

Comparative Binding Energy Analysis Considering Multiple Receptors: A Step toward 3D-QSAR Models for Multiple Targets

Marta Murcia,^{†,‡,^} Antonio Morreale,[‡] and Angel R. Ortiz^{*,‡}

Department of Physiology and Biophysics, Mount Sinai School of Medicine, One Gustave L. Levy Place, P.O. Box 1218, New York, New York 10029, and Bioinformatics Unit, Centro de Biología Molecular “Severo Ochoa” (CSIC-UAM), Universidad Autónoma de Madrid, Cantoblanco, 28049 Madrid, Spain

Received March 26, 2006

Comparative binding energy analysis, a technique to derive receptor-based three-dimensional quantitative structure–activity relationships (3D-QSAR), is herein extended to consider both affinity and selectivity in the derivation of the QSAR model. The extension is based on allowing multiple structurally related receptors to enter the **X**-matrix employed in the derivation of the structure–activity model. As a result, a single model common to all of them is obtained that considers both intra- and inter-receptor affinity differences for a given congeneric series. We applied the technique to a series of 88 3-amidinophenylalanines, binding to thrombin, trypsin, and factor Xa (fXa). A single predictive regression model for the three receptors involving 202 complexes, with a leave-one out (LOO) cross-validated Q^2 of 0.689, was obtained, and selectivity requirements were investigated. We find that total or partial occupancy of any of the three main pockets in the binding site (D-site, P-site, and the rim of the S1-site) leads to higher affinity across the family. However, the fact that thrombin can make stronger interactions in the P-site, as a result of its exclusive 60-loop, makes of this site a specificity pocket for this thrombin. Occupancy of the D-site leads to more active inhibitors toward fXa for the same reason, but the model does not highlight strongly the D-box because inhibitors are too short to fully occupy it. Negative charge density in the neighborhood of position 88 (a Lys insertion in thrombin) is found to be a determinant for thrombin recognition. These results were consistent with previous studies on selectivity in the thrombin/trypsin/fXa system.

Introduction

A drug should discriminate its molecular target from other biomolecules present in the organism. Not only is the affinity for the desired target of relevance, but also the adequate level of cross-reactivity with other proteins, particularly those in the same protein family. Fulfilling selectivity requirements has been found to be of primary importance in a number of cases, including protein kinases,¹ matrix-metalloproteinases (MMPs),² serine proteases,^{3,4} or nuclear receptors.⁵ Because most proteins in the human genome are organized in families as a result of the processes of gene duplication and divergence, this situation is likely to be commonplace with a majority of the new post-genomic targets. As a result, a paradigm shift in drug discovery to the consideration of gene families as a whole, as opposed to individual targets, is starting to emerge.⁶

Selectivity considerations become particularly prominent in the lead optimization phase, when the molecular scaffold needs to be tuned to fulfill constraints other than simply binding to the target.⁷ The availability of the three-dimensional (3D) structures of the proteins in the family can be very helpful to fine-tune the interactions enabling the desired levels of specificity. For example, inhibitors that target highly conserved side chains or protein-backbone atoms are much more likely to be nonspecific than those that make strong interactions with nonconserved residues. Unfortunately, existing methods to estimate binding free energies using macromolecular 3D structures in docking and virtual screening are in many instances too simplistic or insensitive to model ligand selectivity. Some

forms of free energy calculations, such as those based on MM-PBSA or MM-GBSA,⁸ have been documented to provide some encouraging results in the study of selectivity,^{9,10} but they are in most situations too slow to efficiently explore chemical diversity. Recent studies suggest that applying the MM-PBSA energy function to a single, relaxed complex structure might be a feasible strategy to reduce the computational burden,¹¹ the generality of these observations remains to be established.

When activities of a representative set of chemical variations of the basic scaffold are available, it is often more beneficial to resort to the use of three-dimensional quantitative structure–activity relationships (3D-QSARs) that allow the derivation of “tailored” scoring functions. CoMFA¹² or CoMSIA¹³ are two popular 3D-QSAR methods. Both have been used to analyze pairwise selectivity using either the difference or ratio between biological activities (expressed as $-\log K_i$) of a ligand series with respect to two different receptor types as a dependent variable. Examples include serine proteases,^{14,15} matrix metalloproteinases,¹⁶ nuclear receptors,¹⁷ glycine/NMDA and AMPA receptors,¹⁸ or protein kinases.¹⁹ However, this type of analysis has two important shortcomings: first, imposes limitations in multiple receptor comparisons, as they can only be obtained through multiple pairwise analyses; second, no direct use is made of the interactions between the ligand and the protein. More rigorous methods, based on the extended linear response approach (ELR), have also been proposed to study selectivity. In ELR, the average interaction energy of a ligand in water and in the binding site is computed by means of Monte Carlo or molecular dynamics simulations. These energies, together with additional structural descriptors, such as the number of hydrogen bonds formed in the binding process, the change in the hydrophobic surface of the inhibitor upon binding, and the like, enter a multiple linear regression equation to predict affinity. Recently, Tominaga and Jorgensen²⁰ studied the ability of ELR

* Corresponding author. Tel.: 34-91-497-2376. Fax: 34-91-497-4799. Email: aro@cbm.uam.es.

[†] Mount Sinai School of Medicine.

[‡] Universidad Autónoma de Madrid.

[^] Current address: Department of Physiology and Biophysics, Weill Medical College of Cornell University, New York, New York 10021.

to model the binding of 148 inhibitors, belonging to various chemical series, to three different protein kinases. A satisfactory model involving the three receptors, with a Q^2 of 0.67 and four descriptors, could be derived, which underscores its potential to study problems on selectivity. A difficulty with this method, as with the MM-PB(GB)SA, is its computational burden.

Comparative binding energy (COMBINE) analysis²¹ is also a member of the 3D-QSAR family of techniques, but in contrast to CoMFA or CoMSIA, an explicit use is made of the interactions between ligand and protein, and, unlike ELR, only static structures are considered. The key idea is that a simple expression for the differences in binding affinity of a series of related ligand–receptor complexes can be derived by using multivariate statistics to correlate experimental data on binding affinities with components of the ligand–receptor interaction energy computed from energy-minimized 3D structures. COMBINE analysis has been used with success in a number of cases.^{22–25} It has also been demonstrated that regression models derived with COMBINE analysis are suitable for fast virtual screening of compound databases.²⁶

The potential role of COMBINE analysis in modeling selectivity was initially examined by Wang and Wade,²⁴ who as a part of their work analyzed a series of sialic acid and benzoic acid analogues binding to the N2 and N9 subtypes of neuraminidase, having about 50% sequence identity. A predictive joint model for the N2+N9 subtypes, involving 39 complexes, could be developed. Here, we extend these findings and present fully automated computational approaches considering an arbitrary number of receptors of the same protein family and aimed to the direct incorporation of selectivity in lead optimization. The method consists of structural alignment, ligand docking, interaction energy decomposition, and statistical modeling based on COMBINE analysis. As an example, we apply it to a series of 88 benzamidine derivatives (Figure 1) binding to three different serine proteases: thrombin, trypsin, and factor Xa (fXa)¹⁴ and previously studied using CoMSIA by Böhm et al.¹⁴

Serine proteases²⁷ play a key role in a number of diseases.²⁸ Modest changes in sequence and shape of their substrate binding sites confer to this class of enzymes a wide variety of biological functions. For instance, while thrombin and fXa are prominent players in the blood-clotting cascade, trypsin is an enzyme excreted by the pancreas to aid in the digestion of nutrients. A number of the clotting factors in the blood-clotting cascade are inactive serine proteinase zymogens that are proteolytically activated by serine proteinases further up in the cascade. Inhibition at different levels of the cascade has a distinct impact, resulting in varying therapeutic profiles for specific inhibitors. Consequently, there is interest in the design of more selective inhibitors of these enzymes to minimize side effects and to enhance their bioavailability.⁴ Insights from computational studies can be of help to design more specific inhibitors.

Results and Discussion

Docking of the ligands (see Table 1 for a description) in the binding site of the three proteins produced binding modes with the expected set of interactions between ligands and proteins (see Figure 2B for an example). Not all docking experiments were successful, however. Out of the 264 (88 × 3) docking trials attempted, only 202 models successfully fulfilled the distance criteria employed to ensure that correct binding modes enter the analysis (59 complexes in the case of thrombin; 74 in the case of trypsin; and 69 for fXa, see Materials and Methods). In the accepted set of complexes, many of the observed

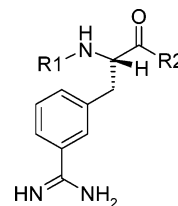


Figure 1. General structure of the 3-amidinophenylalanine¹⁴ series studied in this work (details of the specific ligands can be found in Table 1 and Table 1 of Supporting Information).

interactions are known to involve selectivity determinants (see Figure 2B and Figure 4). Thus, ligands interact with the S1 pocket, where all residues are conserved except for the A190S substitution in trypsin, a trypsin determinant. Similarly, ligands interact with the P-pocket, where the most striking difference between thrombin and the other two proteases, the insertion loop Tyr60A–Pro60B–Pro60C–Trp60D, is found. There are also interactions of some ligands with the “aromatic box” in the D-pocket, formed by residues Phe174, Tyr99, and Trp215, considered the most prominent structural difference between factor Xa and the rest. This set of 202 models was subjected to COMBINE analysis (protocol described in Figure 3) in different conditions (see Materials and Methods).

COMBINE analysis of these complexes yielded statistically significant models in most cases and for a majority of conditions (Table 2 of the Supporting Information). For the case of the screened coulombic potentials–implicit solvent model (SCP-ISM) electrostatic model, a summary of the results is in Table 2. Here, model refers to the type of COMBINE analysis in each case, labeled by the set of receptors entering the analysis. N refers to the total number of complexes that are employed in each one of the analyses. For instance, model Thr+fXa+Trp means that a COMBINE analysis was carried out to explain the affinity differences of the series considering all three receptors at once. Therefore, the total number N of complexes employed in the analysis is in this case 202 (59 + 69 + 74). As can be observed, significant models were obtained for thrombin–fXa–trypsin ($Q^2 = 0.689$), thrombin–fXa ($Q^2 = 0.736$), and fXa–trypsin ($Q^2 = 0.621$), and, to a less extent, for thrombin–trypsin ($Q^2 = 0.486$) and for the single receptor models of thrombin ($Q^2 = 0.439$) and trypsin ($Q^2 = 0.335$). On the other hand, derivation of a predictive model was unsuccessful for fXa inhibition alone ($Q^2 = -0.040$). The effect of the treatment of the electrostatic interactions on the resulting COMBINE analysis models is summarized in Table 3 using the thrombin–fXa–trypsin case (the three-receptor model) as an example (see Table 2 of the Supporting Information for a full account). As can be observed, the main benefit of SCP-ISM is a slight simplification of the regression models, yielding the peak in Q^2 at a lower dimensionality than the standard treatments. This seems to be a general trend (see Table 2 of the Supporting Information), suggesting that slightly more robust models are generated with SCP-ISM. For this reason, the rest of our discussion will be focused only on the SCP-ISM-based COMBINE models.

Coming back to Table 2, a somewhat unexpected conclusion from the analysis of this table is that the use of multiple receptors leads to improved regression models. In fact, while the Q^2 value of the three-receptor model is rather satisfactory, the values for the one-receptor models are poorer than the ones obtained by Böhm et al.¹⁴ using CoMSIA or CoMFA. These authors reported Q^2 values of 0.757, 0.752, and 0.594 (CoMSIA) and 0.687, 0.629, and 0.374 (CoMFA) for thrombin, trypsin, and factor

Table 1. Compound Description, Experimental, and Predicted (LOO) Activities for the Three-Receptor COMBINE Regression Models^a

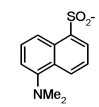
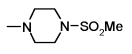
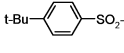
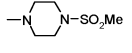
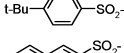
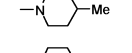

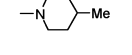
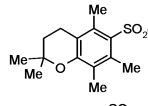
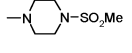

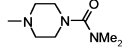

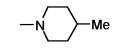

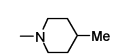

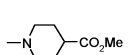

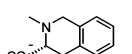

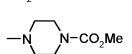

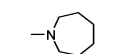

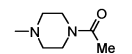
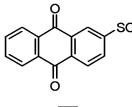
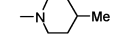
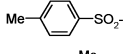
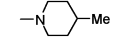
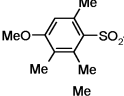
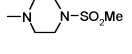
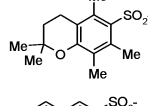
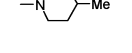

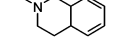

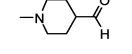

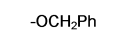

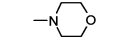

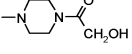

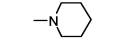

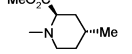

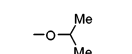
Num	Compound description			Experimental p <i>K</i> _i			Predicted (crossvalidated) p <i>K</i> _i		
	R1	R2	Charge	Thr	Fxa	Trp	Thr	Fxa	Trp
1			+1	8.38	5.41	6.77	7.05	4.88	6.65
2			+1	8.37	5.17	6.80	8.03	4.85	6.40
3			+1	8.30	4.92	6.70	7.24	4.59	6.16
4			+1	8.21	4.39	6.85	7.02	4.54	6.15
5			+1	-	4.13	-	-	5.31	-
6			+1	-	4.62	6.77	-	5.06	6.33
7			+1	7.85	4.85	6.20	7.41	4.04	5.89
8			+1	-	4.38	6.20	-	4.11	5.58
9			+1	7.77	4.37	7.44	7.83	4.86	6.58
10			0	-	4.38	6.89	-	4.11	4.98
11			+1	7.72	4.11	7.70	6.69	4.87	5.98
12			+1	7.68	4.59	6.26	7.39	4.57	5.62
13			+1	-	-	6.85	-	-	6.07
14			+1	7.59	5.64	7.13	8.40	5.01	6.16
15			+1	7.59	4.15	6.28	7.05	4.40	5.94
16			+1	7.50	-	-	6.67	-	-
17			+1	-	-	6.14	-	-	6.10
18			+1	7.43	-	6.59	6.39	-	6.24
19			+1	7.43	4.72	6.66	7.59	4.60	6.49
20			+1	7.38	5.66	6.28	7.49	4.37	6.14
21			+1	7.38	4.80	6.68	6.93	4.51	6.23
22			+1	7.24	4.46	5.96	7.75	5.05	6.14
24			+1	7.19	4.42	6.48	7.14	4.49	6.23
27			+1	7.02	4.24	5.66	7.38	5.19	6.18
28			+1	6.96	5.46	5.85	5.16	4.25	5.65

Table 1 (Continued)

Num	Compound description			Experimental pK_i			Predicted (crossvalidated) pK_i		
	R1	R2	Charge	Thr	Fxa	Trp	Thr	Fxa	Trp
29			0	-	4.27	5.35	-	4.70	5.25
30			+1	6.92	5.64	5.82	6.60	4.64	6.00
31			+1	6.92	4.33	5.40	6.73	4.73	6.13
32			+1	6.82	4.75	6.41	6.80	4.63	6.32
34			+1	6.80	4.12	6.80	7.55	4.86	6.84
35			+1	-	4.08	6.00	-	3.88	5.53
38			+1	6.64	4.77	6.22	6.87	4.38	6.20
40			0	6.59	4.42	6.20	6.64	4.30	5.94
41			+1	6.55	5.60	5.60	5.82	4.80	5.44
42			+1	6.55	4.77	6.92	6.31	4.17	6.69
43			+1	6.50	4.89	5.92	6.62	4.96	5.99
44			+1	6.47	3.75	5.44	7.27	4.74	6.46
45			+1	6.47	4.82	5.92	6.57	4.23	5.72
46			+1	6.46	4.12	5.66	6.21	4.18	5.58
47			+1	6.38	5.50	-	7.49	4.91	-
48			+2	6.30	4.68	6.66	7.14	4.45	6.30
49			+1	6.29	4.80	6.37	6.45	4.60	6.56
50			0	6.24	4.36	6.24	7.40	4.43	5.52
51			+1	6.20	4.70	6.00	6.98	4.67	6.16
52			+1	6.18	3.96	5.09	5.87	4.70	6.01
55			0	5.96	4.36	6.36	6.33	4.47	6.33
56			+1	5.92	4.19	4.85	6.23	4.44	5.78
58			+1	-	5.12	7.10	-	4.70	7.27
59			+1	-	4.24	5.10	-	5.05	5.58
60			0	5.54	3.89	4.80	6.76	4.62	6.07

Table 1 (Continued)

Num	Compound description			Experimental p <i>K</i> _i			Predicted (crossvalidated) p <i>K</i> _i		
	R1	R2	Charge	Thr	Fxa	Trp	Thr	Fxa	Trp
63		-NHMe	+1	5.24	4.59	4.60	5.69	4.72	5.55
64			+1	5.21	3.44	4.80	5.63	3.55	5.66
66	H-		+2	4.89	3.00	4.54	6.48	3.50	4.32
67			+2	-	4.66	6.00	-	4.71	7.17
68		-NHMe	+1	-	3.72	3.85	-	4.80	5.30
69	H-		+2	4.57	3.64	4.54	4.82	3.02	4.02
70		-CO ₂ ⁻	0	4.52	3.89	3.93	5.44	4.45	5.51
71			0	4.46	4.39	4.51	6.41	4.45	5.52
72		-NHMe	+1	-	-	3.00	-	-	5.19
73			+1	8.48	4.66	6.72	7.69	4.25	6.13
74			+1	7.89	5.51	6.59	6.39	4.68	6.03
75			+1	7.59	5.01	6.50	8.20	4.36	5.94
76			+1	7.52	4.66	6.22	6.51	4.69	6.30
77			+2	7.44	4.52	5.89	7.52	4.57	6.47
78			+1	7.28	4.30	6.36	6.78	4.54	6.05
79			+1	7.16	4.34	5.72	7.36	5.07	6.11
80			+1	6.77	4.28	6.15	6.57	4.31	5.90
81			+1	6.59	4.48	6.51	6.36	4.66	6.21
82			+1	-	-	6.01	-	-	4.72
83			+1	6.52	5.07	6.80	6.51	4.42	6.76
84			+1	6.28	4.57	7.57	7.31	4.85	6.81
85			+1	6.28	4.44	5.75	6.21	4.94	6.14
86			+1	-	4.60	7.64	-	4.63	7.05
88		-NHMe	+1	4.75	4.40	4.34	5.39	4.44	5.39

^a Num, number of the compound as it appears in Böhm et al.¹⁴ Dashed entries correspond to failures in the docking protocol. No complex could be derived in these cases, Compounds **26**, **37**, and **61** were found missing in proof. They can be found in Table 1 of the Supporting Information.

Xa, respectively. Although an exact quantitative comparison is not possible, because the set of molecules entering the analysis were not the same in both studies (some molecules failed in our docking procedure, see Materials and Methods), and the

differences may indicate possible inaccuracies in the modeled complexes, as COMBINE analysis is more sensitive to the details of the docked conformation than CoMSIA and CoMFA. A second observation is that inclusion of several receptors leads

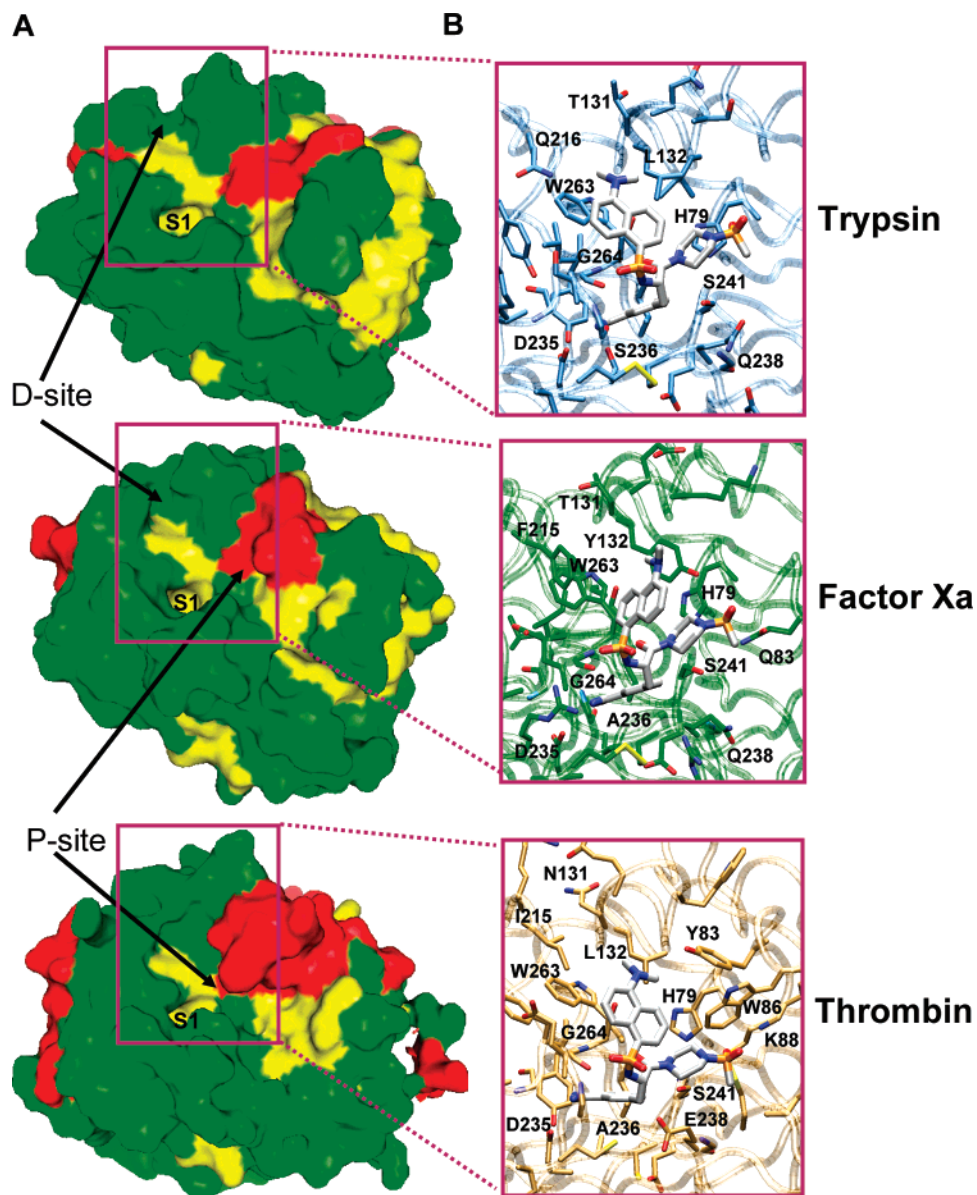


Figure 2. (A) Trypsin-like serine-protease members studied in this work. The solvent accessible surface of the three members—trypsin, fXa and thrombin—is shown, colored by the secondary structure of the corresponding residue (red, helix; yellow, strand; green, coil). Three main subsites in the ligand binding site are highlighted: S1 subsite, formed by a deep, narrow pocket where the conserved Asp189 forms a salt bridge with positively charged moieties; the D (distal) pocket, lined mainly by aromatic residues and particularly evident in factor Xa, to a minor extent for trypsin, and absent in thrombin; and the P (proximal) pocket, particularly evident in thrombin due to the insertion loop Tyr60A–Pro60B–Pro60C–Trp60D, to a minor extent in factor Xa and absent in trypsin. Figure created with PyMol;⁵⁹ (B) the insets show, for each enzyme, the predicted binding mode of compound 1 (see Table 1), with the most relevant interactions labeled. Residue numbers correspond to positions on the multiple sequence alignment (see Figure 4).

to a higher complexity in the regression models. The correlation between the cross-validated or fitted predictions with the observed pK_i values are shown in Figure 5 for the 202 complexes used in the three-receptor model. Excellent predictions are obtained in this case, and no outliers are apparent. A point to note is that, based on the Q^2 values, the single receptor results presented here are significantly worse than the ones obtained previously by our groups analyzing by COMBINE analysis the interaction of a set of 3-amidino-1*H*-indole-2-carboxamide analogues with factor Xa,²⁶ but we emphasize that the series employed in this paper and the one used in the 2004 paper are different. While in the present case we study a series of 3-amidinophenylalanine derivatives, in the 2004 paper, a set of 3-amidino-1*H*-indole-2-carboxamide analogues were employed. Being that the two series are chemically different, with different ranges of activity against fXa, different sets of

interactions, and so on, there is no reason to expect a regression model of similar quality in both cases.

A pertinent question, in light of these results, is to ask whether the two- and three-receptor models appear to give good Q^2 values simply because they distinguish between receptors, or by contrast both intra- and inter-receptor variabilities are being captured. To answer this question, we computed the SDEP values to be expected if they would only model differences among receptors, that is, considering that the COMBINE models were simply assigning to each molecule the average activity of the molecules in the training set observed with each receptor. The average activities of the molecules in the series for each receptor are 6.764 (thrombin), 4.544 (factor Xa), and 6.050 (trypsin). We then computed the expected SDEP values of this so-called “null model”. Results are shown in Table 2, in the SDEP-null column. In all cases, the observed SDEP values

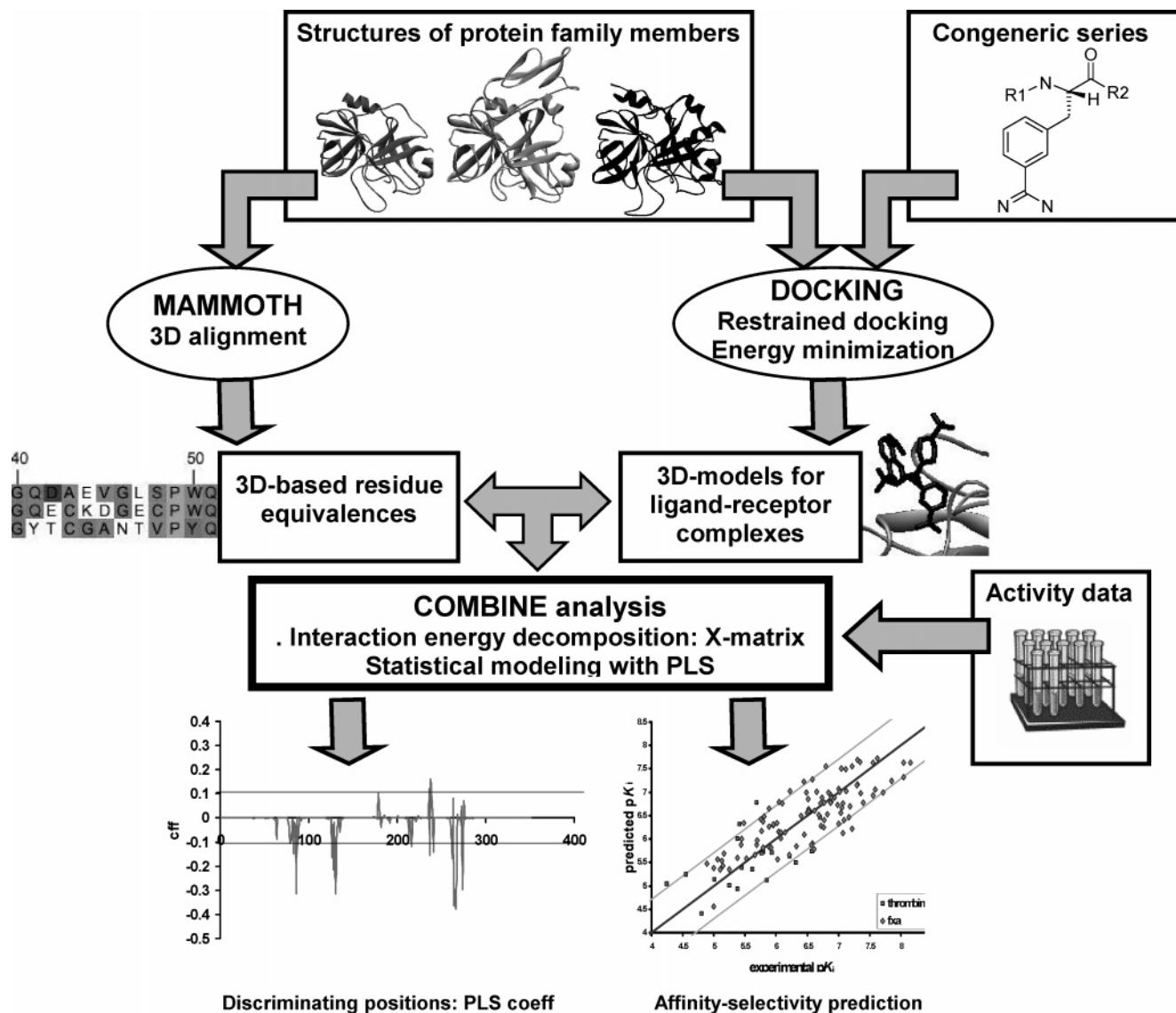


Figure 3. Modeling selectivity with COMBINE analysis.²¹ Initially, the structures in the family (either experimental or modeled structures) are structurally aligned to create a multiple structural alignment.⁵¹ Each ligand in the set is then docked²⁶ to each one of the models. A COMBINE analysis is then carried out, using the previously computed alignment to place the interactions with the different proteins in register. A model of the predicted activity of each ligand in each protein is then generated, which can be assessed by comparison with experimental data. The coefficients of the resulting PLS⁵⁸ model highlight key discriminating interactions. See text for details.

are significantly smaller than the SDEP values from the null models, demonstrating that the COMBINE analyses do model both inter- and intra-receptor variance.

Once the statistical significance of the models is established, we turn our attention to their interpretation. The partitioning scheme in COMBINE analysis allows for a dissection of statistically significant contributions to differences in affinity. These are shown in the form of partial least squares (PLS) coefficients for each interaction type, and can be found in Figure 6 for the three-receptor model. All pockets known to be required for inhibition and selectivity (Figure 2) are selected. This can also be noted by the close match between pockets and important residues labeled in the alignment (Figure 4B). Focusing on the VDW coefficients, three main regions are selected (see also Figure 7): the 83–88 region, corresponding to the thrombin-exclusive loop, forming the P-site; the 126–131 region, participating in the D-site; and the 262–267 region, at the entrance of the S1 pocket, including the interaction with the backbone through residue Gly216 (position 264). The electrostatic coefficients show a large peak for position 238, corre-

sponding to the Glu/Gln 192 substitution (Glu in thrombin; Gln in fXa); for position 236, in the S1 pocket and, therefore, in the neighborhood of Asp189, where thrombin has an Ala residue and fXa has a Ser residue; and to a less extent position 88, where thrombin presents the insertion of a Lys residue (Figure 6). Finally, the desolvation coefficients are found in the region adjacent to the thrombin-exclusive loop (position 79 and 241) and in the D-box (position 132). Figure 7 suggests that both affinity and selectivity can be largely explained by VDW interactions in the P-site, the entrance of the S1 pocket, and, to a lesser extent, in the D-site, the electrostatic interaction with E192Q, and a few, less relevant, desolvation energies in the D-box and the neighborhood of the P-site.

The thrombin versus fXa selectivity features can now be rationalized on the light of the three-receptor model in Figure 5. In general, total or partial occupancy of the three main pockets (D-site, P-site, and the rim of the S1-site) leads to higher affinity across the family. However, the fact that thrombin can make stronger interactions with the ligands in the P-site, as a result of the thrombin-exclusive 60-loop, makes of this site a specific-

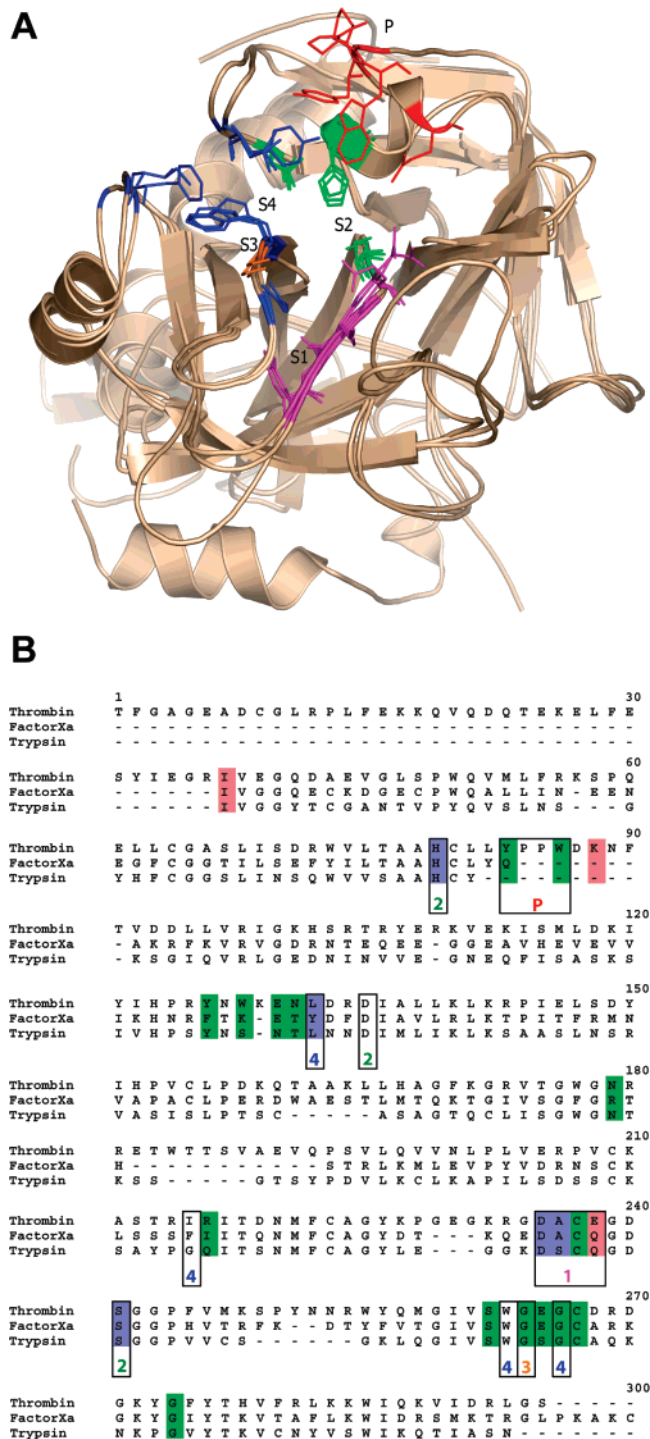


Figure 4. (A) Structural alignment of the three proteins as obtained with MAMMOTH-mult.⁵¹ Graphics created with PyMol.⁵⁹ Pockets in the active site are labeled and they correspond to the boxes in Figure 4B. The implied sequence alignment: (B) structure-based sequence alignment of thrombin, trypsin, and fXa. The last 51 positions are not shown. Alignment positions colored are those found significant to explain the affinity differences, with the coloring scheme according to Figure 6. Boxes correspond to the known selectivity sites (see Figure 2 and the main text). Box labeling refers to the different subsites (1, S1; 2, S2; 3, S3; 4, S4 or D-site; P, P-site). See also the main text.

ity pocket for thrombin. Ligands with hydrophobic moieties occupying the P-pocket should be selective toward thrombin. Occupancy of the D-site leads to more active inhibitors toward fXa for the same reason; although the regression coefficients indicate that placing a hydrophobic moiety in this site increases

Table 2. Summary of the Selected Regression Models^a

model	<i>N</i>	LVs	<i>R</i> ²	<i>Q</i> ²	SDEC	SDEP	SDEP-null
Thr+fXa+Trp	202	7	0.771	0.689	0.588	0.686	0.826
Thr+fXa	128	6	0.842	0.736	0.536	0.693	0.772
Thr+Trp	133	5	0.636	0.486	0.609	0.724	0.945
fXa+Trp	143	5	0.723	0.621	0.560	0.655	0.752
Thr	59	7	0.779	0.439	0.462	0.735	
fXa	69	1	0.152	-0.040	0.485	0.537	
Trp	74	1	0.428	0.335	0.692	0.746	

^a Abbreviations: model, type of COMBINE analysis carried out, according to the number and type of receptors employed; *N*, total number of complexes employed in the derivation of the model; LVs, number of latent variables in the PLS model; *R*², squared correlation coefficient ($R^2 = 1 - [\sum(y_{\text{exp}(t)} - y_{\text{fit}(t)})^2] / [\sum(y_{\text{exp}(t)} - \langle y_{\text{exp}(t)} \rangle)^2]$, where $y_{\text{fit}(t)}$ corresponds to the value of the quantity fitted with the model for complex *i*; $y_{\text{exp}(t)}$ is the experimental value for that quantity and complex, and $\langle y_{\text{exp}(t)} \rangle$ is the average experimental value of the quantity for all *n* complexes); *Q*², LOO squared correlation coefficient (the equivalent of *R*² in LOO cross-validation); SDEC, standard deviation of errors of correlation ($\text{SDEC} = \{[\sum(y_{\text{exp}(t)} - y_{\text{fit}(t)})^2] / n\}^{1/2}$); SDEP, standard deviation of errors of cross-validated prediction (the equivalent of SDEC calculated for LOO cross-validation).

the affinity of the ligand for both thrombin and fXa, the effect is more important for fXa, because this enzyme has stronger interactions with the inhibitors in this site (not shown). However, the model does not highlight strongly the influence of all important residues in the D-box on activity. It only highlights position 132 (see Figures 4B and 6), corresponding to the Tyr residue of the D-box in fXa (see also Figure 2). Probably the reason for this is that the inhibitor series is more selective toward thrombin, and the substitutions are too short and do not fully occupy the D-box. Regarding electrostatics interactions, the analysis indicates that negative charge density in the neighborhood of position 88 (a Lys insertion in thrombin) should be a determinant for thrombin recognition, while fXa selectivity is enhanced by ligands able to harbor negative charge density in the neighborhood of position 238, where thrombin has a Glu residue and fXa has a Gln residue.

As an illustrative example, a detailed view of the interactions for derivative 1, the most potent thrombin inhibitor in the series, can be observed docked into the different enzymes in Figure 2B. Figure 7 shows the same molecule docked onto the thrombin surface, which has the residues selected in Figure 6 mapped following the same color code. A common element of the three complexes is the benzamidino moiety of the ligand inserted in the S1-site. Some subtle differences in binding are apparent, however. The amidino group forms a salt bridge with the side chain of Asp189 (position 235). The distance from the CG atom of Asp189 to the carbon atom of this amidino group is 4.0 Å in thrombin, whereas the insertion of a phenylalanine residue makes this distance slightly larger for trypsin (4.3 Å) and fXa (4.8 Å). In all cases, the terminal amidino group is further stabilized by contacts with the carbonyl oxygen of Gly219 (266) and, additionally, for the trypsin case, by a hydrogen bond with the side chain of Ser190 (position 236, Ala in thrombin and fXa). The inhibitor glycine spacer forms two hydrogen bonds with Gly216 (position 264). One of the oxygen atoms of the sulfonyl group of the inhibitor is also hydrogen bonded to the NH of the Gly219 backbone, while the other remains oriented away from the binding site toward bulk water. More significant differences are found in the disposition of the piperazine and naphthyl moieties. First, while both in thrombin and in trypsin complexes the naphthyl system of the ligand occupies the D-pocket, the presence of an Ile215 in thrombin, instead of the Gly residue found in trypsin, allows for a more notched hydrophobic site and, consequently, for more hydrophobic

Table 3. Effect of the Electrostatic Treatment on the Three-Receptor (Thr+fXa+Trp) COMBINE Model^a

ϵ	desol	LVs																			
		1		2		3		4		5		6		7		8		9		10	
		R^2	Q^2	R^2	Q^2	R^2	Q^2	R^2	Q^2	R^2	Q^2	R^2	Q^2	R^2	Q^2	R^2	Q^2	R^2	Q^2	R^2	Q^2
4	none	0.46	0.44	0.48	0.45	0.50	0.46	0.66	0.59	0.70	0.63	0.74	0.65	0.77	0.67	0.79	0.68	0.80	0.68	0.81	0.69
	PB	0.47	0.44	0.49	0.45	0.52	0.48	0.67	0.62	0.71	0.64	0.74	0.65	0.76	0.67	0.78	0.67	0.80	0.68	0.81	0.68
	ISM	0.46	0.44	0.49	0.45	0.52	0.47	0.64	0.57	0.70	0.62	0.73	0.63	0.75	0.65	0.78	0.66	0.80	0.67	0.81	0.68
d_{ij}	none	0.37	0.31	0.51	0.42	0.63	0.54	0.66	0.58	0.69	0.62	0.73	0.64	0.75	0.64	0.77	0.64	0.79	0.66	0.80	0.66
	PB	0.39	0.34	0.56	0.48	0.66	0.58	0.68	0.61	0.70	0.62	0.73	0.64	0.75	0.63	0.77	0.64	0.78	0.64	0.80	0.64
	ISM	0.37	0.32	0.53	0.43	0.64	0.53	0.67	0.58	0.72	0.64	0.73	0.64	0.75	0.62	0.77	0.64	0.78	0.63	0.79	0.64
ISM	none	0.55	0.54	0.64	0.61	0.69	0.64	0.73	0.67	0.74	0.68	0.75	0.69	0.75	0.69	0.76	0.68	0.76	0.67	0.77	0.66
	PB	0.53	0.52	0.63	0.60	0.66	0.62	0.70	0.63	0.73	0.65	0.74	0.67	0.75	0.68	0.76	0.68	0.76	0.68	0.77	0.67
	ISM	0.40	0.37	0.62	0.58	0.66	0.63	0.68	0.64	0.70	0.64	0.73	0.66	0.74	0.67	0.75	0.68	0.76	0.68	0.76	0.67
	ISMres	0.48	0.46	0.63	0.60	0.68	0.65	0.71	0.66	0.75	0.67	0.76	0.68	0.77	0.69	0.78	0.68	0.78	0.68	0.79	0.65

^a The R^2 and Q^2 values (see Table 2 for definitions) for the 10 first latent variables (LVs) of three-receptor models obtained with different electrostatics models are shown. The different models differ in the dielectric treatment ($\epsilon = 4$; $\epsilon = d_{ij}$; or $\epsilon = D(r)$, the sigmoidal distance dependent screening function of SCP-ISM^{56,57}) and the way to include desolvation terms (none, no desolvation terms are included; PB, ligand and receptor global electrostatic desolvation energies are computed solving the Poisson–Boltzmann equation⁵⁴ and are introduced in COMBINE analysis, as previously described;²⁵ ISM, ligand and receptor global desolvation terms are computed with the SCP-ISM model and included as external variables; ISMres, as before, but receptor desolvation is partitioned in residue contributions). See Table 2 of the Supporting Information for similar results with the other receptor combinations studied in this paper.

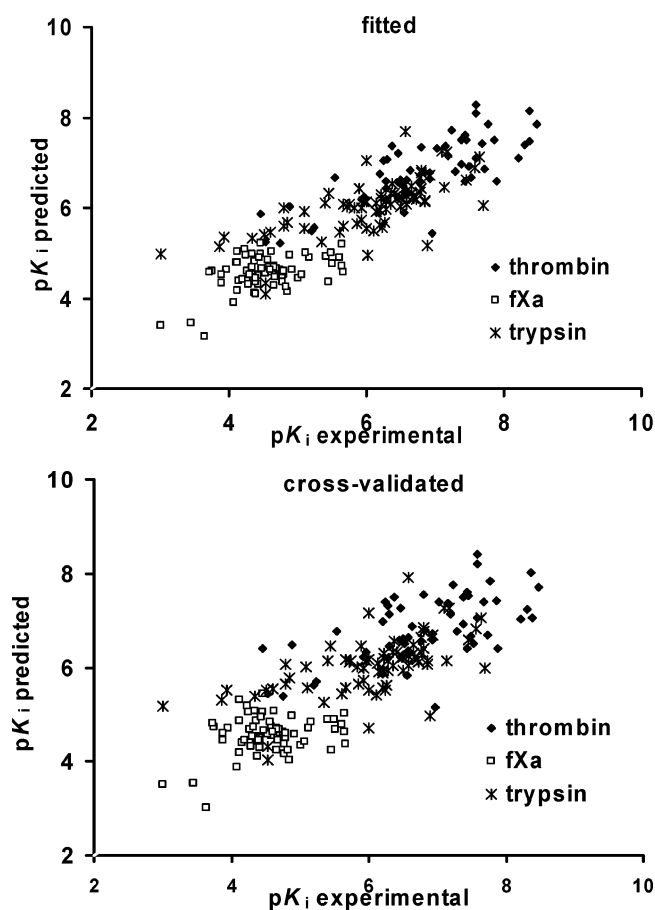


Figure 5. (A) Fitted and (B) cross-validated correlations between observed and predicted activities for the 88 compounds in the 3 proteins (Trp+Thr+fXa model, 202 complexes, see also Table 1).

interactions. On the other hand, the naphthyl group cannot occupy the aromatic cage in fXa and, as a result, remains partly solvent exposed. Second, the piperazine moiety is buried in the thrombin P-site, allowing for strong contacts with the catalytic site (Figure 2b), whereas it remains open to solvent both in trypsin and fXa.

Overall, these results are consistent with previous computational studies^{14,15,29,30} on selectivity in the thrombin/trypsin/fXa system. These studies have highlighted the importance of establishing hydrophobic and steric interactions involving the

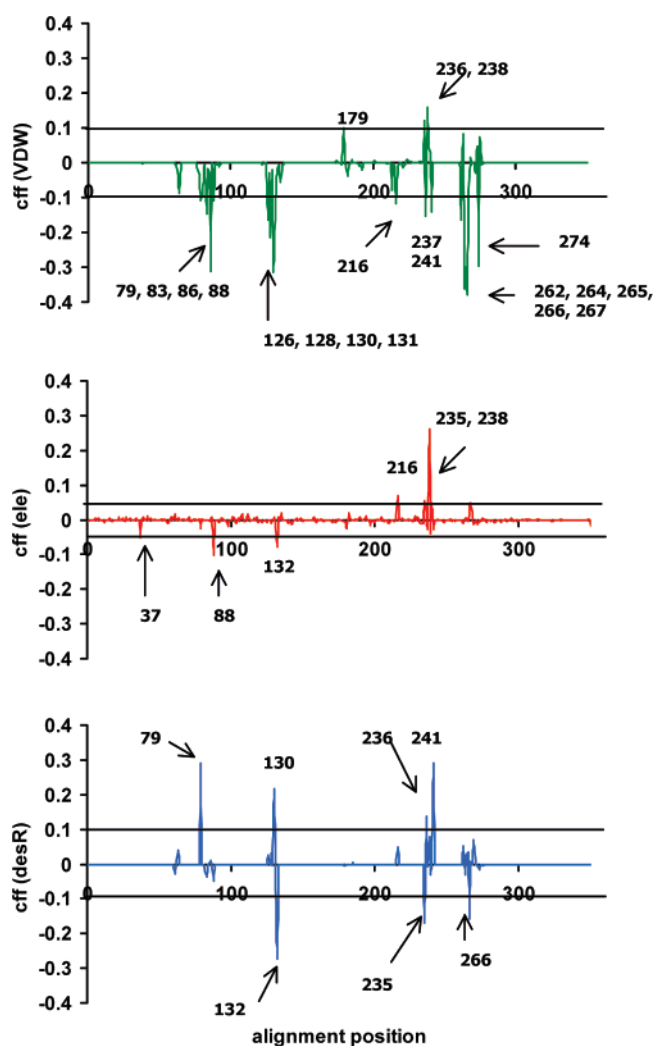


Figure 6. PLS regression coefficients as a function of alignment position. VDW coefficients are in green; electrostatics in red; and residue-based desolvation electrostatic energies in blue. The alignment corresponds to the one found in Figure 4. Labeled are the residue positions considered to provide significant coefficients to explain affinity differences.

thrombin-exclusive 60-loop (P pocket) and the area between this loop and the catalytic triad in the design of thrombin-selective inhibitors. Likewise, the so-called Lys 60f in the S1'

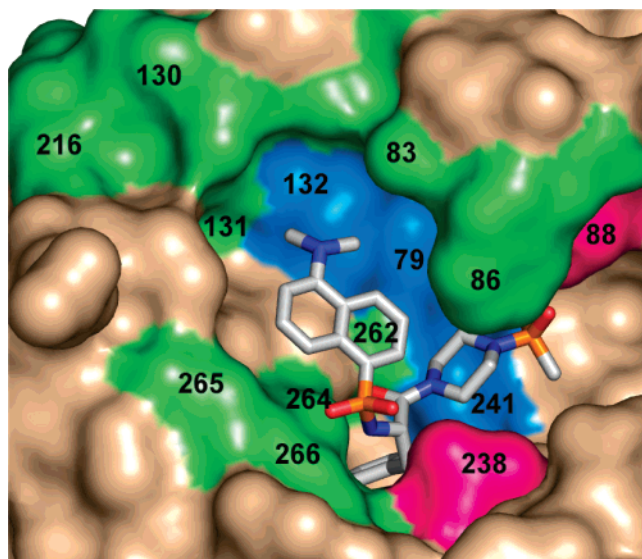


Figure 7. Ligand binding site with the residues selected in Figure 6 mapped onto the thrombin surface. Coloring scheme also according to Figure 6. The complex shown corresponds to thrombin and derivative 1, the most potent thrombin inhibitor in the series. Figure generated with PyMol.⁵⁹

subpocket (primed side of the thrombin active site) has also been implicated in thrombin selectivity.³¹ On the other hand, the incorporation of hydrophobic interactions within the D-pocket in the generation of fXa specific inhibitors is well-recognized. As it is, the interaction of negative charges with the Glu/Gln192 position (position 238 in our alignment) are thought to create unfavorable interactions with the Glu residue in thrombin. This interaction, together with the interaction with glycines 216 and 219, seem to condition a different orientation for the entry of the benzamidine moiety in the S1 subpocket and impose a different interaction pattern with the catalytic triad, which has also been implicated in the rationalization of ligand selectivity. This may be the reason for the emergence of significant PLS coefficients for the conserved positions 237 and 239.

Conclusions

This work constitutes a step forward toward structure–activity models for entire protein families and the incorporation of selectivity in the lead optimization of congeneric series. We present quantitative models of affinity prediction for ligand series across a family of receptors and, as a byproduct, we obtain insights on selectivity determinants from the analysis of the resulting statistical models. There are at least two reasons to prefer such an approach of a simple, energetic, and geometric analysis of representative complexes. First, such dual analysis is not possible, in general, only with plain energetic and structural analysis, and second, it provides information about the reliability of the quantitative predictions through the Q^2 and SDEP values.

Other 3D-QSAR models have tried to take into account selectivity correlations, usually by comparing individual single receptor-based models for the same series of ligands binding to different proteins^{14,18,32–37} or, at most, considering as dependent variables affinity ratios^{38–40} or differences¹⁴ between pairs of proteins. Besides the limitations in the number of receptors to analyze, these approaches may suffer from propagation errors associated with the use of ratios or differences. To our knowledge, this is the first report of a fully automated method where an arbitrary number of members of a given

protein family can be considered in the generation of a single structure–activity model common to all of them, and taking every affinity data as an individual entry in the correlation, for an arbitrary number of ligands.

A pertinent question is the size of the training set required for the analysis. Here, a set of 88 ligands has been employed. It is, however, possible that a smaller number is enough to reach our conclusions, as experience indicates that in most applications COMBINE analysis has been successfully applied with datasets of significantly smaller size.^{22–25} We have used the whole set of 88 ligands for two main reasons: first, because this allows a direct comparison with the results obtained with CoMFA for this same dataset;¹⁴ second, because the use of a large number of ligands is the best way to demonstrate the robustness of the statistical models developed.

Integration of computational methods such as the one presented here with results from high-throughput small molecule–target interaction maps⁴¹ able to provide binding profiles for large numbers of chemically diverse compounds and large number of targets, combined with the phenotypes elicited by these compounds in biological systems, should facilitate the development of inhibitors with appropriate specificity profiles.

Materials and Methods

Materials. The general chemical scaffold of the 88 ligands is schematically shown in Figure 1. Structural variations of the parent structure, present at positions R1 and R2, as well as their experimentally determined biological activities toward the three proteins (pK_i values) were taken from Böhm et al.¹⁴ They are found in Table 1, as well as in Table 1 of the Supporting Information. The affinities toward thrombin and trypsin spread over a satisfactory range, covering 4.0 and 4.7 logarithmic units, respectively, whereas in the case of factor Xa a variation over 3.0 logarithmic units falls close to the lower limit required for COMBINE analysis. Molecules were initially built with Insight-II.⁴²

Crystal structures 1ets,⁴³ 1pph,⁴⁴ and 1hcg⁴⁵ of thrombin, trypsin, and fXa, respectively, were selected from the Brookhaven Protein Data Bank (PDB)⁴⁶ and are shown in Figure 2A. For consistency, these correspond to the same set of structures employed by Böhm et al.,¹⁴ where the receptor structure was indirectly accounted for in the ligand alignment process prior to COMFA/CoMSIA analysis. Crystal structures 1ets and 1pph correspond to the structure of thrombin and trypsin, respectively, complexed with 3-TAPAP (compound number 45 in the series), while the apo form is used for fXa (pdb entry 1hcg), as the corresponding complex with 3-TAPAP is not available. Nevertheless, crystallographic studies with different sets of inhibitors have shown that in these enzymes only small changes in the binding pocket upon binding are observed.

Their binding site can be divided in several subsites (Figure 2A): the deep hydrophobic S1 pocket, where a conserved Asp (Asp189) forms a salt bridge with positively charged moieties; the catalytic triad, or S2 pocket, formed by residues His57, Asp102, and Ser195; the S3 binding subsite, consisting of Gly216 (position 264); the thrombin insertion loop, Tyr60A–Pro60B–Pro60C–Trp60D (positions 83–86), which forms the hydrophobic P (proximal) pocket; and finally the hydrophobic distal S4 region (also called D pocket), lined by residues 99, 174, Trp215, and Gly217 (only for fully conserved residues, residue names are provided).

Parametrization. Structural water molecules in the protein were removed when present and, as a rule, all carboxylic acid groups were ionized, while all basic amino, amidino, and

guanidino groups were protonated. Protein atoms were described by the PARM99⁴⁷ AMBER force field. As ligands are concerned, the benzamidine moiety and basic amino functional groups were protonated, while amides and primary and secondary amino groups adjacent to aromatic portions were treated as uncharged. All carboxylate groups were considered deprotonated. ESP partial charges⁴⁸ were derived with MOPAC⁴⁹ using the AM1 Hamiltonian⁵⁰ starting from the molecular conformation built in Insight-II. AMBER atom types were automatically assigned to the ligands by our in-house docking program, as described.²⁶

Model Building. The procedure is outlined in Figure 3. First, a multiple structural alignment of the three structures was carried out using a multiple alignment version⁵¹ of the MAMMOTH program⁵² to establish structure-based residue equivalences among the three enzymes (Figure 4). This yielded a total of 351 final positions for the matrix of the three superimposed proteins. In a second step, our previously described docking algorithm²⁶ was used to model the interaction of the ligands binding with the three receptor cavities. As in Böhm et al.,¹⁴ the ligand parent structure was fixed according to the bound geometry of 3-TAPAP (R1 = *p*-toluenesulfonyl, R2 = piperidino) in 1pph. The rotameric-based torsional search accounted only for the additional rotatable bonds presented in the R1 and R2 substituents of the ligand structure (Figure 1). A harmonic penalty term was incorporated to the docking scoring function to restrain to 5 Å the interaction distance of the benzamidine group in the ligand with Asp189 (numbering scheme follows chymotrypsin). Similarly, restraints of 3.2 Å for hydrogen bonds with Gly216 were employed between the amidic nitrogen atom of the ligand sulfonamide group and the carbonyl oxygen atom of the mentioned glycine and between the amidic oxygen atom of the ligand and the nitrogen atom of the Gly216 backbone NH group. A force constant of 5 kcal mol⁻¹ Å⁻² was employed. The lowest energy complexes were energy minimized with AMBER 7.0⁵³ as described.²⁶ The distance cutoff of 5 Å between the carboxylic carbon of Asp189 and the amidine group carbon of the ligand was used to filter out geometrically incorrect docking modes prior to COMBINE analysis. A final set of 202 complexes in total resulted (59 complexes in the case of Iets; 74 in the case of 1pph; and 69 for 1hcg) and were subjected to COMBINE analysis.

Energy Calculations. Once the ligand–receptor models are derived, in COMBINE analysis an interaction matrix (the so-called **X**-matrix) summarizing the interaction energy of each ligand with each residue in the protein is derived.²¹ Van der Waals (VDW) and electrostatics interactions are typically employed. The original electrostatic models employed in the COMBINE analysis were based on simple forms of the Coulomb equation. Later, it was shown that more sophisticated electrostatics treatments, such as those based on numerically solving the Poisson–Boltzmann (PB) equation,⁵⁴ could lead to improved regression models.²⁵ However, due to the computational burden, these models cannot be employed in virtual screening calculations. Generalized Born (GB) models⁵⁵ have emerged in recent years as a convenient compromise between computational efficiency and theoretical soundness. Here, besides the basic dielectric models for the computation of electrostatic energy contributions regularly in COMBINE analysis²¹ (Table 2 of the Supporting Information), some statistical models were derived that included electrostatic interactions calculated with a new GB-like⁵⁵ model, the screened SCP-ISM.^{56,57} The SCP-ISM electrostatic binding free energy has the following functional form (Morreale et al., in press):

$$\Delta G_{\text{LR}}^{\text{ele},s} = 332 \sum_{i=1}^{N_L} \sum_{j=1}^{N_R} \frac{q_i q_j}{D(r_{ij}) r_{ij}} + \frac{332}{2} \sum_{i=1}^{N_L+N_R} q_i^2 \left[\left(\frac{1}{D(R_i^c) R_i^c} - \frac{1}{D(R_i^u) R_i^u} \right) + \left(\frac{1}{R_i^u} - \frac{1}{R_i^c} \right) \right]$$

The first term on the right-hand side of the equation corresponds to the screened Coulombic interaction between the charges of the ligand and the receptor, while the second corresponds to the electrostatic desolvation free energy of both ligand and receptor. N_L and N_R are the number of atoms in the ligand and receptor, respectively; q_i is the charge of the i th atom; r_{ij} is the distance between atoms i and j , and R_i corresponds to the Born radius of atom i in the complexed (R_i^c) or in the uncomplexed (R_i^u) form. Finally, $D(r)$ is a sigmoidal distance-dependent screening function. The functional forms for R and $D(r)$ can be found in the original papers by Hassan et al.^{56,57} When distances are in Å and charges in electron units, $\Delta G_{\text{LR}}^{\text{ele}}$ is given in kcal mol⁻¹. The model has been parametrized in our laboratory for its use in the study of the interactions between small organic molecules and proteins. A full account of the performance will be presented in a forthcoming publication (Morreale et al., in press). The main advantage of SCP-ISM with respect to the PB model employed before in COMBINE analysis is that it affords computational speed while allowing a rigorous account of the solvent effect in the evaluation of protein–ligand interactions. In addition, it permits decomposing the desolvation energy of the receptor in residue contributions. These were incorporated into the **X**-matrix (351 new additional variables were added, see below). For comparative purposes, regression models were also derived using the standard Coulombic method as well as incorporating desolvation free energies of ligand and receptor by numerically solving the PB equation, as described.²⁵ However, due to its novelty and improved statistical properties (Table 2 of the Supporting Information), our discussion will be focused on the results obtained with SCP-ISM. Statistics for the optimal models in these conditions are shown in Table 1. Models are further discussed in the Results and Discussion sections. All tested models are in Table 2 of the Supporting Information.

Chemometrics Analysis. The chemometrics analysis was carried out with a multiple receptor adapted version of our in-house program COMBINE,²¹ considering the set of 202 complexes and their corresponding affinity data as dependent variables (Table 1 and Supporting Information). The MAMMOTH alignment (Figure 4B) was used to build a position-based **X**-matrix of energy contributions. Lennard–Jones and electrostatics ligand–receptor interaction energies per residue were computed as usual, but introduced in a global **X**-matrix according to the MAMMOTH alignment. Gaps entered the matrix with a zero value. In the final matrix, each complex was then described by 351 intermolecular electrostatic energy variables and 351 intermolecular VDW energy variables, giving a total of 702 unscaled x -variables. In addition, different conditions were tested with different dielectric models and the introduction of additional external variables (desolvation energies for ligand and receptor). The variables were used directly in the analysis without further pretreatment. The y -variable was assigned as $\text{p}K_i$ toward the appropriate receptor. Similarly, models considering all possible two receptor combinations were derived: thrombin–trypsin (133 complexes), thrombin–fXa (128 complexes), and trypsin–fXa (143 complexes). Finally, one-receptor-based analyses were also developed. All the

different conditions for the PLS⁵⁸ analyses carried out, and the corresponding statistics from 1 to 10 latent variables regression models are found in Table 2 of the Supporting Information. In all cases, leave-one-out cross-validation was employed to evaluate the quality of the models.

Acknowledgment. The datasets are available from the corresponding author (aro@cbm.uam.es). M.M. is grateful to Fundación Ramón Areces for a postdoctoral fellowship. Research at CBMSO has been supported by grants from MEC (BIO2001-3745, BIO2005-0576, and GEN2003-206420-C09-08), Comunidad de Madrid (GR/SAL/0306/2004 and 200520M157), and the CSIC intramural program (PIF2005, project CAR) and by an institutional grant from Fundación Ramón Areces. Generous allocation of computer time at the Barcelona Supercomputer Center is gratefully acknowledged.

Note Added after ASAP Publication. This manuscript was released ASAP on September 23, 2006, with an error in author affiliation information. The correct version was posted on September 25, 2006.

Supporting Information Available: Compound description, receptor, experimental, and predicted (LOO) activities of the multireceptor and the corresponding one-receptor based COMBINE regression models shown in Table 1 can be found in Table 1 of the Supporting Information. The different conditions and results for the derivation of COMBINE models studied in this paper, Trp+Thr+fXa, Thr+fXa, Thr+Trp, fXa+Trp, and the individual receptor-based ones are in Table 2 of this section (up to 10 latent variables). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Bain, J.; McLauchlan, H.; Elliott, M.; Cohen, P. The specificities of protein kinase inhibitors: an update. *Biochem. J.* **2003**, *371*, 199–204.
- Matter, H.; Schudok, M. Recent advances in the design of matrix metalloprotease inhibitors. *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 513–535.
- Bruncko, M.; McClellan, W. J.; Wendt, M. D.; Sauer, D. R.; Geyer, A. et al. Naphthamide urokinase plasminogen activator inhibitors with improved pharmacokinetic properties. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 93–98.
- Walker, B.; Lynas, J. F. Strategies for the inhibition of serine proteases. *Cell Mol. Life Sci.* **2001**, *58*, 596–624.
- Coghlan, M. J.; Elmore, S. W.; Kym, P. R.; Kort, M. E. The pursuit of differentiated ligands for the glucocorticoid receptor. *Curr. Top. Med. Chem.* **2003**, *3*, 1617–1635.
- Caron, P. R.; Mullican, M. D.; Mashal, R. D.; Wilson, K. P.; Su, M. S. et al. Chemogenomic approaches to drug discovery. *Curr. Opin. Chem. Biol.* **2001**, *5*, 464–470.
- Pirard, B. Computational methods for the identification and optimization of high quality leads. *Comb. Chem. High Throughput Screening* **2004**, *7*, 271–280.
- Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S. et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889–897.
- Rizzo, R. C.; Toba, S.; Kuntz, I. D. A molecular basis for the selectivity of thiazazole urea inhibitors with stromelysin-1 and gelatinase-A from generalized born molecular dynamics simulations. *J. Med. Chem.* **2004**, *47*, 3065–3074.
- Laitinen, T.; Kankare, J. A.; Perakyla, M. Free energy simulations and MM-PBSA analyses on the affinity and specificity of steroid binding to antiestradiol antibody. *Proteins* **2004**, *55*, 34–43.
- Kuhn, B.; Gerber, P.; Schulz-Gasch, T.; Stahl, M. Validation and use of the MM-PBSA approach for drug discovery. *J. Med. Chem.* **2005**, *48*, 4040–4048.
- Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Recent advances in comparative molecular field analysis (CoMFA). *Prog. Clin. Biol. Res.* **1989**, *291*, 161–165.
- Klebe, G.; Abraham, U.; Mietzner, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **1994**, *37*, 4130–4146.
- Bohm, M.; Sturzebecher, J.; Klebe, G. Three-dimensional quantitative structure–activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. *J. Med. Chem.* **1999**, *42*, 458–477.
- Bhongade, B. A.; Gouripur, V. V.; Gadad, A. K. 3D-QSAR CoMFA studies on trypsin-like serine protease inhibitors: a comparative selectivity analysis. *Bioorg. Med. Chem.* **2005**, *13*, 2773–2782.
- Matter, H.; Schwab, W. Affinity and selectivity of matrix metalloproteinase inhibitors: a chemometrical study from the perspective of ligands and proteins. *J. Med. Chem.* **1999**, *42*, 4506–4523.
- Wolohan, P.; Reichert, D. E. CoMFA and docking study of novel estrogen receptor subtype selective ligands. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 313–328.
- Baskin, I. I.; Tikhonova, I. G.; Palyulin, V. A.; Zefirov, N. S. Selectivity fields: comparative molecular field analysis (CoMFA) of the glycine/NMDA and AMPA receptors. *J. Med. Chem.* **2003**, *46*, 4063–4069.
- Naumann, T.; Matter, H. Structural Classification of Protein Kinases Using 3D Molecular Interaction Field Analysis of Their Ligand Binding Sites: Target Family Landscapes. *J. Med. Chem.* **2002**, *45*, 2366–2378.
- Tominaga, Y.; Jorgensen, W. L. General model for estimation of the inhibition of protein kinases using Monte Carlo simulations. *J. Med. Chem.* **2004**, *47*, 2534–2549.
- Ortiz, A. R.; Pisabarro, M. T.; Gago, F.; Wade, R. C. Prediction of drug binding affinities by comparative binding energy analysis. *J. Med. Chem.* **1995**, *38*, 2681–2691.
- Knunicek, J.; Luengo, S.; Gago, F.; Ortiz, A. R.; Wade, R. C. et al. Comparative binding energy analysis of the substrate specificity of haloalkane dehalogenase from *Xanthobacter autotrophicus* GJ10. *Biochemistry* **2001**, *40*, 8905–8917.
- Wang, T.; Wade, R. C. Comparative binding energy (COMBINE) analysis of OppA–peptide complexes to relate structure to binding thermodynamics. *J. Med. Chem.* **2002**, *45*, 4828–4837.
- Wang, T.; Wade, R. C. Comparative binding energy (COMBINE) analysis of influenza neuraminidase–inhibitor complexes. *J. Med. Chem.* **2001**, *44*, 961–971.
- Perez, C.; Pastor, M.; Ortiz, A. R.; Gago, F. Comparative binding energy analysis of HIV-1 protease inhibitors: incorporation of solvent effects and validation as a powerful tool in receptor-based drug design. *J. Med. Chem.* **1998**, *41*, 836–852.
- Murcia, M.; Ortiz, A. R. Virtual screening with flexible docking and COMBINE-based models. Application to a series of factor Xa inhibitors. *J. Med. Chem.* **2004**, *47*, 805–820.
- Maryanoff, B. E. Serin inhibitors of serine proteases as potential therapeutic agents: the road from thrombin to trypsin to cathepsin G. *J. Med. Chem.* **2004**, *47*, 769–787.
- Krem, M. M.; Rose, T.; Di Cera, E. Sequence determinants of function and evolution in serine proteases. *Trends Cardiovasc. Med.* **2000**, *10*, 171–176.
- Kastenholz, M. A.; Pastor, M.; Cruciani, G.; Haaksma, E. E. J.; Fox, T. GRID/CPCA: A new computational tool to design selective ligands. *J. Med. Chem.* **2000**, *43*, 3033–3044.
- Sheridan, R. P.; Holloway, M. K.; McGaughey, G.; Mosley, R. T.; Singh, S. B. A simple method for visualizing the differences between related receptor sites. *J. Mol. Graphics Modell.* **2002**, *21*, 217–225.
- Rezaie, A. R.; Olson, S. T. Contribution of lysine 60f to S1' specificity of thrombin. *Biochemistry* **1997**, *36*, 1026–1033.
- Debnath, B.; Samanta, S.; Naskar, S. K.; Roy, K.; Jha, T. QSAR study on the affinity of some arylpiperazines towards the 5-HT_{1A}/alpha₁-adrenergic receptor using the E-state index. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2837–2842.
- Kamath, S.; Buolamwini, J. K. Receptor-guided alignment-based comparative 3D-QSAR studies of benzylidene malonitrile tyrosinase inhibitors as EGFR and HER-2 kinase inhibitors. *J. Med. Chem.* **2003**, *46*, 4657–4668.
- Rivara, S.; Mor, M.; Bordin, F.; Silva, C.; Zuliani, V. et al. Synthesis and three-dimensional quantitative structure–activity relationship analysis of h(3) receptor antagonists containing a neutral heterocyclic polar group. *Drug Des. Discovery* **2003**, *18*, 65–79.
- Brea, J.; Rodrigo, J.; Carrieri, A.; Sanz, F.; Cadavid, M. I. et al. New serotonin 5-HT_{2A} (5-HT_{2B}), and 5-HT_{2C} receptor antagonists: synthesis, pharmacology, 3D-QSAR, and molecular modeling of (aminoalkyl)benzo and heterocycloalkanones. *J. Med. Chem.* **2002**, *45*, 54–71.
- Fichera, M.; Cruciani, G.; Bianchi, A.; Musumarra, G. A 3D-QSAR study on the structural requirements for binding to CB(1) and CB(2) cannabinoid receptors. *J. Med. Chem.* **2000**, *43*, 2300–2309.

- (37) Moron, J. A.; Campillo, M.; Perez, V.; Unzeta, M.; Pardo, L. Molecular determinants of MAO selectivity in a series of indolyl-methylamine derivatives: biological activities, 3D-QSAR/CoMFA analysis, and computational simulation of ligand recognition. *J. Med. Chem.* **2000**, *43*, 1684–1691.
- (38) Bakken, G. A.; Jurs, P. C. QSARs for 6-azasteroids as inhibitors of human type 1 5 α -reductase: prediction of binding affinity and selectivity relative to 3-BHSD. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1255–1265.
- (39) Ravina, E.; Negreira, J.; Cid, J.; Masaguer, C. F.; Rosa, E. et al. Conformationally constrained butyrophenones with mixed dopaminergic (D(2)) and serotonergic (5-HT(2A), 5-HT(2C)) affinities: synthesis, pharmacology, 3D-QSAR, and molecular modeling of (aminoalkyl)benzo- and -thienocycloalkanones as putative atypical antipsychotics. *J. Med. Chem.* **1999**, *42*, 2774–2797.
- (40) Huang, Q.; Liu, R.; Zhang, P.; He, X.; McKernan, R. et al. Predictive models for GABAA/benzodiazepine receptor subtypes: studies of quantitative structure–activity relationships for imidazobenzodiazepines at five recombinant GABA_A/benzodiazepine receptor subtypes [α^1 , α^2 , α^3 , α^4 , α^5 , and α^6] via comparative molecular field analysis. *J. Med. Chem.* **1998**, *41*, 4130–4142.
- (41) Fabian, M. A.; Biggs, W. H., III; Treiber, D. K.; Atteridge, C. E.; Azimioara, M. D. et al. A small molecule–kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **2005**, *23*, 329–336.
- (42) *Insight-II*, version 2000; Molecular Simulations, Inc.: San Diego, CA.
- (43) Brandstetter, H.; Turk, D.; Hoeffken, H. W.; Grosse, D.; Sturzebecher, J. et al. Refined 2.3 Å X-ray crystal structure of bovine thrombin complexes formed with the benzamidine and arginine-based thrombin inhibitors NAPAP, 4-TAPAP, and MQPA. A starting point for improving antithrombotics. *J. Mol. Biol.* **1992**, *226*, 1085–1099.
- (44) Turk, D.; Sturzebecher, J.; Bode, W. Geometry of binding of the *N*- α -tosylated piperidides of *m*-amidino-, *p*-amidino-, and *p*-guanidino phenylalanine to thrombin and trypsin. X-ray crystal structures of their trypsin complexes and modeling of their thrombin complexes. *FEBS Lett.* **1991**, *287*, 133–138.
- (45) Padmanabhan, K.; Padmanabhan, K. P.; Tulinsky, A.; Park, C. H.; Bode, W. et al. Structure of human des(1–45) factor Xa at 2.2 Å resolution. *J. Mol. Biol.* **1993**, *232*, 947–966.
- (46) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N. et al. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (47) Wang, J.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* **2000**, *21*, 1049–1074.
- (48) Besler, B. H.; Merz, K. M.; Kollman, P. A. Atomic charges derived from semiempirical methods. *J. Comp. Chem.* **1990**, *11*, 431–439.
- (49) Stewart, J. J. MOPAC: a semiempirical molecular orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1–105.
- (50) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
- (51) Lupyán, D.; Leo-Macias, A.; Ortiz, A. R. A new progressive-iterative algorithm for multiple structure alignment. *Bioinformatics* **2005**, *21*, 3255–3263.
- (52) Ortiz, A. R.; Strauss, C. E.; Olmea, O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.* **2002**, *11*, 2606–2621.
- (53) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E., III; Wang, J.; Ross, W. S.; Simmerling, C.; Darden, T.; Merz, K. M.; Stanton, R. V.; Cheng, A.; Vincent, J. J.; Crowley, M.; Tsui, V.; Gohlke, H.; Radmer, R.; Duan, Y.; Pitner, J.; Massova, I.; Seibel, G. L.; Singh, U. C.; Weiner, P. P.; Kollman, A. *AMBER 7*; University of California: San Francisco, CA.
- (54) Honig, B.; Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **1995**, *268*, 1144–1149.
- (55) Still, W.; Tempczyk, A.; Hawley, R.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
- (56) Hassan, S. A.; Guarnieri, F.; Mehler, E. L. A general treatment of solvent effects based on screened Coulomb potentials. *J. Phys. Chem. B* **2000**, *104*, 6478–6489.
- (57) Hassan, S. A.; Guarnieri, F.; Mehler, E. L. Characterization of hydrogen bonding in a continuum solvent model. *J. Phys. Chem. B* **2000**, *104*, 6490–6498.
- (58) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J., III. The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses. *SIAM J. Sci. Stat. Comp.* **1984**, *5*, 735–743.
- (59) DeLano, W. *The PyMOL Molecular Graphics System*; DeLano Scientific, LLC: San Carlos, CA, <http://www.pymol.org>.

JM060350H