studies of peptide analogues. J Med Chem 38:2705-2713.

Doytchinova IA, Walshe V, Borrow P et al (2005) Towards the chemometric dissection of peptide–HLA-A* 0201 binding affinity: comparison of local and global QSAR models. J Comput-Aided Mol Des 19:203-212

Eriksson L, Johansson E, Müller M, et al (2000) On the selection of the training set in environmental QSAR analysis when compounds are clustered. J Chemom 14:599-616

Furusjö E, Svenson A, Rahmberg M et al (2006) The importance of outlier detection and training set selection for reliable environmental QSAR predictions. Chemosphere 63:99-108

Gedeck P, Rohde B, Bartels C (2006) QSAR-how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. J Chem Inf Model 46:1924-1936

Golbraikh A, Tropsha A. Beware of q^2!. J Mol Graphics Modell 20:269-276

Hellberg S, Eriksson L, Jonsson J et al (1991) Minimum analogue peptide sets (MAPS) for quantitative structure-activity relationships. Int J Pept Protein Res 37:414-424

Hemmateenejad B, Yousefinejad S, Mehdipour AR (2011) Novel amino acids indices based on quantum topological molecular similarity and their application to QSAR study of peptides. Amino acids 40:1169-1183

Hemmateenejad B, Miri R, Elyasi M (2012) A segmented principal component analysis-regression approach to QSAR study of peptides. J Theor Biol 305:37-44

Hou T, Li N, Li Y et al (2012) Characterization of domain–peptide interaction interface: Prediction of SH3 domain-mediated protein–protein interaction network in yeast by generic structure-based models. J Proteome Res 11:2982-2995

Hou T, McLaughlin W, Lu B, et al (2006) Prediction of binding affinities between the human amphiphysin-1 SH3 domain and its peptide ligands using homology modeling, molecular dynamics and molecular field analysis. J Proteome Res 5:32-43

Hou T, Xu Z, Zhang W, et al (2009) Characterization of Domain-Peptide Interaction Interface A Generic Structure-based Model to Decipher the Binding Specificity of SH3 Domains. Mol Cell Proteomics 8:639-649

Hou T, Zhang W, Case DA, et al (2008) Characterization of domain–peptide interaction interface: a case study on the amphiphysin-1 SH3 domain. J Mol Bio, 376:1201-1214

Journel AG, Huijbregts CJ (1978) Mining geostatistics. Academic press, London

Kawashima S, Kanehisa M (2000) AAindex: amino acid index database. Nucleic Acids Res 28:374-374

Kidera A, Konishi Y, Oka M et al (1985) Statistical analysis of the physical properties of the 20 naturally occurring amino acids. J Protein Chem 4:23-55

Leardi R (2000) Application of genetic algorithm-PLS for feature selection in spectral data sets. J Chemom 14:643-655

Li J, Gao XB, Jiao LC (2005) A new feature weighted fuzzy clustering algorithm. In Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing 412-420. Springer Berlin Heidelberg

Liang G, Yang L, Kang L et al (2009) Using multidimensional patterns of amino acid attributes for QSAR analysis of peptides. Amino Acid 37:583-591

Liang GZ, Zhou P, Zhou Y et al (2006) New descriptors of amino acids and their applications to peptide quantitative structure-activity relationship. Acta Chim Sin 64:393-396

Lin ZH, Long HX, Bo Z et al (2008) New descriptors of amino acids and their application to peptide QSAR study. Peptides 29:1798-1805

Sandberg M, Eriksson L, Jonsson J et al (1998) New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J Med Chem 41:2481-2491

Sewald N, Jakubke HD (2002) Peptides: chemistry and biology (Vol. 2). Weinheim, Wiley-Vch

Sheridan RP, Feuston BP, Maiorov VN et al (2004) Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. J Chem Inf Model 44:1912-1928

Tian F, Zhou P, Li Z (2007) T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. J Mol Struct 830:106-115

Tomii, K and Kanehisa, M (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng 9:27-36

Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci 22:69-77

Vivencio DP, Hruschka ER, Nicoletti MC et al (2007, April) Feature-weighted k-nearest neighbor classifier. In Foundations of Computational Intelligence, 2007. FOCI 2007. IEEE Symposium on (pp. 481-486). IEEE

Wölfel M, Ekenel HK (2005, September) Feature weighted Mahalanobis distance: improved robustness for Gaussian classifiers. In 13th European Signal Processing Conference

Xu Z, Hou T, Li N, et al (2012) Proteome-wide detection of Abl1 SH3-binding peptides by integrating computational prediction and peptide microarray. Mol Cell Proteomics 11:O111.010389

Yousefinejad S, Hemmateenejad B, Mhedipour AR (2012) New autocorrelation QTMS-based descriptors for use in QSAM of peptides. J Iran Chem Soc 9:569-577

Zhang H, Wang H, Dai Z, et al (2012) Improving accuracy for cancer classification with a new algorithm for genes selection. BMC bioinformatics 13:1-20

Zhou P, Chen X, Wu YQ et al (2010) Gaussian process: an alternative approach for QSAM modeling of peptides. Amino Acid 38:199-212