Journal of Medicinal Chemistry

ARTICLE

pubs.acs.org/jmc

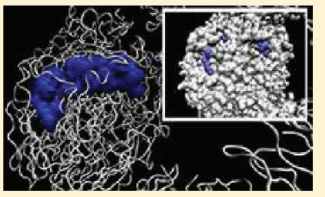
Novel Peptide-Specific Quantitative Structure−Activity Relationship (QSAR) Analysis Applied to Collagen IV Peptides with Antiangiogenic Activity

Corban G. Rivera,[*,†,‡] Elena V. Rosca,[†] Niranjan B. Pandey,[†] Jacob E. Koskimaki,[†] Joel S. Bader,[†,‡] and Aleksander S. Popel[†]

[†]Department of Biomedical Engineering, 613 Traylor Building, Johns Hopkins University, 720 Rutland Avenue, Baltimore, Maryland 21205, United States

[‡]High-Throughput Biology Center, Broadway Research Building, Room 359, Johns Hopkins School of Medicine, 733 N. Broadway Street, Baltimore, Maryland 21205, United States

ABSTRACT: Angiogenesis is the growth of new blood vessels from existing vasculature. Excessive vascularization is associated with a number of diseases including cancer. Antiangiogenic therapies have the potential to stunt cancer progression. Peptides derived from type IV collagen are potent inhibitors of angiogenesis. We wanted to gain a better understanding of collagen IV structure−activity relationships using a ligand-based approach. We developed novel peptide-specific QSAR models to study the activity of the peptides in endothelial cell proliferation, migration, and adhesion inhibition assays. We found that the models produced quantitatively accurate predictions of activity and provided insight into collagen IV derived peptide structure−activity relationships.



BACKGROUND

Excessive vascularization is a hallmark of many diseases including cancer, rheumatoid arthritis, diabetic nephropathy, pathologic obesity, age-related macular degeneration, and asthma. Compounds that inhibit angiogenesis represent potential therapeutics for many diseases. Judah Folkman performed pioneering research in the field of angiogenesis;[1] his work led to the identification of a number of polypeptides with antiangiogenic activity.[2] One of polypeptides called endostatin was derived from the noncollagenous (NC1) domain of collagen XVIII.[3] Work led by Raghu Kalluri resulted in the development of small antiangiogenic peptides from the NC1 domain of collagen IV including canstatin,[4] arrestin,[5] and tumstatin.[6] These collagen IV derived fragments were reviewed in the context of other angiogenesis modulating compounds.[7−9] On the basis of these parent compounds, work in our laboratory identified more than 100 similar peptide sequences from diverse parent proteins throughout the proteome.[10] The set of parent proteins included collagen IV, CXC chemokines, type I thrombospondin domain (TSP-1) containing proteins, serpins, somatotropins, and tissue inhibitors of metalloproteinases (TIMPs). Work carried out in our group experimentally validated in vitro inibition of endothelial cell (EC) proliferation and migration by peptides derived from type IV collagens,[11] thrombospondin domain-containing proteins,[12,13] and CXC chemokines.[14] These studies showed that a large fraction of the peptides have antiangiogenic potential. Subsequently, our laboratory tested some of these peptides in vivo using mouse xenograft models of breast and lung cancer[15,16] and ocular models.[17] The peptides derived from type IV collagen are attractive targets because of their efficacy against multiple angiogenic properties (i.e., endothelial cell proliferation, migration, and adhesion).[18]

A better understanding of the structure−activity relationship of type IV collagen peptides could help us better understand the mechanism of action and produce more active peptides. For many of these peptides, the receptor had not been elucidated. When the receptor is unknown, ligand-based modeling approaches must be used. Examples of ligand-based design methods include pharamcophore modeling[19−22] and quantatitive structure−activity relationship (QSAR)[23−26] analysis. These methods correlate diverse aspects of molecular structure and flexibility with a quantatitive measure of activity. Some work has been done on developing peptide-specific feature sets for QSAR.[27,28] Others make use of position weight matrices to describe a family of peptides.[29] Many of these methods require solving NP-hard[30] problems. That means a polynomial time algorithm is not known for solving these problems. For large data sets, these methods must resort to using inexact approaches and heuristics.

To continue developing the type IV collagen-derived peptides, we aimed to (i) develop techniques for computationally efficient, peptide-specific QSAR analysis, (ii) enable predictions of peptide activity, and (iii) gain a better understanding of the structure−activity relationship of collagen IV derived peptides. In this work, we described several novel peptide-specific QSAR methods that helped us address these aims. We formulated the models using convex optimization in a way that could be solved quickly to global optimality. We used experimentally determined activity data from collagen IV peptides to develop individual models for endothelial cell proliferation, migration, and adhesion. We validated the QSAR models by making activity predictions and

Received: January 31, 2011
Published: August 25, 2011

Table 1. Compound Database[a]

Ref.	Compound	Structure	Proliferation	Migration	Adhesion
0	SP2000	LRRFSTMPFMFCNINNVCNF	67.55 ± 2.58	91.02 ± 0.63	96.25 ± 0.34
2	SP2002	LRRFSTMPFMFGNINNVGNF	25.71 ± 0.78	87.90 ± 0.70	97.80 ± 0.15
4	SP2004	LRRFSTMPFMF*********	2.97 ± 2.62	-14.20 ± 7.85	-2.60 ± 0.32
6	SP2006	LRRFSTMPFMFXNINV****	59.70 ± 4.10	26.55 ± 3.15	95.62 ± 1.22
7	SP2007	LRRFSTMPFMFX********	4.85 ± 1.62	-24.38 ± 5.87	11.04 ± 1.81
8	SP2008	LRRFSTMP**********	7.02 ± 0.69	-20.09 ± 4.67	-0.14 ± 0.00
9	SP2009	**********NINNVXNF	24.25 ± 12.25	31.65 ± 0.85	70.78 ± 0.05
10	SP2010	********FMFXNINNVXNF	28.30 ± 3.39	-16.60 ± 9.83	66.18 ± 1.83
11	SP2011	****STMPFMFXNINNVXNF	13.83 ± 7.64	-24.46 ± 6.89	6.82 ± 2.87
12	SP2012	LRRFSTMPFMFXNINNVXNF	69.72 ± 2.25	96.97 ± 0.19	98.46 ± 0.32
13	SP2013	LNRFSTMPF**********	6.35 ± 4.17	-16.05 ± 0.69	0.32 ± 0.86
14	SP2014	LRRFSTNLPFNLF*******	2.95 ± 3.01	-22.77 ± 1.28	-24.59 ± 15.37
15	SP2015	LRRFSTMPAMFXNINNVXNF	65.40 ± 1.20	99.75 ± 0.25	95.60 ± 0.31
16	SP2016	LRRFSTMPFAFXNINNVXNF	63.60 ± 0.35	99.60 ± 0.40	98.33 ± 1.14
17	SP2017	LRRFSTMPFMAXNINNVXNF	65.10 ± 0.57	99.60 ± 0.40	99.27 ± 0.21
20	SP2020	**********FXNINNVXN*	20.55 ± 0.46	40.35 ± 3.45	77.51 ± 2.36
21	SP2021	**********FXNIN*****	8.17 ± 2.27	-23.68 ± 0.59	3.45 ± 2.25
22	SP2022	LRRFSTMPFMFSNINNVSNF	50.47 ± 0.66	92.56 ± 0.97	97.95 ± 1.27
23	SP2023	LRRFSTMPFMFANINNVANF	48.61 ± 0.53	99.59 ± 0.13	98.19 ± 0.03
24	SP2024	LRRFSTMPFMFININNVINF	73.20 ± 0.78	92.92 ± 1.78	98.58 ± 0.12
25	SP2025	LRRFSTMPFMFTNINNVTNF	59.10 ± 0.85	96.08 ± 0.15	98.69 ± 0.08
27	SP2027	LRRFSTMPFMFVNINNVVNF	62.20 ± 0.64	98.78 ± 0.14	95.12 ± 0.62
35	SP2035	LRRFSTMPFAFININNVINF	46.58 ± 4.23	69.33 ± 0.14	94.80 ± 0.93

[a] A data set of 23 collagen IV derived compounds tested for endothelial cell proliferation (at 100 μM), migration (at 50 μM), and adhesion (at 100 μM) inhibition. The table gives the mean % inhibition for each assay and the standard error of the mean (SEM). The screening of all compounds was done with $n = 2$ and normalized to a vehicle control. We use the single letter X to represent L-α-amino-n-butyric acid (Abu). Peptides **27** and **35** were held out as an external validation set.

performing experiments for an external set of peptides. The activity of the external set of peptides was verified by endothelial cell proliferation, migration, adhesion, and tube formation assays.

RESULTS

Peptide Activity in Vitro Using EC Proliferation, Migration, and Adhesion Assays. This study is based on a libary of 23 collagen IV derived peptides. The founding peptide **0** (SP2000)[10] was found as a homologue of tumstatin[6] in the human proteome. These peptides consisted of a series of truncations and selected amino acid substitutions designed to improve translational potential. In Table 1 we present the activity of the 23 (21 training + 2 external verification) peptides in endothelial cell proliferation (at 100 μM), migration (at 50 μM), and adhesion (at 100 μM) . Peptide concentrations were chosen to provide diversity in activity measurements. All experiments were performed in duplicate, and the result of each experiment was the average of three replicates on the same plate. Activity measurements are given as a percentage of the vehicle control.

Modeling Overview. In Figure 1, we outline the peptide modeling procedure. The methods are based on data that associate

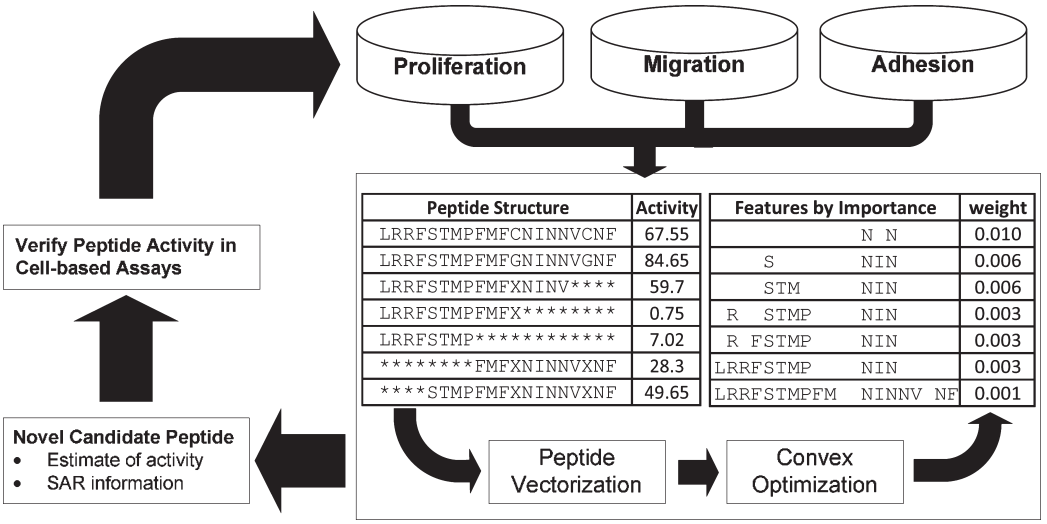


Figure 1. Overview of the peptide optimization framework. The procedure is based on data that associate peptides with an activity score (e.g., endothelial cell proliferation inhibition activity). The peptides are converted into unique sparse vectors. We use convex optimization to select features that differentiate high activity peptides from low performing peptides. The selected features can be used to help understand the structure−activity relationships of the peptides. New peptides can be synthesized based on the SAR information. A feedback loop can be created by adding new experimentally tested peptides to the database.

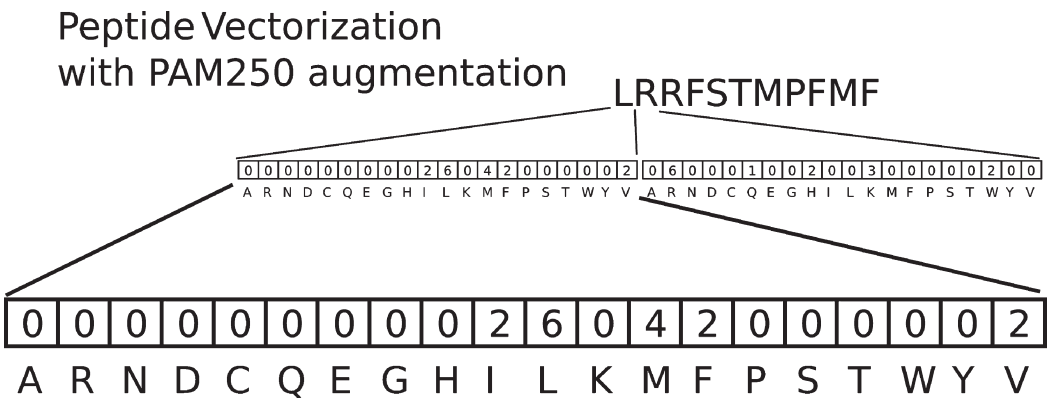


Figure 2. Peptide vectorization. Each peptide is converted into a sparse vector that uniquely maps specific amino acids to positions in the peptide. The mapping is augmented by the PAM250 amino acid substation matrix. PAM matrices are based on the empirical mutation rate of amino acids in evolutionarily related proteins. For example, the figure shows the vectorization of the peptide LRRFSTMPFMF. The first amino acid leucine (L) can mutate to isoleucine (I), methionine (M), phenylalanine (F), and valine (V) at rates greater than expected by chance. The weights assigned to these amino acids are given by log of the odds ratio in the PAM250 matrix. All other amino acids mutate from leucine at a lower rate than expected by chance. As a result, their values are set to zero. The PAM matrix gave us a principled way to associate common amino acids based on their chemical and structural properties.

peptide features with a quantitative activity score (e.g., endothelial cell (EC) proliferation inhibition activity). Peptides are converted into unique sparse vector of features. For example, Figure 2 shows the vectorization of the short peptide LRRFSTMPFMF. In the simplest methodology that we consider, each feature uniquely identifies an amino acid at a single position. We use convex optimization to select features that differentiate highly active and inactive peptides. We formulate the convex optimization objective in a way that can be solved quickly to global optimality.

Peptide-Specific QSAR Method Comparison. We developed four approaches to model the data in Table 1 and learn about the structure−activity relationship of type IV collagen peptides. The approaches were based on the least absolute shrinkage and selection operator (Lasso).[31] The approaches differed in the features that they consider and the weight assigned to training examples. The specific details of these approaches can be found in the section Materials and Methods.

In Table 2, we compared four methods for their ability to predict peptide efficacy. We compared each of these methods to a naive featureless method that always predicted the average activity from the training set. The methods were evaluated on three data sets that measured the ability of peptides to inhibit endothelial cell proliferation (A), migration (B), and adhesion (C). To compare these approaches, we took a leave-one-out cross-validation (LOOCV) approach. The concept of LOOCV is that we use all but a single peptide to train the model. We then use that model to predict the efficacy of the single peptide, which was left out. This allowed us to compute the error between predicted and observed activity measurements. To determine

Table 2. Comparison of Algorithms for Predicting Peptide Efficacy[a]

(A) proliferation model	Lasso	nonlinear Lasso	local Lasso	local nonlinear Lasso	featureless
Lasso		0.004	0.580	0.218	0.003
nonlinear Lasso			0.496	0.815	0.001
local Lasso				0.343	0.018
local nonlinear Lasso					0.007
featureless					
(B) migration model	Lasso	nonlinear Lasso	local Lasso	local nonlinear Lasso	featureless
Lasso		0.034	0.340	0.000	0.000
nonlinear Lasso			0.179	0.651	0.000
local Lasso				0.034	0.000
local nonlinear Lasso					0.000
featureless					
(C) adhersion model	Lasso	nonlinear Lasso	local Lasso	local nonlinear Lasso	featureless
Lasso		0.033	0.205	0.098	0.017
nonlinear Lasso			0.862	0.432	0.004
local Lasso				0.740	0.005
local nonlinear Lasso					0.007
featureless					

[a] We tested five methods for their ability to predict peptide efficacy. A complete description of each method can be found in Materials and Methods. The methods were evaluated on three data sets that measured the ability of peptides to inhibit endothelial cell proliferation (A), migration (B), and adhesion (C). For each method and data set, we compute LOOCV test error. To determine which methods were superior to others, we conducted *t*-tests for all pairs of methods based on their squared test errors. Significantly low test errors indicate better performance. The table gives the *p*-value associated two-tailed paired *t*-test. At the 0.05 level, the naive featureless method had significantly higher error than all other methods. Also, the nonlinear Lasso method had significantly less error than the Lasso method.

which methods were statistically superior to others, we conducted *t* tests for all pairs of methods based on their squared test errors. Significantly low test errors indicate better performance. The table gives the *p*-value associated two-tailed paired *t* test. At the 0.05 level, all of the models had lower error than the naive featureless method. Also, the nonlinear Lasso method had significantly less error than the Lasso method. These results held over all three data sets. On the basis of these results, the rest of the study was performed using nonlinear Lasso. In Figure 3, we show the observed and leave-one-out predictions for each method for all peptides in the data set in endothelial cell proliferation, migration, and adhesion assays. The figure illustrates that no single method had the least error in all trials and that the predictive performance is good even in cases where percent inhibition is negative as seen in the migration and adhesion data sets.

QSAR Analysis for Type IV Collagen Derived Peptides. In the previous sections we make extensive use of leave-one-out cross-validations to estimate generalization error. We concluded from these analyses that nonlinear Lasso had statistically lower generalization error than Lasso. Low generalization error is an indication that the features used in the models may be useful for understanding the structure−activity relationship of type IV collagen peptides.

In this section and unlike the previous sections, we train models for endothelial cell proliferation, migration, and adhesion based on all of the data in Table 1 except for the external validation set consisting of **27** and **35**. The models are structured such that important features receive high weight. The model features (first column) and weights (second column) are given in decreasing order in Table 3. The features are indicated for each row by the change in sequence from the preceding row. The weights were determined using the nonlinear Lasso method (as described in Materials and Methods). We analyze these features for QSAR analysis. This approach gives us a way of indirectly identifying putative pharmacophores for the collagen IV derived peptides.

When multiple amino acids are viable options in a position, they are shown in decreasing order of importance. In the migration model (Table 3C) in the 18th position, L-α-amino-*n*-butyric acid (indicated by X) is preferred with a weight of 0.018 over alanine with a weight of 0.016. The proliferation model (Table 3A) makes it clear that there are important regions on the N-terminus (LRRF) and the C-terminus (NINNVXN). In the adhesion model (Table 3B), the highly weighted asterisks in the 20th position indicate that truncation of the phenylalanine may improve the antiadhesion activity of the peptide. Like the proliferation model, the regions on the N-terminus (LRRF) and C-terminus (NINNVX) are selected. Unlike the proliferation model, the L-α-amino-*n*-butyric acid in the 12th position is one of the most important features for antiadhesion activity. The migration model (Table 3C) highlights the C-terminal (ANINNVXN) as a useful indicator of antimigration activity; however, for full antimigration activity the LRRF sequence is also required. From all three models we found that both the C-terminal sequence LRRF and the N-terminal sequence XNINNVXN are required for full activity.

Structural Association. We examined the structure of peptide **0** as it exists in the native type IV collagen NC1 domain (PDB code 1T60). In Figure 4, we show the conformation of the peptide in the native protein. By computing the solvent accessible surfaces of the protein, we found two exposed regions

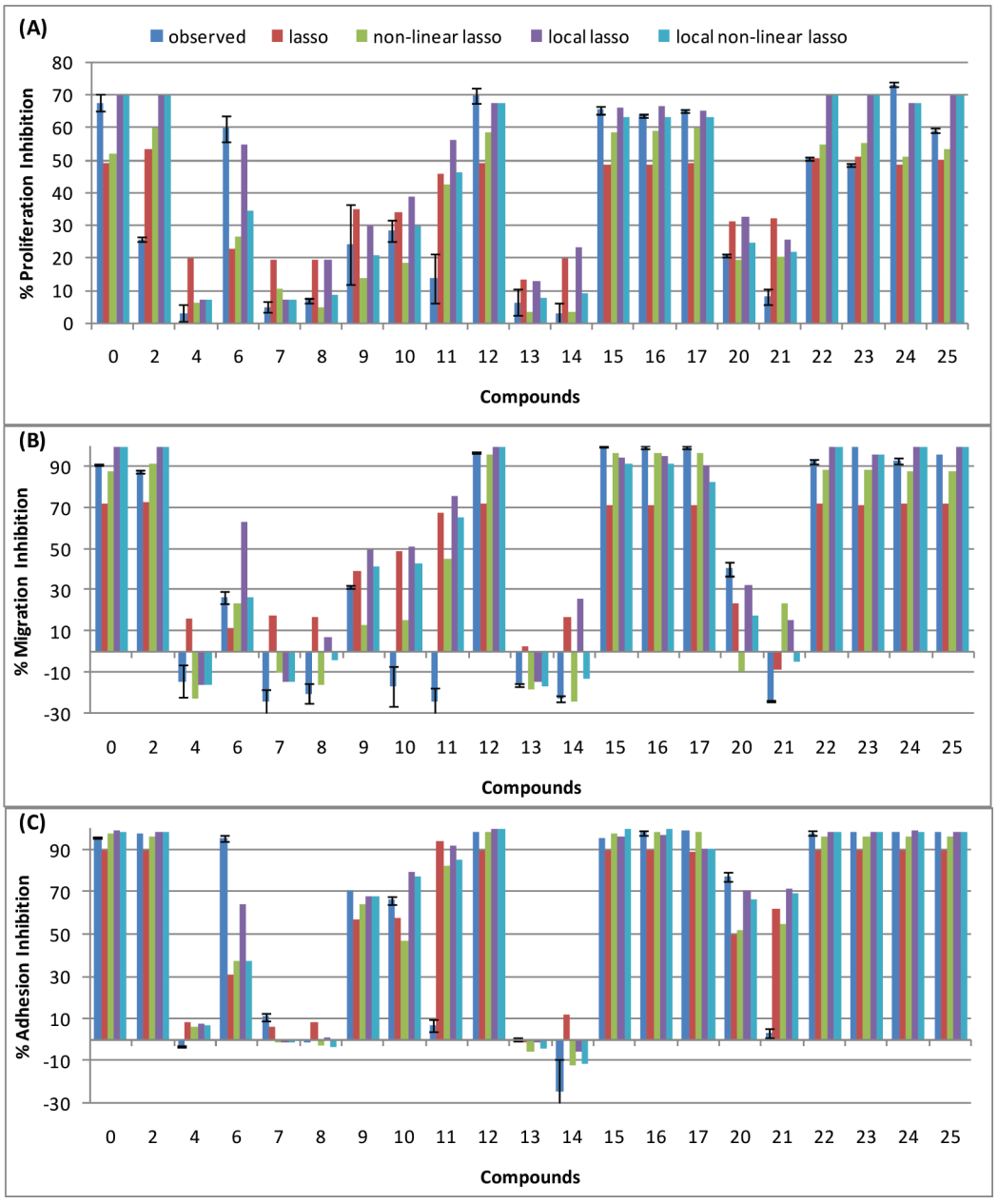


Figure 3. Quantitative predictions of peptide activity using nonlinear Lasso. The observed and predicted activity of the 21 training peptides screened in endothelial cell (A) proliferation, (B) migration, and (C) adhesion assays. Compounds are given in Table 1. Predictions are made using LOOCV to assess the generalization error of the method. Predictions are shown for the four methods described in Materials and Methods. The results imply an average error of between 14% and 20% depending on the assay.

corresponding to the N-terminal (LRR) and C-terminal (INN). These regions correlate with the peptide motifs needed for antiangiogenic activity.

Experimental Model Validation. Two peptides, **27** and **35**, were held out as an external validation set. Models for proliferation, migration, and adhesion were trained using all other peptides from Table 1. On the basis of these models, peptides **27** and **35** were predicted to have similar activity. They were predicted to have 54.15%, 93.35%, and 97.54% proliferation, migration, and adhesion inhibition, respectively. On the basis of the experimentally determined activities given in Table 1 and predicted activities, R^2 values on the external validation set were 0.84, 0.85, and 0.99 for the proliferation, migration, and adhesion models, respectively.[32] From the R^2 values on the external validation set, we could conclude that the models were predictive for anti-angiogenesis phenotypes. In Figure 5, endothelial cell tube formation assays at 100 μM confirmed the potency of peptides **27** (Figure 5C) and **35** (Figure 5D), relative to a vehicle control (Figure 5A) and a weaker peptide **8** (SP2008) (Figure 5B).

DISCUSSION AND CONCLUSIONS

Type IV collagens are basement membrane proteins that are essential for binding cells to the extracellular matrix.[33] Type IV collagen derived peptides have proven to be effective inhibitors of angiogenesis.[34] Using the models trained using the data from

dx.doi.org/10.1021/jm200114f |J. Med. Chem. 2011, 54, 6492–6500

Table 3. Model Details for Proliferation, Migration, and Adhesion Inhibition[a]

(A) Proliferation Inhibition Model		(C) Migration Inhibition Model	
N N	0.009		
NIN	0.009		
NIN	0.009		
L NIN	0.009		
LR NIN	0.009		
LR F NIN	0.009		
LRRF NIN	0.009	* N	0.018
LRRF NIN	0.009	* N N	0.018
LRRF NIN X	0.003	** N N	0.018
LRRF NIN XN	0.003	** NV N	0.018
LRRF NIN VXN	0.003	** NVXN	0.018
LRRF NINNVXN	0.003	** A NVXN	0.016
LRRF P NINNVXN	0.001	** A NVAN	0.016
LRRF MP NINNVXN	0.001	F* * NVAN	0.013
LRRF MP NINNVXN	0.001	R F* * I NVAN	0.013
(B) Adhesion Inhibition Model		RR F* *NI NVAN	0.013
X *	0.034	RRF F* *NI NVAN	0.013
L X I *	0.017	RRF F* *NINNVAN	0.013
LR XNI *	0.017	LRRF F* *NINNVAN	0.013
LRR XNI *	0.017	LRRF F* *NINNVAN	0.011
LRR XNIN *	0.017	LRRF F* *NINNVANF	0.011
LRRF XNIN *	0.017	LRRF F* *NINNVXNF	0.007
LRRF * XNIN N*	0.011	LRRF F* *NINNVXN*	0.006
LRRF ** XNIN N*	0.011	LRRF F* *NINNVXN*	0.006
LRRF*** XNINN N*	0.011	LRRF F* SNINNVXN*	0.006
LRRF*** XNINNV N*	0.011	LRRF F* SNINNVSN*	0.006
LRRF*** XNINNVXN*	0.011	LRRF T F* ININNVSN*	0.002
LRRF**** XNINNVXN*	0.011	LRRFST F* ININNVSN*	0.002
LRRF**** XNINNVXN*	0.011	LRRFST FM *NINNVSN*	0.002
LRRF**M* XNINNVXN*	0.009	LRRFST FMF*NINNVSN*	0.002
LRRF**MP XNINNVXN*	0.009	LRRFS* FMF*NINNVSN*	0.001
LRRF**MP **XNINNVXN*	0.003	LRRFS** FMF*NINNVSN*	0.001
LRRF**MP ***NINNVXN*	0.003	LRRFS***FMF*NINNVSN*	0.001
LRRF**MP ***NINNVXN*	0.003	LRRFS***FMF*NINNVXN*	0.001
LRRF**MPF***NINNVXN*	0.002	LRRF****FMF*NINNVXN*	0.001
LRRF**MPF**XNINNVXN*	0.002	LRRFS***AMF*NINNVXN*	0.001
LRRF**MPF**XNINNVXN*	0.001	LRRFST**AMF*NINNVXN*	0.001
LRRF**MP****NINNVXN*	0.001	LRRFST**AMF*NINNVXN*	0.001
LRRFSTAPFMFXNINNVXNF	weights	LRRFSTAPFMFXNINNVXNF	weights

[a] Feature selection using the nonlinear Lasso method and trained on all of the data from Table 1. The table gives the features and weights selected for the (A) proliferation, (B) migration, and (C) adhesion inhibition models, respectively. The features are indicated for each row by the change in sequence from the preceding row. For each model, pairs of amino acid features (summarized in the first column) were given a weight by a linear model (shown in the second column). Asterisks indicate a preference for a missing amino acid. A full length amino acid sequence (i.e., SP2012) is given at the base of each model for reference.

Table 1, we that found a pair of regions, namely, LRRF at the C-terminus and XNINNVXN at the N-terminus, are needed for full activity. This pair of important regions indicates that secondary structure or multiple binding sites may be important for the endothelial cell proliferation, migration, and adhesion inhibition activity of type IV collagen derived peptides. These results are consistent with those of a previous study on the tumstatin peptide by Eikesdal et al.[35] They found that the mutations to the NINN region resulted in a significant change in EC proliferation inhibition. These results also indicate that truncations to the 20-mer peptide with the exception of the phenylalanine in the 20th position would be detrimental to the activity of the collagen IV derived peptides.

In this article, we describe four novel peptide-specific QSAR approaches. We compared these approaches by testing their ability to predict the outcome of in vitro experiments. The comparison indicated that one approach called nonlinear Lasso had statistically lower generalization error than Lasso (Table 2). We showed the individual predictions made by this approach in Figure 3. We found that the predictions made using the all four approaches were statistically significant compared to a method based on naive predictions. These results gave us confidence in the utility of the peptide-specific QSAR models. We analyzed the features of these models to learn about the structure—activity relationship of collagen IV derived peptides. By analyzing the structure of the collagen IV NC1 domain, we found that the solvent accessible regions of the peptide in the parent protein correlated with the motifs needed for antiangiogenic activity.

MATERIALS AND METHODS

Peptide Data Set. All peptides were synthesized by New England Peptide with at least 95% purity evaluated using both HPLC and MALDI by the manufacturer. Table 1 gives the compound structures in terms of the one letter amino acid codes. Truncated amino acids are indicated by asterisks. The error in the activity measurements was based on two biological replicates each derived from the mean of three technical replicates. The data are shown as percent inhibition relative to a vehicle control. A single dose was selected for each data set that produced a diverse set of activities for the candidate peptides. Proliferation and adhesion measurements were taken at a peptide concentration of 100 μM, while migration measurements were taken with a compound dose of 50 μM.

Cell Culture. Human umbilical vein endothelial cells (HUVEC) were purchased from Lonza and were grown under the manufacturer's recommendation using endothelial basal medium (EBM-2) supplemented with the Bullet Kit (EGM-2, Lonza). Cells of passages 2−7 were used for experiments. Cells were grown at 37 °C in a humidified incubator with 5% CO_2.

Proliferation Assays. Colorimetric WST-1 reagent (Roche, IN) was used to perform the proliferation assays. HUVECs were plated in 96-well plates at a 2000 cells/well density. Peptides at 100 μM in fully supplemented media were added to the adherent cells and incubated for 72 h. WST-1 reagent was added in serum free media for 4 h, and the color intensity was measured at 450 nm with Victor-V plate reader (Perkin-Elmer, MA).

Migration Assay. The effect of the migration inhibition of the peptides on the cells was determined using electrical impedance measurements with a continuous and real time migration assay (RT-CIM, ACEA Biosciences, CA). The top compartment of the CIM plate was coated with fibronectin (20 μg/mL), and 45 000 HUVEC/well were added in either the presence or absence of the peptide at 50 μM. Fully

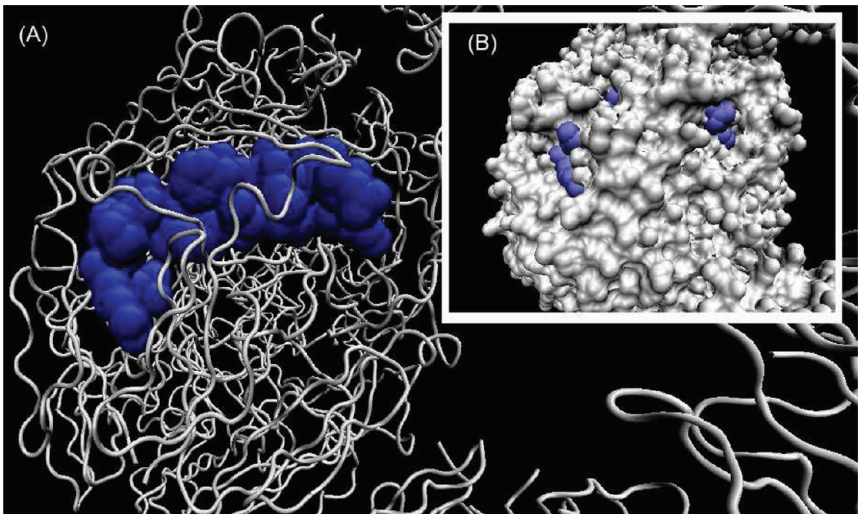


Figure 4. Solvent accessible surfaces of the peptide **0** in noncollagenous (NC1) domain of collagen IV: (A) location of the peptide **0** in the NC1 domain of collagen IV; (B) solvent exposed surfaces of peptide **0**. The regions at the N-terminus and C-terminus are solvent accessible.

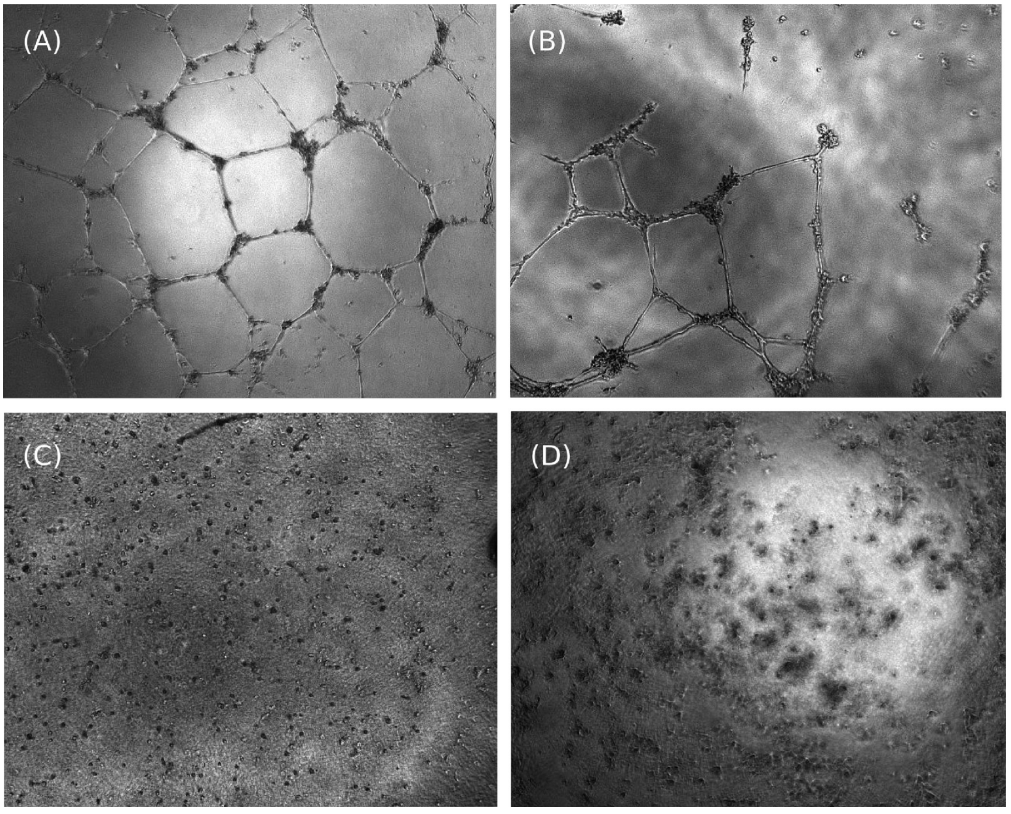


Figure 5. Endothelial tube formation assay. Endothelial cell tube formation assays are useful indicators of angiogenesis potential. (A) Tube formation for the positive control (vehicle control). Entothelial cell tube formation without an added compound. HUVECs form robust tube structures. (B) Endothelial cell tube formation with the addition of 100 μM **8**. The figure shows only partial inhibition of tube structures. (C) 100 μM **27** completely inhibits the formation of tube structures. (D) 100 μM **35** completely inhibits the formation of tube structures.

supplemented medium was added to the bottom compartment serving as chemoattractant. The migration of the cells is measured by the integrated sensors in the bottom side of the porous membrane which divides the two chambers. This technology allows for easy quantification of cell migration by monitoring the cell index (derived from the measured impedances).

Adhesion Assay. The adhesion inhibitory potential of the peptides was also measured using RT-CIM technology. In this instance single compartment E-plates (ACEA, Biosciences,CA) were used in which 25 000 HUVEC/well were plated in the presence or absence of the peptides at 100 μM and the adhesion was measured by the changes in the cell index amplitude for 3 h.

Tube Formation. Tube formation assay was performed by following the published protocol by Arnaoutva et al.[36] Briefly, 96-well plates were coated with Geltrex, reduced growth factor basement membrane matrix (Invitrogen, CA) (50 μL/well), and incubated at 37 °C for 30 min to allow gelation to occur. HUVECs were added to the top of the gel at a density of 15 000 cells/well in the presence or absence of the peptide (100 μM). The positive control included the same amount of solvation vehicle (i.e., DMSO) as the experimental condition. Cells were incubated at 37 °C with 5% CO_2 overnight, and pictures were captured with a CCD Sensicam camera mounted on a Nikon inverted microscope.

Peptide-Specific QSAR Approaches. We took as input a set of peptide sequences along with an experimentally measured efficacy for each peptide. The method returned a model that could be used to predict the efficacy of hypothetical peptides from the same class. The method worked by converting each peptide sequence into an input space of amino acids and positions. Those were the explanatory variables in the peptide-specific QSAR modeling framework. A weight for each feature was learned using non-negative Lasso regression[37] with the peptide efficacies as response variables. The scaling term for the L1-norm regularizer was determined using leave-one-out cross-validation. Despite evaluation of many features, the use of L1-norm regularization allowed the model to avoid overfitting. The convex nature of the optimization problem allowed the method to quickly reach the globally optimal solution without a combinatorial search of input space. The software which was implemented in Matlab using CVX[38] is freely available upon request.

Lasso with an Amino Acid Substation Matrix. Without loss of generality, we describe the method in terms of the 20 common amino acids. Given m peptides of length n, let p_{ij} be amino acid j in peptide i. Let $\mathbf{r}$ be a list of all 20 natural amino acids. Let $\mathbf{S}$ be a 20 × 20 amino acid association matrix; in this study we use the PAM250 matrix[39] such that $\mathbf{S}(a,b)$ gives the association between amino acids a and b. We use the PAM250 matrix as a principled approach to give weight to amino acids with similar biochemical properties. Let $\mathbf{A}$ be an $m \times 20n$ matrix that encodes the amino acid sequences such that

$$\mathbf{A}_{i,j.k} = \mathbf{S}(p_{i,j},r_k) \quad (1)$$

Let $\mathbf{b}$ be a vector of length m representing the activity of each peptide. In this study, the quantitative measure of activity is given by percent endothelial cell proliferation, migration, or adhesion inhibition. Our goal is to learn values in the weight vector $\mathbf{x}$ of length $20n$. The values in the weight vector $\mathbf{x}$ correspond to the relative importance of the features considered in the model. Using this formulation, we solve the standard Lasso objective subject to $\mathbf{x} \geq 0$. Lasso is composed of the least-squares objective regularized by the L1-norm of the weight vector. The parameter λ influences the sparsity of the weight vector $\mathbf{x}$

$$\min ||\mathbf{Ax} - \mathbf{b}||_2 + \lambda||\mathbf{x}||_1 \quad (2)$$

Nonlinear Lasso. In the previous section, we described the linear version of Lasso using only the input space described in $\mathbf{A}$. As an alternative approach, we expand on the input space given in the previous approach to a feature space consisting of pairs of features. Let $\mathbf{A}'$ be an $m \times (20n)^2$ matrix. Although the number of features is large, we use sparse matrices to eliminate unused variables and reduce the problem size. We make use of aggressive regularization to avoid overfitting. The Lagrange multiplier λ is selected automatically by leave-one-out cross-validation. We use the objective from eq 2 except that we make use of $\mathbf{A}'$ and the $\mathbf{x}$ vector is of length $(20n)^2$.

Locally Weighted Methods. We extend both linear and nonlinear Lasso to construct locally weighted variants of both methods. The idea is that we will weight training examples in $\mathbf{A}$ by their proximity to the vectorized peptide $\mathbf{y}$ to be predicted. The intuition is that we prefer to make smaller training errors for points close to the test point $\mathbf{y}$. The weight $\mathbf{w}$ assigned to each training example in $\mathbf{A}$ is given in eq 3.

$$\mathbf{w}_j = \exp[-||\mathbf{y} - \mathbf{A}_j||_2] \quad (3)$$

The weighted objective for the linear version of Lasso is given in eq 4.

$$\min \sum_i w_i(\mathbf{A}_i\mathbf{x} - b_i)^2 + \lambda||\mathbf{x}||_1 \quad (4)$$

Statistical Significance and Cross-Validation. To evaluate the quality of the predictions given by the peptide-specific QSAR approaches, we perform leave-one-out cross-validation. For each of the m peptide examples, we split the examples into a test set containing the ith peptide and a training set containing all other peptides. We use the training set of peptides to obtain the weight vector $\mathbf{x}$. Let $\mathbf{p}_i$ be the vector of length $20n$ that encodes the ith peptide. The predicted activity q_i for the ith peptide is given by

$$\mathbf{q}_i = \mathbf{p}_i^T\mathbf{x} \quad (5)$$

The statistical significance of the predictions is determined by comparing the set of residuals generated using our model predictions with residuals generated using naive model predictions. We test the null hypothesis that the residuals between the observed and predicted values are equal to the residuals between the observed and naive model predictions (i.e., a model that always predicts the mean training efficacy). The alternative hypothesis is that the residuals between the observed and predicted values are less than the residuals between the observed and naive model predictions. We generate a p-value for each model using a one-sided paired t test. We used R^2 as a metric of model performance on the external validation set.

$$R^2 = 1 - \sum_{i=1}^{n} \frac{(y - \hat{y})^2}{(y - \bar{y})^2} \quad (6)$$

In this metric, experimentally observed values y are compared with predicted values $\hat{y}$ relative to the mean observed value from the training set $\bar{y}$.

AUTHOR INFORMATION

Corresponding Author

*Telephone: 410-955-1787. Fax: 410-614-8796. E-mail: cgrivera@jhu.edu.

Author Contributions

C.G.R. designed the method, performed the analysis, and wrote the paper. E.V.R., J.E.K., and N.B.P. performed the in vitro experiments. J.S.B. and A.S.P. motivated investigation of the problem, providing guidance for the analysis and manuscript. All coauthors edited the paper.

Notes

The authors declare no competing interests.

ACKNOWLEDGMENT

The work was supported by NIH Grants R01 CA138264, R01 HL101200, and U54RR020839.

ABBREVIATIONS USED

QSAR, quantitative structure−activity relationship; NC1, noncollagenous domain; CXC, N-terminal cysteine-X-cysteine domain; TSP-1, thrombospondin 1 domain; TIMP, tissue inhibitor of metalloproteinases; EC, endothelial cell; NP-hard, nondeterministic polynomialtime hard; Lasso, least absolute shrinkage and selection operator; LOOCV, leave-one-out cross-validation; HUVEC, human umbilical

vein endothelial cells; L1-norm, one norm; PAM250, 250% point accepted mutation matrix

REFERENCES

(1) Folkman, J. Tumor angiogenesis: therapeutic implications. *N. Engl. J. Med.* **1971**, *285*, 1182–1186.

(2) Folkman, J. Angiogenesis: an organizing principle for drug discovery? *Nat Rev Drug Discovery* **2007**, *6*, 273–286

(3) O'Reilly, M. S.; Boehm, T.; Shing, Y.; Fukai, N.; Vasios, G.; Lane, W. S.; Flynn, E.; Birkhead, J. R.; Olsen, B. R.; Folkman, J. Endostatin: an endogenous inhibitor of angiogenesis and tumor growth. *Cell* **1997**, *88*, 277–285.

(4) Kamphaus, G. D.; Colorado, P. C.; Panka, D. J.; Hopfer, H.; Ramchandran, R.; Torre, A.; Maeshima, Y.; Mier, J. W.; Sukhatme, V. P.; Kalluri, R. Canstatin, a novel matrix-derived inhibitor of angiogenesis and tumor growth. *J. Biol. Chem.* **2000**, *275*, 1209.

(5) Nyberg, P.; Xie, L.; Sugimoto, H.; Colorado, P.; Sund, M.; Holthaus, K.; Sudhakar, A.; Salo, T.; Kalluri, R. Characterization of the anti-angiogenic properties of arresten, an [alpha] 1 [beta] 1 integrin-dependent collagen-derived tumor suppressor. *Exp. Cell Res.* **2008**, *314*, 3292–3305.

(6) Maeshima, Y.; Sudhakar, A.; Lively, J. C.; Ueki, K.; Kharbanda, S.; Kahn, C. R.; Sonenberg, N.; Hynes, R. O.; Kalluri, R. Tumstatin, an endothelial cell-specific inhibitor of protein synthesis. *Science* **2002**, *295*, 140.

(7) Kalluri, R. Basement membranes: structure, assembly and role in tumour angiogenesis. *Nat. Rev. Cancer* **2003**, *3*, 422–433.

(8) Mundel, T. M.; Kalluri, R. Type IV collagen-derived angiogenesis inhibitors. *Microvasc. Res.* **2007**, *74*, 85–89.

(9) Nyberg, P.; Xie, L.; Kalluri, R. Endogenous inhibitors of angiogenesis. *Cancer Res.* **2005**, *65*, 3967.

(10) Karagiannis, E. D.; Popel, A. S. A systematic methodology for proteome-wide identification of peptides inhibiting the proliferation and migration of endothelial cells. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 13775–13780.

(11) Karagiannis, E. D.; Popel, A. S. A theoretical model of type I collagen proteolysis by matrix metalloproteinase (MMP) 2 and membrane type 1 MMP in the presence of tissue inhibitor of metalloproteinase 2. *J. Biol. Chem.* **2004**, *279*, 39105–39114.

(12) Karagiannis, E. D.; Popel, A. S. Anti-angiogenic peptides identified in thrombospondin type I domains. *Biochem. Biophys. Res. Commun.* **2007**, *359*, 63–69.

(13) Karagiannis, E. D.; Popel, A. S. Peptides derived from type I thrombospondin repeat-containing proteins of the CCN family inhibit proliferation and migration of endothelial cells. *Int. J. Biochem. Cell Biol.* **2007**, *39*, 2314–2323.

(14) Karagiannis, E. D.; Popel, A. S. Novel anti-angiogenic peptides derived from ELR-containing CXC chemokines. *J. Cell. Biochem.* **2008**, *104*, 1356–1363.

(15) Koskimaki, J. E.; Karagiannis, E. D.; Rosca, E. V.; Vesuna, F.; Winnard, P. T., Jr.; Raman, V.; Bhujwalla, Z. M.; Popel, A. S. Peptides derived from type IV collagen, CXC chemokines, and thrombospondin-1 domain-containing proteins inhibit neovascularization and suppress tumor growth in MDA-MB-231 breast cancer xenografts. *Neoplasia* **2009**, *11*, 1285–1291.

(16) Koskimaki, J. E.; Karagiannis, E. D.; Tang, B. C.; Hammers, H.; Watkins, D. N.; Pili, R.; Popel, A. S. Pentastatin-1, a collagen IV derived 20-mer peptide, suppresses tumor growth in a small cell lung cancer xenograft model. *BMC Cancer* **2010**, *10*, 29.

(17) Cano Mdel, V.; Karagiannis, E. D.; Soliman, M.; Bakir, B.; Zhuang, W.; Popel, A. S.; Gehlbach, P. L. A peptide derived from type 1 thrombospondin repeat-containing protein WISP-1 inhibits corneal and choroidal neovascularization. *Invest. Ophthalmol. Visual Sci.* **2009**, *50*, 3840–3845.

(18) Rosca, E. V.; Koskimaki, J. E.; Rivera, C. G.; Pandey, N. B.; Tamiz, A. P.; Popel, A. S. Anti-angiogenic peptides for cancer therapeutics. *Curr. Pharm. Biotechnol.* **2011**, *12*, 1101–1116.

(19) Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A. PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 647–671.

(20) Schneidman-Duhovny, D.; Dror, O.; Inbar, Y.; Nussinov, R.; Wolfson, H. J. PharmaGist: a webserver for ligand-based pharmacophore detection. *Nucleic Acids Res.* **2008**, *36*, W223–228.

(21) Güner, O. F., Ed. *Pharmacophore Perception, Development, and Use in Drug Design*; International Universit Line: La Jolla, CA, 1999.

(22) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251–3264.

(23) Blankley, C. Introduction: A Review of QSAR Methodology. In *Quantitative Structure—Activity Relationships of Drugs*; Topliss, J. G., Ed.; Academic Press: New York, 1983.

(24) Dudek, A. Z.; Arodz, T.; Galvez, J. Computational methods in developing quantitative structure—activity relationships (QSAR): a review. *Comb. Chem. High Throughput Screening* **2006**, *9*, 213–228.

(25) Hansch, C.; Leo, A.; Hoekman, D. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.

(26) Kubiny, H. Variable selection in QSAR studies. I. An evolutionary algorithm. *Quant. Struct.—Act. Relat.* **1994**, *13*, 285–294.

(27) Lin, Z. H.; Long, H. X.; Bo, Z.; Wang, Y. Q.; Wu, Y. Z. New descriptors of amino acids and their application to peptide QSAR study. *Peptides* **2008**, *29*, 1798–1805.

(28) Zhou, P.; Chen, X.; Shang, Z. Side-chain conformational space analysis (SCSA): a multi conformation-based QSAR approach for modeling and prediction of protein—peptide binding affinities. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 129–141.

(29) Doytchinova, I. A.; Blythe, M. J.; Flower, D. R. Additive method for the prediction of protein—peptide binding affinity. Application to the MHC class I molecule HLA-A* 0201. *J. Proteome Res.* **2002**, *1*, 263–272.

(30) Finn, P.; Halperin, D.; Kavraki, L.; Latombe, J. C.; Motwani, R.; Shelton, C.; Venkatasubramanian, S. Geometric manipulation of flexible ligands. *Appl. Comput. Geom. Geom. Eng.* **1996**, 67–78.

(31) Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc., Ser. B* **1994**, *58*, 267–288.

(32) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579–586.

(33) Khoshnoodi, J.; Pedchenko, V.; Hudson, B. G. Mammalian collagen IV. *Microsc. Res. Tech.* **2008**, *71*, 357–370.

(34) Maeshima, Y.; Manfredi, M.; Reimer, C.; Holthaus, K. A.; Hopfer, H.; Chandamuri, B. R.; Kharbanda, S.; Kalluri, R. Identification of the anti-angiogenic site within vascular basement membrane-derived tumstatin. *J. Biol. Chem.* **2001**, *276*, 15240–15248.

(35) Eikesdal, H. P.; Sugimoto, H.; Birrane, G.; Maeshima, Y.; Cooke, V. G.; Kieran, M.; Kalluri, R. Identification of amino acids essential for the antiangiogenic activity of tumstatin and its use in combination antitumor activity. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 15040.

(36) Arnaoutova, I.; George, J.; Kleinman, H. K.; Benton, G. The endothelial cell tube formation assay on basement membrane turns 20: state of the science and the art. *Angiogenesis* **2009**, *12*, 267–274.

(37) Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc., Ser. B* **1996**, *58*, 267–288.

(38) Grant, M.; Boyd, S. Graph implementations for nonsmooth convex programs. *Recent Adv. Learn. Control* **2008**, 95–110.

(39) Jones, D. T.; Taylor, W. R.; Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **1992**, *8*, 275.