



## Comparative molecular field analysis and energy interaction studies of thrombin-inhibitor complexes

Roberta Bursi\* & Peter D.J. Grootenhuis\*,\*\*

*Department of Molecular Design & Informatics, N.V. Organon, P.O. Box 20, 5340 BH Oss, The Netherlands*

Received 6 May 1998; Accepted 15 September 1998

**Key words:** CoMFA, drug design, inhibitor, molecular modelling, protease, scoring function

### Summary

A Comparative Molecular Field Analysis (CoMFA) and an interaction energy-based method were applied on a database holding the 3D structures of 29 thrombin-inhibitor complexes. Several parameters were optimized in both methods in order to obtain the best correlation between theoretical and experimentally determined binding ( $K_i$ ) data. CoMFA, which only uses the information of the inhibitors, performed best ( $r = 0.99$ ,  $q^2 = 0.46$ ,  $N = 29$ ) when HF 6-31G charges were used in combination with a pharmacophore-based alignment. Inclusion of hydrophobic fields did not lead to improvements. The interaction energy-based approach uses the information of the whole thrombin-inhibitor complex. A statistically significant correlation ( $r = 0.74$ ,  $N = 14$ ) could only be obtained for a subset of the database containing the high resolution structures. Geometry optimization of the ligand only in combination with downscaled electrostatics performed best.

### Introduction

The availability of three-dimensional structures of target enzyme structures is mostly considered as a great help for the design of inhibitors. Also for the interpretation of binding data the structural analysis of the interactions between the enzyme and the inhibitor plays an important role. This explains in part the large efforts of pharmaceutical companies to obtain crystal structures of key enzymes for particular diseases. In the field of pharmacologically important proteases the efforts on HIV protease (AIDS), elastase (cystic fibrosis), and thrombin (thrombosis) are very illustrative. Although for each of these proteases substrates, substrate analogues and peptidomimetic inhibitors were known before the first protease crystal structures were solved, successful contributions to several drug candidates by computational methods have been published [1,2]. Unfortunately, false predictions and suggestions leading to inactive compounds can only be published in particular cases since inactive compounds

are seldomly patented, thus hampering their publication. Therefore, the question remains how effective structure-based design efforts really are.

One of the main factors that limits the impact of computational drug design methodologies is the absence of general, robust scoring functions with high predictive abilities. Nearly every recent study on structure-based design has emphasized this issue [3]. The state of the art of computational scoring methods that assess the binding affinity between proteins and (putative) ligand based on the structure has been reviewed recently [4–6]. Fortunately, progress with the development of scoring methods is being made as the methods become more sophisticated and take into account (de)solvation effects, protein and ligand flexibility, entropy, etc.

Completely different approaches towards analysing and predicting the binding affinities of ligands are followed when the structure of the target receptor is not known, as is the case for most diseases. 3D-QSAR techniques such as Comparative Molecular Field Analysis (CoMFA) [7] but also neural network and evolutionary algorithms are frequently applied nowadays [8].

\*To whom correspondence should be addressed.

\*\*Present address: CombiChem, Inc., 9050 Camino Santa Fe, San Diego, CA 92121, U.S.A.

In the present study we compare and evaluate a representative example of both types of methodologies in order to evaluate the impact of knowing the 3D structure of the protein-ligand complex. Thus CoMFA and several scoring function-based analyses were applied on a database of noncovalent thrombin-inhibitor complex structures. Thrombin is very much suited as a test case for design methodologies since its well-defined active site cleft has both polar and apolar regions. In previous studies [9,10] we used a mostly modelled database of thrombin-inhibitor complexes based on the crystal structures of thrombin in complex with the inhibitors Argatroban, NAPAP, and TAPAP. In the present study a database holding 29 crystal structures of thrombin-inhibitor complexes was used. Although earlier work suggests that simple scoring methodologies only work within congeneric series of ligands [11], we decided to construct a chemically diverse database consisting of inhibitors from several compound series (see below). The CoMFA and scoring function-based methodologies were optimized for the current thrombin inhibitor database and subsequently the predictive power of both approaches was evaluated.

### Computational methodology

#### *The thrombin-inhibitor database*

A database of 29 thrombin-inhibitor complexes was compiled. All inhibitors bind non-covalently to thrombin. The binding mode of a typical thrombin inhibitor, Argatroban, is displayed in Figure 1. As Table 1 shows, all complexes are structurally defined by X-ray crystallographic analysis at resolutions varying from 3.2 to 2.2 Å; 14 structures had a resolution smaller than 2.5 Å. Water molecules and ions (if present) were removed from the coordinate sets. In the case of the human thrombin-DAPA complex, crystallization was achieved by co-crystallization with a short peptide from the C-terminus of hirudin in order to decrease autocatalytic degradation of thrombin. This peptide was also removed from the dataset. Residues at the two termini of the A-chain and the C-terminus of the B-chain of thrombin with partially undefined coordinates were excluded. Since these three termini are distant from the actual binding site of the inhibitor (more than 15 Å), their omission is unlikely to affect the binding energy.

All molecular mechanics calculations were performed using QUANTA/CHARMm version 4.1. The

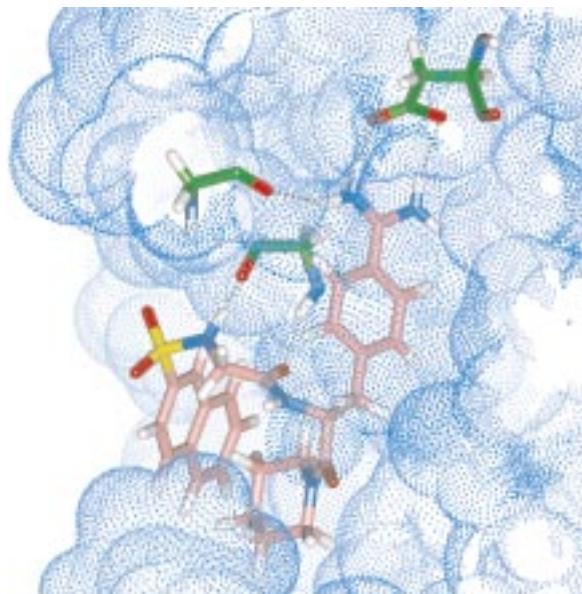


Figure 1. X-ray structure of Argatroban: binding mode.

all-atom force field was used. Hydrogen atoms of the protein moiety were added by the HBUILD option of the CHARMm program. The default charge states of titratable sites on the protein were chosen. All histidines, including His 57, were uncharged. The inhibitor molecules were modelled with QUANTA's molecular editor. The charge template method was used to assign partial charges to the inhibitor atoms. A charge smoothing scheme was applied that distributes residual partial charges on carbon and neighbouring nonpolar hydrogen atoms. The raw crystal structures were further refined by minimizing the energy of the inhibitors while keeping all protein atoms at fixed positions. We will refer to these refined structures as the X-ray min set.

#### *CoMFA*

Tripos standard steric and electrostatic molecular fields were used in all CoMFA analyses [12]. The field values were not smoothed and the cut-off values of the steric and electrostatic interactions were kept to the default values of 30.0 kcal/mol. A smooth transition was chosen between the cut-off plateaus for the steric and electrostatic calculations and every Coulombic electrostatic energy calculation was performed using a distance dependent dielectric  $\epsilon = R$ .

Most of the partial least squares (PLS) analyses were performed at a grid distance of 2 Å, but for every type of alignment the translational and rotational invariance was subsequently checked by setting

Table 1. Thrombin database

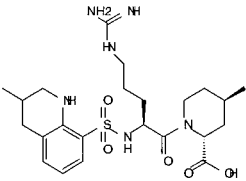
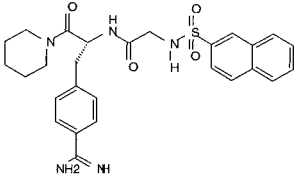
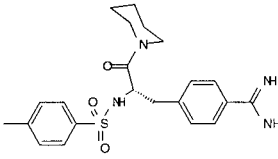
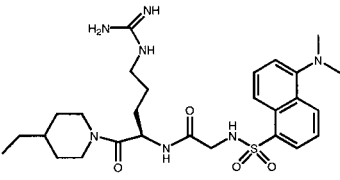
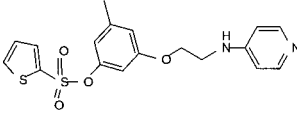
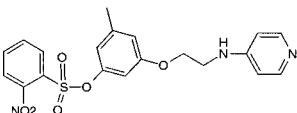
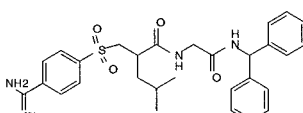
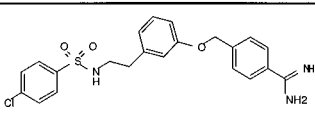
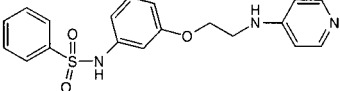
No.	Structure	R [Å]	Name, Code	Observed pKi
1		2.2	Argatroban <sup>1</sup>	7.7
2		2.3	NAPAP <sup>2</sup>	8.5
3		2.5	4-TAPAP <sup>3</sup>	5.9
4		2.3	DAPA <sup>4</sup>	7.0
5		3.1	BM12.1668	8.3
6		2.8	BM12.1684	8.2
7		2.8	BM14.1196	5.1
8		2.9	BM14.1203	5.8
9		2.2	BM14.1224	6.5

Table 1. (continued)

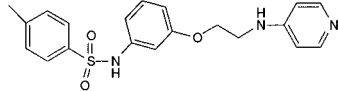
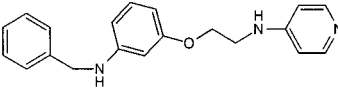
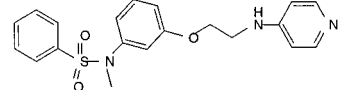
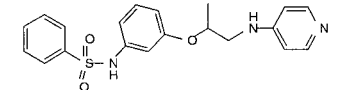
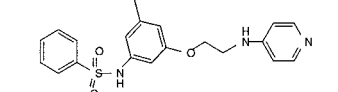
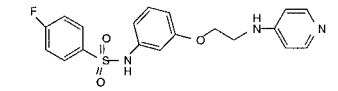
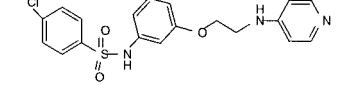
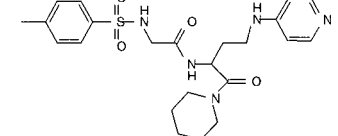
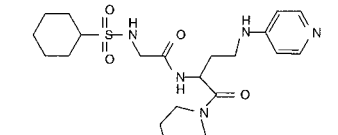
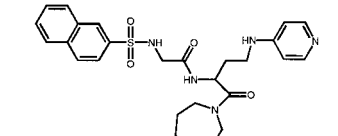
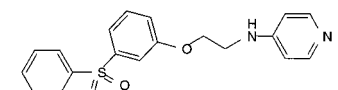
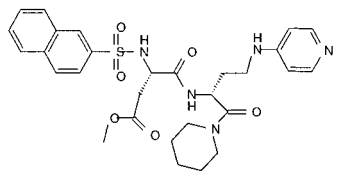
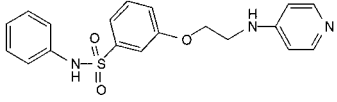
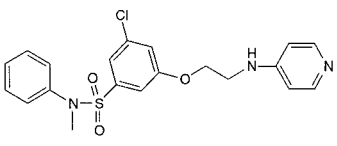
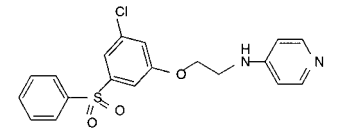
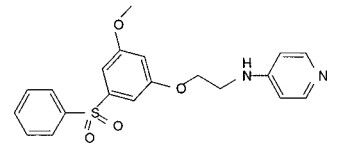
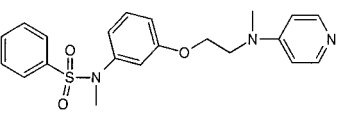
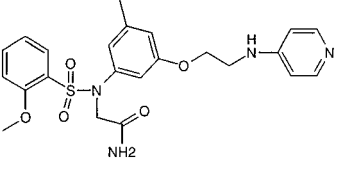
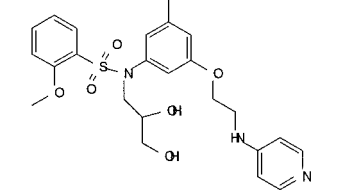
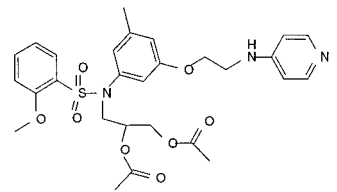
10		2.7	BM14.1238	5.8
11		2.5	BM14.1241	5.7
12		2.5	BM14.1243	7.2
13		2.4	BM14.1244	5.8
14		2.5	BM14.1248 <sup>5</sup>	7.6
15		2.9	BM18.0537	6.5
16		3.2	BM18.0540	6.2
17		2.8	BM51.0986	6.0
18		2.7	BM51.1011 <sup>6</sup>	5.4
19		2.9	BM51.1012	6.1
20		2.5	BM51.1022	6.5

Table 1. (continued)

21		2.8	BM51.1023	6.3
22		2.5	BM51.1031	5.8
23		2.4	BM51.1037	7.4
24		3.0	BM51.1045	7.3
25		2.5	BM51.1047	6.8
26		2.9	BM51.1059	6.2
27		2.5	BM51.1081	8.4
28		2.6	BM51.1110	8.2
29		2.8	BM51.1116	8.5

PDB reference code: <sup>1</sup>1ETR; <sup>2</sup>1ETS; <sup>3</sup>1ETT; <sup>4</sup>1FPC; <sup>5</sup>1UVT; <sup>6</sup>1UVS. The other compounds were taken from the Boehringer Mannheim unpublished database.

the grid distance to 1 Å. The CoMFA standard scaling was used throughout the analyses and during the leave-one-out cross validation procedure the columns filtering  $\sigma$  values of 2.0 and 0.0 (no filtering) kcal/mol were used. Because of the lengthy duration of the leave-one-out analysis performed at a grid distance of 1 Å, the cross validated analyses were performed with  $\sigma$  values equal to 2.0 kcal/mol.

Correlation chances were checked by randomizing several times the experimental binding constants and by repeating the PLS analysis. The results indicated that the correct structure-activity PLS models were unaffected by overfitting.

The hydrophobic molecular field calculations were performed by means of the HINT program [13] using the default settings. The refined crystal structures as described in the previous section were used in all CoMFA calculations.

#### *Interaction energy-based scoring*

Correlations between binding affinities and interaction energies were investigated by variation of several parameters during the geometry optimization and interaction energy evaluation phase. We started with taking no electrostatics into account during the geometry optimization. The effect of protein flexibility during the optimisation phase was studied by allowing atoms within  $M$  Å ( $M = 0, 2, 4, 6, 8$ ) of the ligand to move during the optimisation, while the rest of the protein was kept fixed. In all cases the ligand geometry was optimized without constraints. The other atoms that are allowed to move were harmonically constrained with a force constant of  $F$  kcal/mol Å ( $F = 0, 2, 4, 6, 8, 10$ ). During the subsequent interaction energy evaluation phase the effect of the electrostatics was varied by applying a distance dependent dielectric  $\epsilon = NR$  ( $N = 1, 2, 3, 4$ ) leading to  $(4 \times 6 + 1) \times 4 = 100$  combinations ('grid points') of  $M$ ,  $F$ , and  $N$ . In the case of a constant dielectric  $\epsilon = E$  ( $E = 1, 20, 39, 58, 77$ ) 125 grid points were defined. The calculations with the distance dependent dielectric were also performed while applying the same function during the geometry optimization, leading to another 100 grid points. For every grid point on this 325 point grid, the 29 thrombin-inhibitor complexes were energy minimised and the interaction energy was evaluated. Finally the correlation between the interaction energies and the experimentally determined binding constants was calculated using linear regression analysis.

## **CoMFA of the thrombin-inhibitor database**

### *The effect of different charge models*

Since it is known that the type of charges employed in CoMFA strongly affects the results of the electrostatic interaction calculations [14], we analysed which set of charges performs best for the thrombin-inhibitor database. CHARMM, Gasteiger [15], semiempirical Mulliken [16] AM1 [17], ESP [18] AM1 and ab initio Mulliken HF 6-31G charges [19] were calculated for the 29 thrombin inhibitors. The results are displayed in Table 2. When no filter was used in the analyses, the resulting  $q^2$  values varied considerably from 0.16 to 0.46. The optimum number of components was 3 in all analyses and the highest  $q^2$  value was obtained when the HF 6-31G charges were used. ESP and Mulliken AM1 6-31G charges yield the second best  $q^2$  values followed by the Gasteiger and CHARMM sets of charges. The small  $q^2$  values of the latter sets were mostly due to two outliers, NAPAP and BM14.1196, and no obvious explanation for this could be found. While it was not surprising that ESP AM1 charges performed better than the Mulliken AM1 charges, it was interesting to see that the simple split valence 6-31G basis set in the absence of polarization and diffuse functions provided the best model. Significant models ( $q^2 \geq 0.3$ – $0.4$ ) were obtained only when Mulliken and ESP AM1 and HF 6-31G charges were used.

Since the HF 6-31G charge calculations required several CPU hours on a Power Challenge R4000 for each molecule, while the Mulliken AM1 calculations only required 1–2 CPU minutes, all subsequent analyses were performed with Mulliken AM1 charges.

### *The effect of various alignment rules*

CoMFA is only possible after superposition of the studied molecules. If no experimental data are available on the binding mode of a ligand, an 'alignment rule' has to be applied and this is generally regarded as one of the most serious problems. In our case, the binding mode of the ligands is experimentally known and the alignment rule is simply reduced to the rigid superposition of the known protein structures as they are present in the corresponding complexes. We decided to compare this experimental alignment with other – theoretically derived – alignment methods. A total of four different alignment rules were considered and the results are displayed in Table 3.

Firstly, the experimental alignments obtained with the initial (X-ray) and minimised (X-ray min) crystal structures of the thrombin inhibitors were considered.

Table 2. Comparison of different charge models (alignment rule: X-ray min. grid = 2.0 Å, N=29)

Model	$q^2$ <sup>a</sup>	$s$ <sup>b</sup>	comp <sup>c</sup>	$\sigma$ <sup>d</sup>	$r^2$ <sup>e</sup>	$s$ <sup>f</sup>	$F^g$	$\sigma$	# <sup>h</sup>
CHARMm	0.220	0.970	3	2.0					29
CHARMm	0.156	1.009	3	0.0	0.936	0.278	121.3	0.0	29
Gasteiger	0.296	0.922	3	2.0					29
Gasteiger	0.293	0.924	3	0.0	0.928	0.295	107.4	0.0	29
AM1	0.393	0.856	3	2.0					29
AM1	0.370	0.872	3	0.0	0.946	0.255	146.2	0.0	29
ESP AM1	0.404	0.848	3	2.0					29
ESP AM1	0.399	0.851	3	0.0	0.948	0.251	151.8	0.0	29
HF 6-31G	0.462	0.806	3	2.0					29
HF 6-31G	0.460	0.807	3	0.0	0.949	0.248	155.8	0.0	29

<sup>a</sup> Correlation coefficient in prediction.

<sup>b</sup> Statistical error in prediction.

<sup>c</sup> Number of components.

<sup>d</sup> Column filtering in kcal/mol.

<sup>e</sup> Correlation coefficient in fitting.

<sup>f</sup> Statistical error in fitting.

<sup>g</sup> F factor.

<sup>h</sup> Number of molecules.

Secondly, the refined structures were aligned by means of a shape-based alignment with superimposed centres of mass, which is implemented in the SPERM program [20]. Finally, three centroids were defined in every ligand refined structure corresponding to the three sites which interact with the P<sub>1</sub>, P<sub>2</sub> and P<sub>3</sub> pockets of thrombin. The three centroids can easily be described as three pharmacophore centers, two hydrophobic centers (interacting with the P<sub>2</sub> and P<sub>3</sub> pockets) and one positively charged center (interacting with the P<sub>1</sub> pocket), as given by the SYBYL definition of pharmacophores.

As discussed in the Methods section, the translational and rotational invariance of these alignment rules was analysed by reducing the grid steps of the CoMFA box from 2 to 1 Å. With the exception of the SPERM alignment, the grid refinement quantitatively affects the  $q^2$  and corresponding  $s$  values of the different alignment rules. Qualitatively, however, the ranking of these models does not change: the Centroid alignment provides the best  $q^2$ , while SPERM yields the lowest one. The two experimental alignments yield models very similar to each other and of intermediate  $q^2$  value. The values of the final correlation coefficients,  $r^2$ , are not significantly different, with the exception of the value corresponding with the SPERM-based alignment method.

The quality of a model is not only assessed by the  $q^2$  and corresponding  $s$  values but also by the number

of components it requires; the fewer the components, the better is the model. Since the Centroid alignment requires 5 components while the X-ray and X-ray min alignments require only 3, we have repeated the PLS analysis for the Centroid alignment with 3 components. The results reported in Table 3 show that when 3 components are considered, the Centroid and the experimentally determined alignments provide models of similar quality.

#### Molecular field contour analysis

It can be helpful to look at these models not only from a statistical point of view, but also in terms of their corresponding steric and electrostatic molecular field contours. The steric and electrostatic fields of the X-ray refined and Centroid models, 1 Å grid distance, were compared and found to describe similar regions in space as important and/or unimportant to the binding. The CoMFA model based on the SPERM alignment, on the contrary, provides a different description of the space surrounding the ligands and therefore of the binding mode as deduced from CoMFA. The differences in the theoretically derived alignments can be explained by observing that the Centroid-based alignment tends to superimpose substituent anchor-points while the SPERM-based alignment superimposes the overall molecular shape. Because of that, structural variation necessary to describe the activity is closely defined in the former case while it is more spread in the

Table 3. Comparison of alignment methods (Mulliken AM1 charge model, N = 29)

Method	$q^2$ <sup>a</sup>	$s^b$	comp <sup>c</sup>	$\sigma^d$	$r^2$ <sup>e</sup>	$s^f$	F <sup>g</sup>	$\sigma$	Grid <sup>h</sup>
X-ray	0.387	0.860	3	2.0					2.0
	0.389	0.859	3	0.0	0.911	0.329	84.8	0.0	2.0
	0.330	0.899	3	2.0	0.912	0.326	86.0	0.0	1.0
X-ray min	0.393	0.856	3	2.0					2.0
	0.370	0.872	3	0.0	0.946	0.255	146.2	0.0	2.0
	0.302	0.918	3	2.0	0.940	0.269	130.4	0.0	1.0
SPERM	0.204	0.961	2	2.0					2.0
	0.194	0.967	2	0.0	0.791	0.493	49.1	0.0	2.0
	0.208	0.958	2	2.0	0.782	0.503	46.6	0.0	1.0
Centroid	0.529	0.804	6	2.0					2.0
	0.521	0.810	6	0.0	0.993	0.096	540.3	0.0	2.0
	0.432	0.863	5	2.0	0.986	0.137	315.6	0.0	1.0
	0.342	0.891	3	2.0	0.934	0.283	117.6	0.0	1.0

<sup>a–g</sup>See Table 1.<sup>h</sup>Grid distance in Å.

latter, producing a larger uncertainty (larger  $s$  values, smaller  $q^2$  values) in the analysis.

The CoMFA fields visualize suggested regions in space which are suitable or unsuitable to increase the affinity of the ligands towards the receptor. However, they can also be interpreted as a fingerprint of the receptor. Yellow/green regions (unsuitable/suitable regions for steric bulk on the ligand), for instance, suggest the presence/absence of the receptor functional groups. The X-ray refined and the Centroid models are characterised by features which can be recognised in the receptor structure, while this is not the case for the SPERM alignment.

#### *The effect of adding a hydrophobic field*

Finally, the 29 thrombin-inhibitors were analysed in terms of the hydrophobic field. A hydrophobic calculation was performed with HINT for the refined ligand structures. The HF 6-31G set of charges were used and the alignment was the X-ray min alignment rule. The results are shown in Table 4.

When only one molecular field is used in CoMFA, each molecular field yields  $q^2$  and corresponding  $s$  values comparable to the other fields, although the steric field yields the largest  $q^2$  and the best correlation coefficient  $r^2$ . In case two molecular fields are combined, the best model is given by the conventional CoMFA analysis, i.e., the combination of steric and electrostatic fields. In combination with the steric or the hydrophobic field, the electrostatic contribution remains the smallest contribution and of comparable magnitude in both models. The combination of all three

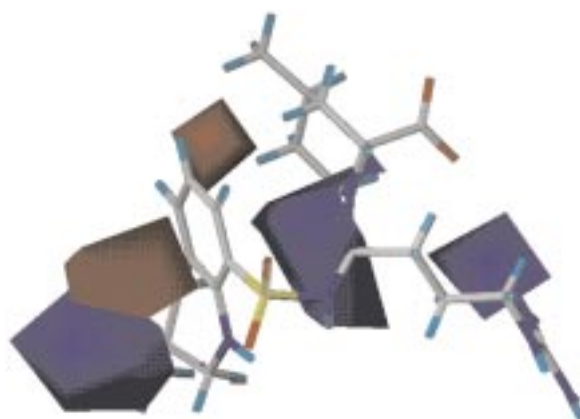


Figure 2. 3-Fields analysis: electrostatic contours. The red and blue regions correspond to regions favourable and unfavourable to a negative charge. Argatroban is displayed.

fields, the 3-Fields analysis, unfortunately did not significantly improve the model. The  $q^2$  value dropped from 0.462 to 0.423. In this analysis, the electrostatic field contribution is intermediate in size to the other two contributions and does not change remarkably with respect to the previous two-fields analyses. The steric and hydrophobic fields possess the smallest and the largest contributions, respectively, and the steric field contribution undergoes the largest changes. The three molecular fields results of the 3-Fields analysis are displayed in Figures 2–4.



Table 4. CoMFA: Molecular fields contribution (alignment rule: X-ray min,  $n = 29$ , charge mode: HF 6-31G,  $\sigma = 2.0$  kcal/mol)

Field(s)	$q^2$ <sup>a</sup>	$s^b$	comp <sup>c</sup>	$r^2$ <sup>d</sup>	$s^e$	ster <sup>f</sup>	elect <sup>f</sup>	hydr <sup>f</sup>
ster	0.350	0.886	3	0.902	0.345	1.00	–	–
electr	0.308	0.896	2	0.800	0.481	–	1.00	–
hydr	0.324	0.869	1	0.720	0.559	–	–	1.00
ster+electr	0.462	0.806	3	0.949	0.248	0.62	0.38	–
ster+hydr	0.362	0.860	2	0.857	0.408	0.37	–	0.63
electr+hydr	0.402	0.833	2	0.867	0.393	–	0.44	0.56
ster+electr+hydr	0.423	0.818	2	0.877	0.377	0.24	0.34	0.42

<sup>a–e</sup>See Table 1.

<sup>f</sup>Molecular field contribution.

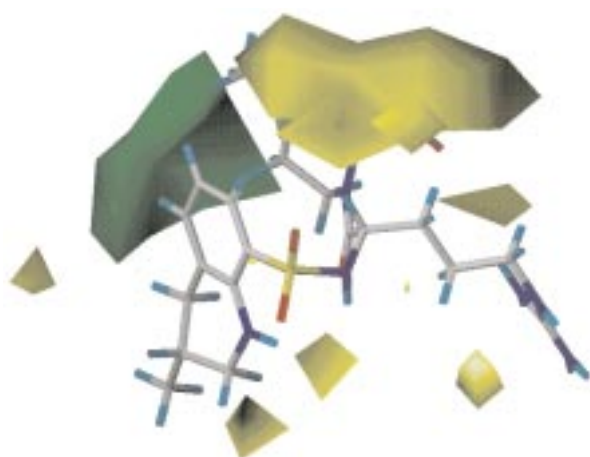


Figure 3. 3-Fields analysis: steric contours. The green and yellow regions are regions favourable and unfavourable to steric bulk. Argatroban is displayed.

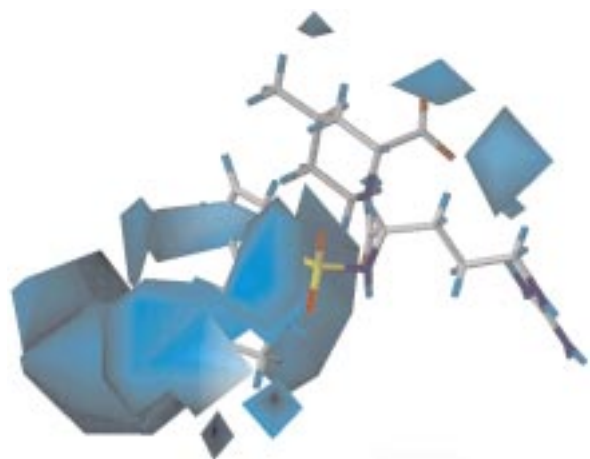


Figure 4. 3-Fields analysis: hydrophobic contours. Argatroban is displayed.

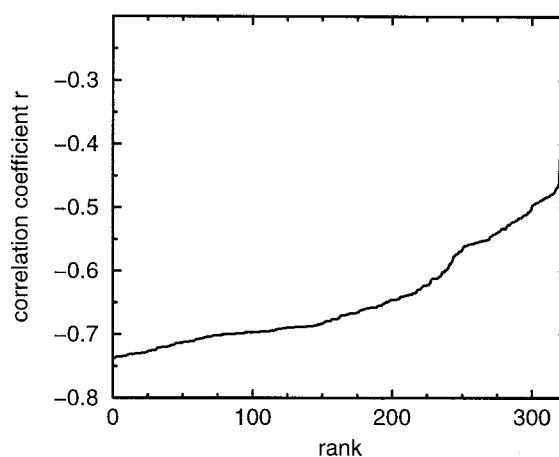


Figure 5. Correlation coefficients  $r$  versus the rank number.

### Interaction energy analysis of the thrombin-inhibitor database

The interaction energies between thrombin and each of the 29 inhibitors were calculated using different protocols (see Computational methodology section). When trying to correlate the calculated interaction energies with the corresponding experimentally determined  $K_i$  values, it turned out that none of the protocols lead to a statistically significant correlation for the 29 molecule database. However, for the smaller set of 14 high resolution structures with resolution smaller than  $2.5 \text{ \AA}$ , meaningful correlations could be obtained. The protocols leading to the best and worst correlations are given in Table 5, while the correlation coefficients for the 325 grid points have been depicted in Figure 5.

The correlation coefficients  $r$  vary from  $-0.26$  to  $-0.73$ . From Figure 5 it can be learned that the relatively good correlations are not very dependent on the protocol used, although the most effective proto-

Table 5. Interaction energy results

Rank	F <sup>a</sup>	M <sup>b</sup>	$\epsilon^c$	r <sup>d</sup>	Type <sup>e</sup>
1	0.0	0.0	4.0	-0.738	R
2	0.0	0.0	39.0	-0.736	C
3	0.0	0.0	20.0	-0.735	C
4	0.0	0.0	58.0	-0.735	C
5	2.0	2.0	20.0	-0.735	C
6	2.0	2.0	39.0	-0.735	C
7	0.0	0.0	77.0	-0.734	C
8	2.0	2.0	58.0	-0.734	C
9	8.0	2.0	20.0	-0.734	C
10	2.0	2.0	77.0	-0.733	C
11	8.0	2.0	39.0	-0.732	C
12	4.0	2.0	20.0	-0.731	C
13	4.0	2.0	39.0	-0.731	C
14	8.0	2.0	4.0	-0.731	R
15	8.0	2.0	58.0	-0.731	C
.	.	.	.	.	.
.	.	.	.	.	.
34	10.0	4.0	58.0	-0.721	C
35	8.0	2.0	3.0	-0.720	R
36	8.0	2.0	4.0	-0.720	RR
37	10.0	4.0	20.0	-0.720	C
38	10.0	4.0	77.0	-0.720	C
.	.	.	.	.	.
.	.	.	.	.	.
315	4.0	2.0	1.0	-0.478	RR
316	2.0	6.0	1.0	-0.476	R
317	8.0	4.0	1.0	-0.470	R
318	0.0	8.0	1.0	-0.468	RR
319	0.0	6.0	3.0	-0.463	R
320	0.0	8.0	2.0	-0.412	R
321	0.0	6.0	2.0	-0.392	R
322	0.0	8.0	1.0	-0.370	C
323	0.0	6.0	1.0	-0.321	C
324	0.0	8.0	1.0	-0.275	R
325	0.0	6.0	1.0	-0.261	R

<sup>a</sup> F: harmonic constrain force constant.<sup>b</sup> M: range within which the ligand is allowed to move during optimisation.<sup>c</sup>  $\epsilon$ : dielectric constant.<sup>d</sup> r: correlation coefficient.<sup>e</sup> Type: R and C denote distance dependent and constant dielectric, respectively. RR denotes that during minimisation and interaction energy evaluation a distance dependent dielectric was used.

cols share certain characteristics. The protocol with the best correlation coefficient ( $r = -0.738$ ) was obtained with  $\epsilon = 4R$  and only the ligand allowed to move during the minimization, which is similar to the one found previously for describing a set of modelled thrombin-inhibitor complexes [10], although in that study  $\epsilon = R$  during the energy evaluation appeared to be optimal. The electrostatics are downscaled in all the top-scoring protocols. Only the ligand is allowed to move or at most a limited number of (harmonically constrained) protein atoms. These conclusions are in general agreement with our earlier studies in this area [10].

Interestingly, in the top-scoring protocols no electrostatics were applied during the geometry optimization phase. The first protocol with a distance dependent dielectric during this phase is found at rank 36 ( $r = -0.72$ ) and also there the electrostatics are scaled down ( $\epsilon = 4R$ ). Typically this protocol leads to a significantly poorer performance than the one in which no electrostatics are applied during the geometry optimization phase.

Generally, the low scoring protocols (see Table 5) are characterized by an important role for the electrostatics in combination with many flexible protein atoms. Turning on electrostatics during the geometry optimization leads to poor results, in particular when the electrostatics are not scaled down, e.g.  $\epsilon = R$  or  $\epsilon = 2R$ .

The optimum protocol found in this study (no protein flexibility, down-scaled electrostatics) may seem counter-intuitive. However, we feel that the reduced electrostatics is a way to compensate for the fact that no explicit (de)solvation effects are taken into account. In the absence of explicit water molecules the calculated strength of hydrogen bonds and other electrostatic interactions tends to be overestimated when using low dielectrics. The limited flexibility of the protease context is probably due to the fact that thrombin displays only small structural changes when comparing various thrombin-inhibitor complexes. For other proteins that are known to undergo important conformational changes upon binding to a ligand, such as HIV-protease, other protocols allowing more flexibility may be found to be more effective.

## Discussion

CoMFA and the interaction energy-based method are two computational approaches that describe binding

in similar terms, i.e. by a steric contribution using the Lennard-Jones potential, and an electrostatic (Coulombic) contribution. Other terms describing effects such as the (de)solvation energy and/or the entropic factors are neglected by both approaches as used in the current paper. Subsequently, the two techniques conceptually diverge to rejoin at the end in a common objective, i.e. the prediction of the binding affinities,  $pK_i$ 's.

Within the CoMFA approach, the  $pK_i$ 's are determined as a linear combination of several statistically determined variables which have been obtained from the molecular field differences existing among the structures of the inhibitors only. Within the interaction energy approach the steric and electrostatic contributions to the interaction energies are calculated using the 3D structures of the protease-inhibitor complexes. From this study it is clear that despite the application of energy minimization protocols the interaction energy approach demands high quality initial structures in order to yield satisfactory results.

Another factor that has to be considered when the performance of the two approaches is analysed is the multi-variable (multi-component) nature of CoMFA. This property of CoMFA also explains why the correlation coefficients  $r^2$  (and, therefore,  $r$ ) of its significant models are generally higher than the correlation coefficients  $r$  obtained by the best interaction energy computational protocols. It was somewhat disappointing to find that for this set of inhibitors the inclusion of additional variables such as the hydrophobic field did not improve the model. Clearly, the choice of the 'optimum' molecular field(s) remains a critical issue in CoMFA studies, as recently discussed by Kellogg [21].

The efficacy and utility of the CoMFA technique in QSAR studies is by now well established in the literature [22]. It is also recognised that the requirement of a molecular alignment method can be problematic when no experimental information is available on the binding modes of the ligands under investigation [23]. For a given set of molecules, usually many theoretical alignments can be designed that yield statistically significant models. This may lead to erroneous conclusions, e.g. we found that the  $q^2$  and  $s$  values of the Centroid alignment are better than the corresponding values of the experimental alignments. In this particular case, the steric and electrostatic contours of these alignments, i.e. the physical information included in these models, are qualitatively not very different from each other, but, in general, great care must be taken in

selecting the 'best' alignment. Experimental information, when available, or simply pure chemical intuition should in our opinion always be part of the 'best' model selection process.

Despite the conceptual differences and different performances of the CoMFA and interaction energy approaches, some similarities can be found. For instance, when all 29 inhibitors charged by the CHARMM set of charges are considered, both CoMFA and the interaction energy-based method lead to the same conclusion, i.e., no significant correlation can be found between the experimental and the predicted  $pK_i$  values. Our studies clearly confirm that in both methods electrostatics plays a crucial role and calibrating work is required to obtain significant correlation coefficients.

## Conclusions

In this study we have applied CoMFA and an interaction energy-based approach on a database of thrombin-inhibitor complexes. In CoMFA the effects of several parameters including charge models, alignment rules and inclusion of hydrophobic fields on the quality of the models, were analysed. The best CoMFA model was obtained when the Mulliken HF 6-31G set of charges was used in combination with a pharmacophore-based type of alignment rule (the Centroid alignment). Surprisingly, the alignment based on the experimentally determined 3D structures of the thrombin-inhibitor complexes was – in statistical terms – not the best model. The inclusion of a hydrophobic field did not lead to improved models for our database. The interaction energy study involved an energy minimisation of the inhibitors with and without (harmonical) constraints on the surrounding protein. Different values of a distance dependent or constant dielectric variable were also considered. The best results were obtained by damping electrostatics and by keeping a rigid protein environment during the energy minimizations of the complexes. The interaction energy-based method only yielded statistically significant correlations for high resolution structures.

We feel that there are some important lessons to be learned from the current study. First of all, it is clear that the statistically best model does not necessarily lead to the best description of the system or, in other words, the highest predictive power. The fact that the pharmacophore-based alignment method and not the experimentally observed modes of binding yields the

highest correlation illustrates this point. It is also clear that the interaction energy-based approach is a relatively poor scoring method that only works to some extent for high resolution structures. This finding confirms the continuous need in drug design projects for experimental verification of models by determining high-resolution structures of key compounds in complex with their receptor. As shown above, the statistical engine behind CoMFA is very powerful, and we feel that the high correlations found may sometimes lead to unrealistic expectations regarding the predictive power of the method. In this respect, the simple statistics used to evaluate the interaction energy-based approach may provide a more realistic picture on the strengths and weaknesses of the method. Finally, we note that in the practice of drug design it is most effective to apply various drug design techniques and approach the problem from various points of view.

### Acknowledgements

We thank Dr. W. van der Saal and Dr. R. Engh, Boehringer Mannheim, for making available experimental data and the 3D structures of the test database.

### References

1. Kuntz, I.D., Meng, E.C. and Shoichet, B.K., *Acc. Chem. Res.*, 27 (1994) 117.
2. Charifson, P.S. and Kuntz, I.D., In Charifson, P.S. (Ed.), *Practical Application of Computer-Aided Drug Design*, Marcel Dekker, New York, NY, 1997, pp. 1–37.
3. Böhm, H.J., *Curr. Opin. Biotechnol.*, 7 (1996) 433.
4. Knegtel, R.M.A. and Grootenhuys, P.D.J., *Perspect. Drug Discov. Des.*, 9/10/11 (1998) 99.
5. Ajay, Murcko, M.A. and Stouten, P.F.W., In Charifson, P.S. (Ed.), *Practical Application of Computer-Aided Drug Design*, Marcel Dekker, New York, NY, 1997, pp. 355–410.
6. Timms, D. and Wilkinson, A.J., In van Gunsteren, W.F., Weiner, P.K. and Wilkinson, A.J. (Eds.), *Computer Simulation of Biomolecular Systems*, Vol. 3, KLUWER/ESCOM, Dordrecht, 1997, pp. 466–493.
7. Cramer III, R.D., Patterson, D.E. and Bunce, J.D., *J. Am. Chem. Soc.*, 110 (1988) 5959.
8. Devillers, J. (Ed.) *Neural Networks in QSAR and Drug Design*, Academic Press, London, U.K., 1996.
9. Grootenhuys, P.D.J. and van Helden, S.P., In Wipff (Ed.), *Computational Approaches in Supramolecular Chemistry*, Kluwer Academic Publishers, Dordrecht, 1994, pp. 137–149.
10. Grootenhuys, P.D.J. and van Galen, P.J.M., *Acta Crystallogr.*, D51 (1995) 560.
11. Ajay and Murcko, M.A., *J. Med. Chem.*, 38 (1995) 4953.
12. SYBYL 6.3, Tripos Assoc., St. Louis, MO.
13. HINT, version 2.1S, Kellogg, G.E. and Abraham, D.J., *J. Med. Chem.*, 34 (1991) 758.
14. Kroemer, R.T., Hecht, P. and Liedl, K.R., *J. Comput. Chem.*, 17 (1996) 1296.
15. Gasteiger, J. and Marsili, M., *Tetrahedron*, 36 (1980) 3219.
16. Mulliken, R.S.L., *J. Chem. Phys.*, 23 (1955) 1833.
17. Stewart, J.J.P., MOPAC 6.0, Quantum Chemical Program Exchange 455, 1990.
18. a. Singh, U.C. and Kollman, P.A., *J. Comput. Chem.*, 5 (1984) 129.  
b. Besler, B.H., Merz, K.M., Jr. and Kollman P. A., *J. Comput. Chem.*, 11 (1990) 431.
19. GAUSSIAN 94, Frisch, M.J., Trucks, G.W., Schlegel, H.B., Gill, P.M.W., Johnson, M.A., Robb, M.A., Cheeseman, J.R., Keith, T.A., Petersson, G.A., Montgomery, J.A., Raghavachari, K., Al-Laham, M.A., Zakrzewski, V.G., Ortiz, J.V., Foresman, J.B., Cioslowski, J., Stefanov, B.B., Nanayakkara, A., Challacombe, M., Peng, C.Y., Ayala, P.Y., Chen, W., Wong, M.W., Andres, J.L., Repogle, E.S., Gomperts, R., Martin, R.L., Fox, D.J., Binkley, J.S., Defrees, D.J., Baker, J., Stewart, J.P., Head-Gordon, M., Gonzalez, C. and Pople, J.A., Gaussian, Inc., Pittsburgh, PA, 1995.
20. Perry, N.C. and van Geerestein, V.J., *J. Chem. Inf. Comput. Sci.*, 32 (1992) 607.
21. Kellogg, G.E., *Med. Chem. Res.*, 7:6/7 (1997) 417.
22. Oprea, T.I. and Waller, C.L., *Rev. Comput. Chem.*, 11 (1997) 127.
23. Klebe, G., Mietzner, T. and Weber, F., *J. Comput.-Aided Mol. Design*, 8 (1994) 751.