Local Indices for Similarity Analysis (LISA)—A 3D-QSAR Formalism Based on Local Molecular Similarity

Jitender Verma,[†] Alpeshkumar Malde,[†,§] Santosh Khedkar,[†,⊥] Radhakrishnan Iyer,[‡] and Evans Coutinho*[,†]

Department of Pharmaceutical Chemistry, Bombay College of Pharmacy, Kalina, Santacruz (E), Mumbai 400 098, India, and Spring Bank Pharmaceuticals, Inc., 113 Cedar Street, Milford, Massachusetts 01757

Received June 26, 2009

A simple quantitative structure activity relationship (QSAR) approach termed local indices for similarity analysis (LISA) has been developed. In this technique, the global molecular similarity is broken up as local similarity at each grid point surrounding the molecules and is used as a QSAR descriptor. In this way, a view of the molecular sites permitting favorable and rational changes to enhance activity is obtained. The local similarity index, calculated on the basis of Petke's formula, segregates the regions into "equivalent", "favored similar", and "disfavored similar" (alternatively "favored dissimilar") potentials with respect to a reference molecule in the data set. The method has been tested on three large and diverse data sets—thrombin, glycogen phosphorylase *b*, and thermolysin inhibitors. The QSAR models derived using genetic algorithm incorporated partial least square analysis statistics are found to be comparable to the ones obtained by the standard three-dimensional (3D)-QSAR methods, such as comparative molecular field analysis and comparative molecular similarity indices analysis. The graphical interpretation of the LISA models is straightforward, and the outcome of the models corroborates well with literature data. The LISA models give insight into the binding mechanisms of the ligand with the enzyme and allow fine-tuning of the molecules at the local level to improve their activity.

INTRODUCTION

Quantitative structure activity relationship (QSAR) remains one of the most widely used tools in the drug design arena. Since its first report by Hansch and co-workers[1−4] and subsequently by Free and Wilson,[5] the concept has been continuously revamped to overcome its many limitations. Likewise, molecular similarity has also been extensively used in drug design, e.g., in the selection of analogs for chemicals, in the estimation of molecular properties, in the rational selection of candidates from large databases, and in some QSAR approaches.[6−9] Generally these studies use distance or angular information to characterize the level of resemblance between the molecules. The similarity between a pair of molecules is estimated based on the overlap of the analogous fields of the two molecules, taken as a sum over all components on a three-dimensional (3D) grid, using various 3D-molecular similarity indices.[10−14] The electron density, electric field, electrostatic potential, molecular lipophilicity potential, molecular fields, shape, etc., have all been used for similarity assessment.[6,7,15−18]

One of the most popular current methods of analyzing 3D molecular similarity is the technique of comparative molecular field analysis (CoMFA), where the shape information of the field is indirectly coded as attribute index numbers signifying the value of the field at the sampled grid point.[19] Comparative molecular similarity index analysis (CoMSIA) uses a SEAL-based molecular similarity index derived from steric, electrostatic, hydrophobic, hydrogen-bond acceptor and donor fields as 3D descriptors for QSAR analysis.[20] Shape similarity and differences in molecular shape analysis (MSA) are described quantitatively in terms of common-overlap steric volumes (CoSV) between pairs of molecules, by representing atoms as spheres of standard van der Waals radii.[21] Comparative molecular moment analysis (CoMMA) uses the lower order moments of the molecular mass and the charge distribution for comparison.[22] Molecular quantum similarity measure (MQSM) compares the first-order molecular density functions of the two molecules.[23] In Hopfinger's four-dimensional (4D)-QSAR method, the molecular similarity is estimated as a function of conformation, alignment, and atom type and is applied to study chiral and isosteric compounds as well as for the identification of common pharmacophores.[24]

With the exception of CoMSIA, all methods discussed above take into consideration the molecule as whole (i.e., a global approach) for calculating the molecular similarity. However, since these methods use integral similarity indices, they provide limited information about the spatial molecular features responsible for the variation of activity with the 3D structure. The local similarity measure overcomes this problem by estimating similarity at the local level and, thus, provides useful insights to rationally interpret the 3D-QSAR results based on the structural features. On the other hand, Hopfinger's method allows molecular similarity to be

* Corresponding author. Telephone: +91-22-26670871. Fax: +91-22-26670816. E-mail: evans@bcpindia.org.
† Bombay College of Pharmacy.
‡ Spring Bank Pharmaceuticals.
§ Present Address: School of Chemistry and Molecular Biosciences, University of Queensland, St. Lucia, QLD 4072, Australia.
⊥ Present Address: Laboratory for Drug Discovery and Target Validation, Department of Medicine, BIDMC-RN#227F, Harvard Medical School, Boston, MA 02215.

10.1021/ci900224u
Published on Web 12/08/2009

measured with respect to the whole molecule (global) as well as with respect to functional segments of the molecule (local). Another method that, in a way, considers molecular similarity is Comparative residue interaction analysis (CoRIA) which involves computing the interaction energy of each unit of the ligand (local) with individual residues in the receptor active site (local).[25−28] Similarly, the HomoSAR method is based on the calculation of a similarity index for every position (local) in each peptide in the data set by comparison against a reference peptide.[29]

In most cases of grid-based calculation of similarity indices, the potentials at all the grid points are summed up to give an overall global molecular similarity between the molecules. Herein, we report a method termed local indices for similarity analysis (LISA), which involves the calculation and the comparison of similarity between the molecules (in terms of a local similarity index, LSI) at each and every grid point in a 3D space. This permits rational, site-specific alteration of the molecules to improve their activity. The local similarity indices at all grid points surrounding the molecules are then used as descriptors in the QSAR formalism, and the models are derived with genetic algorithm incorporated partial least square analysis (G/PLS) statistics. The LISA methodology has been tested and validated on three large and diverse data sets—thrombin, glycogen phosphorylase *b* (GPB), and thermolysin inhibitors.

LISA FORMALISM

The principles of LISA formalism are based on the following facts:

1. The reference molecule is one which, in addition to possessing the appropriate size and shape complementary to the entire active site volume, also embraces the essential pharmacophoric features necessary to interact with the crucial receptor residues. Therefore, one of the prerequisites and decisive factors for the LISA methodology is the selection of a suitable reference molecule, with respect to which the local similarity index (LSI) for the remaining molecules can be calculated. According to the general convention, the most active molecule in a given series can be considered as the reference molecule. For decades, researchers have been successfully using the concept of a reference molecule for drug design, particularly in the areas of molecular alignment and pharmacophore modeling, and it is being followed here.[30]
2. The biological activity is directly related to the structural properties of the molecule, and the molecular structure can be measured and represented with a set of numbers usually called descriptors. Molecules with common or related structures generally have similar physicochemical properties by virtue of which they have similar binding modes and consequently comparable biological activities. The reverse also holds true.
3. Structural properties, which lead to an observed biological response, are most commonly determined by the nonbonding (or non-covalent) forces, mainly steric and electrostatic.
4. Not each segment of the reference molecule (the most potent molecule in the series) contributes toward its biological activity, in a positive or additive manner.
5. Not each segment of the remaining molecules (other than the reference one) is responsible for their relatively moderate or lower activity. They may contain groups that may otherwise contribute positively to the activity but are not being exploited to that extent due to their current location/orientation in the molecule or due to the presence/absence of other groups in the molecule.
6. The primary objective is to extract not only from the reference molecule but also from the remaining ones in the series, the crucial features which contribute positively to the biological activity and to utilize them to design novel compounds and/or optimize existing leads. Hence LISA formalism allows comparison of the degree of similarity between the molecules at the local level (at each and every grid point in a 3D system) rather than at the global level, as is the case in the common 3D-QSAR methods.

The LSI at each grid point for the molecules under consideration is calculated by an adaption of the formula given by Petke in 1993 for 3D-molecular similarity:[31]

$$\mathrm{LSI}_i = 2P_{\mathrm{A}i}P_{\mathrm{B}i}/(P^2_{\mathrm{A}i} + P^2_{\mathrm{B}i}) \quad (1)$$

LSI_i = Local similarity index at the grid point *i*.

$P_{\mathrm{A}i}$ = Interaction energy (steric/electrostatic/lipophilic) between the probe atom and the molecule A (reference molecule) at the grid point *i*.

$P_{\mathrm{B}i}$ = Interaction energy (steric/electrostatic/lipophilic) between the probe atom and the molecule B (target molecule) at the grid point *i*.

The numerator in eq 1 is the overlap of the property (*P*) of molecules A and B, while the denominator is the normalization factor. This index measures similarity at discrete points in space, thereby allowing more control over the nature of molecular similarity calculations. The use of such discrete similarities permits both graphical[32] as well as statistical[31] analysis to be undertaken on the resultant similarity grid. The overall molecular similarity can be obtained as the ratio of the sum of similarities over all points in space to the total number of points considered, thereby leading to an average similarity over the measured space.

The standard probe atoms, like H^+ for electrostatic interactions and CH_3 for steric interactions, are used with an energy cutoff of ± 30.0 kcal/mol. The method is flexible enough to accommodate additional standard validated probe atoms. A typical data set gives LSI values within the range of +1.0 to −1.0 (because of the normalization factor), with the LSI values of the reference molecule equal to unity at all grid points. A positive value for the LSI at a particular grid point indicates "similarity", while negative values for the LSI suggest "dissimilarity" with the reference molecule. The magnitude of the LSI indicates the extent of similarity or dissimilarity. However, it is the sign of the coefficient of the LSI descriptors in the QSAR equations which will ultimately govern whether an increase (favor) or decrease (disfavor) in the similarity of the target molecule with the reference is needed in order to improve the biological activity of the target molecule. A cross-section of the LISA descriptors is given in Scheme 1.

COMPUTATIONAL DETAILS

The application and outcome of the LISA formalism has been discussed comprehensively in the subsequent sections for only the thrombin data set; due to space restrictions, its substantiation on the other two data sets, GPB and thermolysin inhibitors, has been conferred only briefly. The proce-

Scheme 1

Molecule	pK_i	Interaction energy values at each GRID point with the probe atoms					
		Ele_1	Ele_2	…	Ste_1	Ste_2	…
Reference	11.85	-0.138	-0.158	…	0.000	-0.001	…
Mol 1	6.84	-0.049	-0.082	…	-0.003	-0.005	…
Mol 2	7.10	0.539	0.562	…	0.000	-0.001	…
…	…	…	…	…	…	…	…
Mol n	10.80	0.509	0.521	…	-0.001	-0.002	…

$$LSI_i = 2P_{Ai}P_{Bi} / (P_{Ai}^2 + P_{Bi}^2)$$

Molecule	pK_i	LISA descriptors (LSI_i)					
		Ele_1	Ele_2	…	Ste_1	Ste_2	…
Reference	11.85	1.000	1.000	…	1.000	1.000	…
Mol 1	6.84	0.627	0.819	…	0.321	0.222	…
Mol 2	7.10	-0.481	-0.522	…	1.000	0.999	…
…	…	…	…	…	…	…	…
Mol n	10.80	-0.506	-0.556	…	0.587	0.476	…

dures of the descriptor calculation, the validation controls, and the other statistical parameters (for the LISA, CoMFA, and CoMSIA models) are similar in all three cases and consequently are explained in detail solely for the thrombin data set.

Thrombin Data Set. Thrombin is a serine protease enzyme with a fundamental role in blood coagulation. It cleaves various substrates involved in coagulation and activates cell surface receptors via a novel proteolytic action.[33] Thrombin stimulates aggregation and secretion in blood platelets at the site of vascular injury and also has inflammatory and reparative actions, stimulating chemotaxis in monocytes, proliferation of fibroblasts and lymphocytes, and inducing endothelium-dependent relaxation of blood vessels.[34] Inhibitors of the thrombin enzyme are of immense importance in primary and secondary prevention of deep vein thrombosis, pulmonary embolism, myocardial infarctions, and strokes, in those who are predisposed.[35,36]

A set of 66 high-quality X-ray structures of the non-covalent inhibitors complexed with the enzyme thrombin in the Protein Data Bank (PDB)[37] was used. This data set was chosen because of the good structural diversity and the wide range of potency of these inhibitors. Based on chemical (defined by the Daylight fingerprint) and biological (based on pK_i values) diversity, the data set was divided into a training set of 44 molecules and a test set of 22 molecules. The pK_i values ranged from 3.00 to 11.85, spanning ~9 log units. The ligands in the protein complexes were rectified for correct atom and bond types in Sybyl v7.1,[38] and all hydrogen atoms were added. Partial atomic charges were assigned to the atoms, and the structures were subsequently minimized with the MMFF94 force field,[39] keeping heavy atoms fixed until the gradient reached 0.01 kcal/mol/Å. Molecule 2CF8 with the highest resolution (1.3 Å) was chosen as the template, and the remaining PDB structures were superimposed on it. Subsequent removal of the protein gave 66 ligands aligned according to their bioactive conformation in the active site (Figure 1). This configuration of the 66 molecules was used for further calculations. A standard grid with 2.0 Å spacing was constructed around the aligned molecules using Cerius2[40] running on a Silicon Graphics O2 workstation. The total grid points were found to be 2184. The electrostatic and steric interaction energies at each grid point were calculated with the standard H+ and CH_3 probes, respectively. The LSI descriptors (4368 columns, 2184 columns each for the electrostatic and steric fields) were calculated from this data using eq 1, as depicted in Scheme 1, and utilized for subsequent statistical analysis. The most

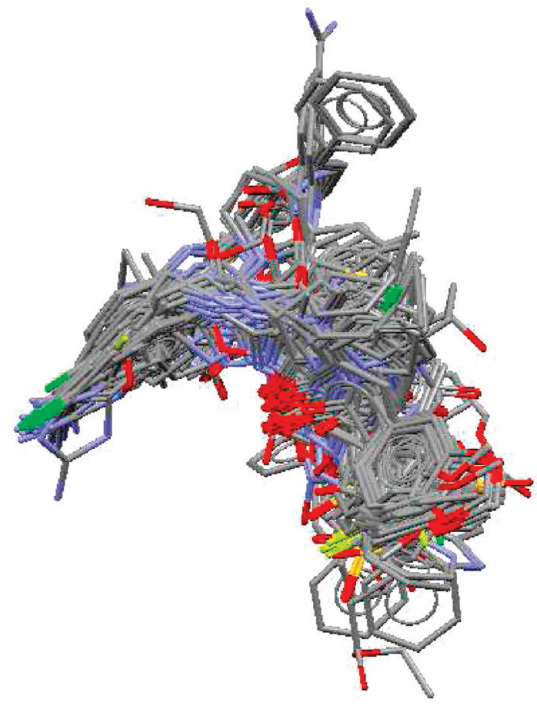

Figure 1. Alignment of the molecules in the thrombin data set.

Table 1. Biological Activities and PDB Codes of Thrombin Inhibitors Used in the Study

mol ID	PDB code	pK_i	set	mol ID	PDB code	pK_i	set
1	1BHX	6.84	training	34	1O2G	6.12	training
2	1BMM	7.10	training	35	1OYT	7.24	test
3	1BMN	8.44	test	36	1QBV	5.39	test
4	1C1U	8.25	training	37	1RIW	7.55	training
5	1C1V	8.64	training	38	1SB1	6.89	training
6	1C4U	10.37	training	39	1SL3	11.85	training
7	1C4V	10.80	training	40	1T4U	7.68	training
8	1C4Y	7.92	training	41	1TA2	8.52	training
9	1C5N	4.70	test	42	1TA6	9.13	training
10	1D6W	7.96	training	43	1TOM	8.31	training
11	1D9I	9.11	training	44	1UVS	5.40	test
12	1DWB	6.52	test	45	1VZQ	7.44	test
13	1DWC	7.41	test	46	1WAY	3.40	training
14	1DWD	7.59	test	47	1WBG	3.00	training
15	1EB1	10.43	test	48	1YPE	8.10	training
16	1FPC	7.00	training	49	1YPG	8.00	training
17	1G30	6.85	training	50	1YPJ	7.02	test
18	1G32	6.11	training	51	1YPL	4.97	training
19	1GHV	7.35	training	52	1YPM	5.72	training
20	1GHW	5.40	training	53	1Z71	9.18	test
21	1GHY	8.10	training	54	1ZGI	8.34	training
22	1GJ4	4.22	training	55	1ZGV	6.00	training
23	1GJ5	5.26	test	56	1ZRB	9.11	training
24	1K21	8.38	test	57	2ANM	9.00	test
25	1K22	8.40	training	58	2C8W	8.43	training
26	1KTS	8.03	training	59	2C8X	6.66	training
27	1KTT	5.82	training	60	2C8Y	4.00	training
28	1MU6	8.38	training	61	2C90	3.48	training
29	1MU8	9.00	test	62	2C93	4.92	test
30	1MUE	8.64	test	63	2CF8	8.10	test
31	1NM6	10.05	test	64	2CF9	7.82	training
32	1NT1	8.89	training	65	2FEQ	8.86	test
33	1NZQ	7.96	training	66	2FES	9.01	test

active molecule, i.e., 1SL3 (pK_i 11.85) was used as the reference molecule (R1). In order to investigate the influence of the reference molecule in the LISA formalism, few other molecules with decreasing biological activities were also used as the reference to construct various QSAR models. They are 1C4U (R2, pK_i 10.37), 1D6W (R3, pK_i 7.96), 1KTT (R4, pK_i 5.82), and 1WBG (R5, pK_i 3.00). The biological activity and PDB codes of the molecules considered in this study are given in Table 1, and the ligand structures are shown in the Supporting Information (Table S1).

GPB Data Set. The enzyme glycogen phosphorylase catalyzes the first (or the rate-limiting) step in the phosphorolysis of glycogen to glucose-1-phosphate. Inhibition of this enzyme prevents the conversion of glycogen into glucose and, thus, helps in attenuating hyperglycemia in type 2 diabetes, by shifting the balance between glycogen breakdown and synthesis in favor of the latter.[41] Glycogen phosphorylase exists in two interconvertible states: the inactive *b* form (GPB), which gets phosphorylated to its active form *a* (GPA) during the glycogen metabolism. Furthermore, phosphorylase *a* and *b* each exist in two subforms—a T (tense) inactive state and a R (relaxed) state. Phosphorylase *b* is usually in the T state, inactive due to the physiological presence of ATP and glucose-6-phosphate, and phosphorylase *a* is normally in the R state (active).[42]

A data set of 66 GPB inhibitors assembled by Klebe et al.[43] was used for validation of the LISA methodology. For the sake of comparison with the original publication, the alignment reported by Klebe et al. (Figure 2) was used as such. The crystal structures of inhibitors complexed with GPB (PDB codes: 1A8I, 1AXR, 1B4D, 1E1Y, 1NOJ, 1NOK, 2GPB, 2PRJ, 3GPB, 4GPB, 5GPB, and 6GPB), available in the protein data bank, were used as templates to construct and superimpose the remaining GPB inhibitors. This was followed by energy minimization inside the binding pocket of 2GPB held rigid with the MAB force field as implemented in the program MOLOC.[44] As reported by Klebe et al.,[43] the same group of 58 compounds was used for the training set, and 8 compounds were used for the test set. The experimentally determined activities (pK_i values) of these inhibitors, which cover around 5.5 log units (from 1.30 to 6.80), are given in Table 2, while their structures are shown in the Supporting Information (Table S2).

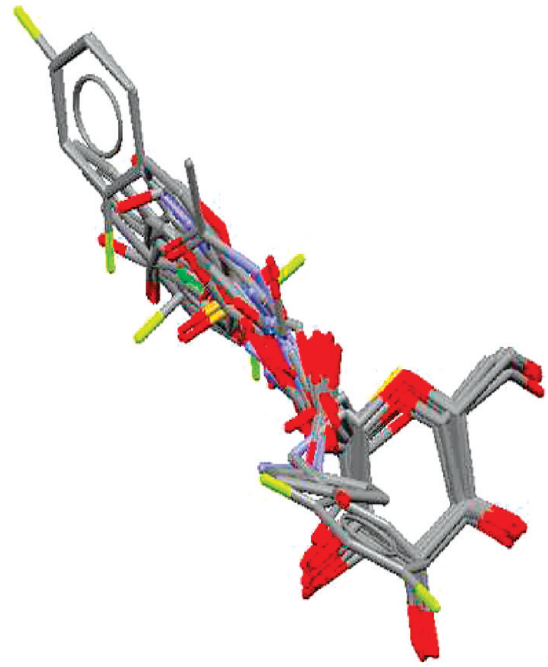

Figure 2. Alignment of the molecules in the GPB data set.

Thermolysin Data Set. Thermolysin is an extracellular, thermostable, zinc-containing, neutral endopeptidase/metalloproteinase enzyme produced by the gram-positive bacteria *Bacillus thermoproteolyticus*.[45] Thermolysin specifically catalyzes the hydrolysis of peptide bonds containing hydrophobic amino acids. As a member of zinc metalloproteinases, the enzyme has served as a model system for the development of inhibitor design strategies that can be translated to zinc proteases of physiological importance.[46] Inhibitors of thermolysin have considerable potential as therapeutic agents in the biosynthesis and metabolism of different bioactive peptides.

The data set employed for validation of the LISA formalism consisted of 76 thermolysin inhibitors compiled by Klebe et al.[20,43] Like in case of GPB, the alignment reported by Klebe et al. (Figure 3) was used without any modification. The crystal structures of thermolysin−inhibitor complexes deposited in the protein data bank (PDB codes: 1TLP, 1TMN, 2TMN, 4TLN, 4TMN, 5TLN, 5TMN, and 6TMN) were used as templates for building and aligning the remaining thermolysin inhibitors. Finally, the compounds were minimized inside the binding pocket of 1TLP held rigid with the MAB force field in the program MOLOC.[44] The

Table 2. Biological Activities of GPB Inhibitors Used in the Study

mol ID	pK_i	set	mol ID	pK_i	set
01	2.8	training	34	3.7	training
02	1.3	training	35	3.4	training
03	1.9	training	36	4.4	training
04	1.6	training	37	4.5	training
05	1.6	training	38	4.4	training
06	1.6	training	39	4.4	training
07	1.8	training	40	4.1	training
08	2.1	training	41	4.0	training
09	2.8	training	42	3.9	training
10	1.7	training	43	3.4	training
11	1.8	training	44	2.3	training
12	2.3	training	45	4.8	training
13	2.4	training	46	3.5	training
14	1.5	training	47	3.0	training
15	1.8	training	48	2.7	training
16	2.3	training	49	2.6	training
17	2.0	training	50	2.4	training
18	2.1	training	51	2.4	training
19	1.8	training	52	1.7	training
20	3.4	training	53	2.1	training
21	2.7	training	54	2.9	training
22	3.0	training	55	2.6	training
23	2.1	training	56	5.5	training
24	2.6	training	57	6.8	training
25	2.5	training	58	4.2	training
26	1.7	training	59	2.1	test
27	3.4	training	60	2.1	test
28	1.9	training	61	1.8	test
29	1.4	training	62	1.8	test
30	2.3	training	63	1.6	test
31	1.6	training	64	4.1	test
32	3.7	training	65	3.3	test
33	2.6	training	66	4.4	test

same group of 61 and 15 molecules were used for formulation of the training and test sets, respectively. The activities (pK_i values) of the thermolysin inhibitors (Table 3) are spread over a sufficiently large range from 0.52 to 10.17, spanning around 9.7 log units. The structures of these inhibitors are shown in the Supporting Information (Table S3).

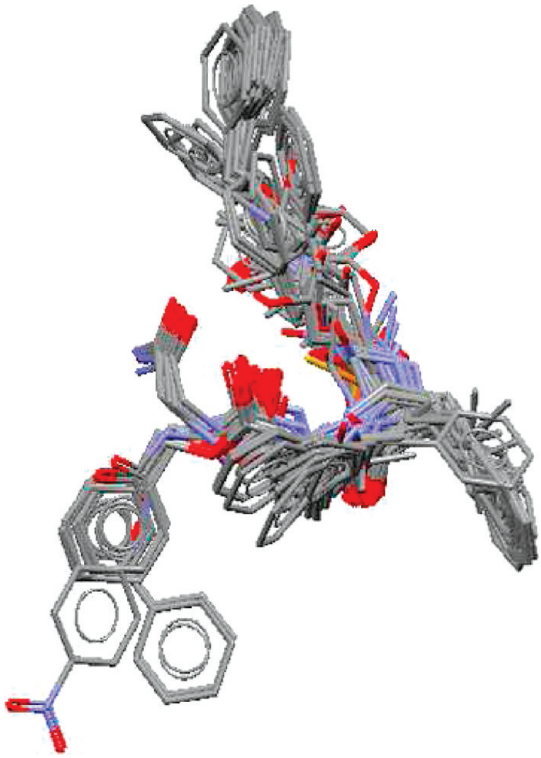

Figure 3. Alignment of the molecules in the thermolysin data set.

Table 3. Biological Activities of Thermolysin Inhibitors Used in the Study

mol ID	pK_i	set	mol ID	pK_i	set
01	2.47	training	39	5.74	training
02	6.12	training	40	6.07	training
03	7.28	training	41	5.74	training
04	8.82	training	42	3.46	test
05	5.84	training	43	10.17	training
06	3.29	training	44	7.35	training
07	2.51	training	45	4.41	training
08	0.52	training	46	3.03	training
09	2.47	training	47	3.6	training
10	7.47	training	48	1.68	training
11	7.96	training	49	4.89	training
12	6.22	training	50	2.65	training
13	5.55	training	51	6.39	training
14	6.66	training	52	7.73	test
15	5.77	training	53	7.18	test
16	2.42	training	54	6.52	test
17	2.54	training	55	6.74	training
18	6.37	training	56	5.85	test
19	6.18	training	57	7.78	training
20	6.18	training	58	7.12	test
21	4.70	training	59	6.57	test
22	6.32	training	60	8.04	training
23	3.72	training	61	6.12	test
24	2.96	training	62	4.89	training
25	3.38	training	63	4.27	test
26	4.06	training	64	3.64	test
27	7.55	training	65	5.05	training
28	4.10	training	66	3.18	test
29	7.72	test	67	4.32	training
30	0.52	training	68	2.51	test
31	5.59	training	69	6.17	training
32	4.14	training	70	3.66	test
33	2.79	test	71	4.62	training
34	6.44	training	72	6.32	training
35	0.52	training	73	4.52	training
36	5.64	training	74	4.38	training
37	5.16	training	75	3.42	training
38	2.37	training	76	5.57	training

Statistics. The LSI descriptors were scaled to zero mean and unit standard deviation, by subtracting each value in a given column from the column mean and then dividing by the standard deviation of that column. This was done to assign equal weight to all the descriptors and to place them on the same platform for a meaningful statistical analysis. G/PLS (genetic function approximation GFA[47] in conjunction with partial least square PLS[48]), as implemented in Cerius2,[40] was used as the chemometric method to derive the QSAR equations. The GFA algorithm develops an initial population of individuals; a fitness function, which is a measure of least-square error, is then applied as an estimate of the quality of each individual. Individuals with the best fitness scores are allowed to mate and propagate their genetic material to offspring through the crossover and/or mutation operations. After repeatedly performing these steps, the average fitness of the individuals in the population increases, as good combination of "genes" (LSI descriptors in the present case) are discovered and spread through the population. The best combinations of the LSI descriptors are then subjected to PLS for regression analysis. Only linear terms were used to develop the QSAR models, and the optimal number of components selected was six at which the cross-validated r^2

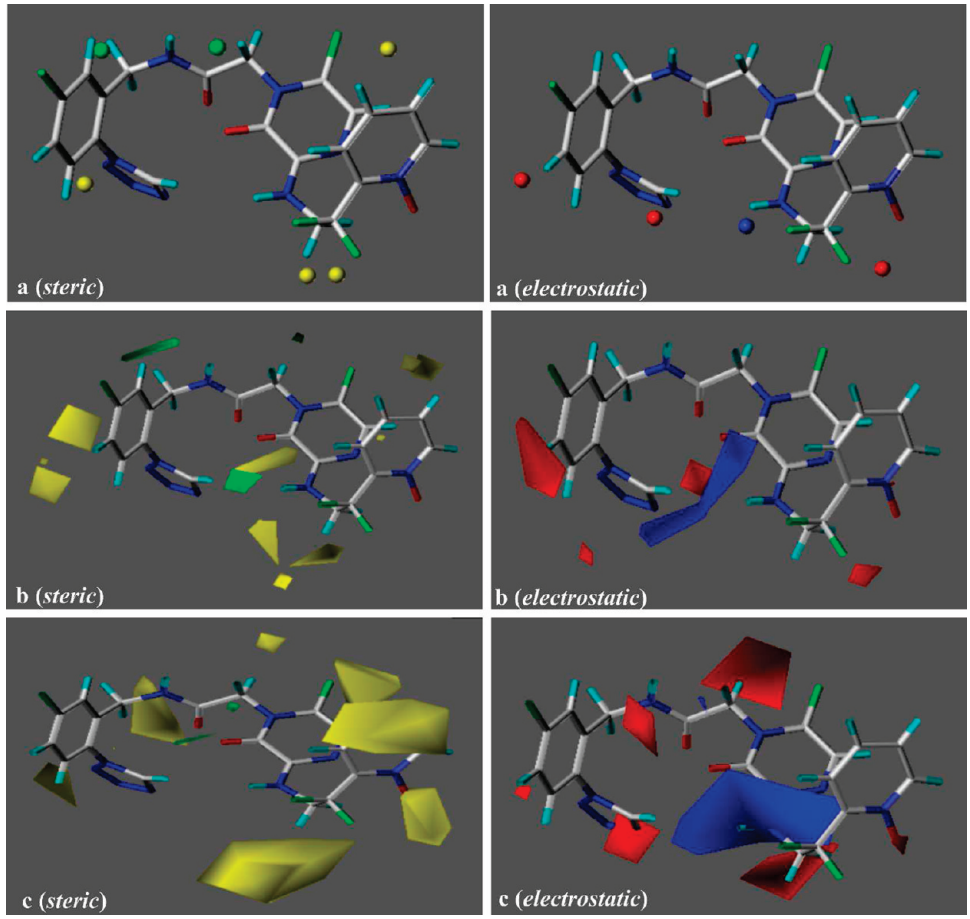

Figure 4. Graphical representation of the steric and electrostatic contours/spheres around thrombin inhibitor 39 for the best LISA (a), CoMFA (b), and CoMSIA (c) models, respectively. Green color signifies that steric bulk is favored at these sites, whereas yellow color shows that the bulk is disfavored. Similarly red color indicates the regions where electronegative substituents are favored, while blue color is associated with the positions where electropositive substituents are favored.

(i.e., q^2) was found to be the maximum. In order to facilitate simple interpretation and easy use of the models in designing new ligands, the length of the equations was confined to six terms (including the constant) for which the r^2 and predictive residual sum of squares (PRESS) values were found to be the optimum. The number of generations was set to 10000, and the population size was set to 500. Crossover and mutation probabilities of 50% (default settings) were employed with a smoothness parameter of 1.0 (the smoothness function penalizes the equations on their size and, thus, controls the bias in the scoring factor between equations with different numbers of terms). Internal cross-validation using randomization at 99% confidence interval, leave-one-out and leave-group-out (in groups of 5) was performed to calculate the q^2 for the QSAR models.[49] To further assess the robustness and statistical confidence of the derived models, bootstrapping analysis for 100 runs was carried out. Externally, the best QSAR models were evaluated for their predictive power on the test set molecules.

CoMFA and CoMSIA Models. For comparison of the LISA formalism, CoMFA and CoMSIA models were developed for the same training and test set molecules, using only the steric and electrostatic fields in Sybyl v7.1.[38] The CoMFA and CoMSIA models are generally represented as 3D "coefficient contours". Colored contours in the map represent the areas in 3D space where changes in the steric and electrostatic field values of a compound correlate strongly with the concomitant changes in its biological activity. The steric and electrostatic contour plots of the best CoMFA and CoMSIA models for the thrombin data set are shown in Figure 4b and c, respectively. For ease of comparison, the results of the LISA models have also been analyzed and represented graphically similar to that of the colored steric and electrostatic contour plots of CoMFA and CoMSIA models, except that contours have been replaced with the corresponding colored spheres, as shown in Figure 4a.

RESULTS AND DISCUSSION

The conformations of the thrombin inhibitors have been derived from the crystal complexes of the inhibitors bound to the enzyme. All the inhibitors are, thus, in their "bioactive" conformation. Superposition of the protein complexes gave all inhibitors in a bound orientation in the active site. Thus, all problems associated with the binding pose, the conformational preference, and the resulting overall alignment are avoided in the data set used in the present study. This data set is also diverse in terms of chemical structure and biological activity. All these factors contribute to the confidence in the derived 3D-QSAR models.

The best QSAR equations generated by the LISA methodology for the thrombin data set are given in Table 4 and displayed graphically in Figure 4a. The best LISA models for the GPB and thermolysin data sets are provided

Table 4. Best LISA Models Developed for the Thrombin Dataset Using Different Reference Molecules[a]

LISA model	reference molecule	pK_i	best QSAR equation
R1	1SL3	11.85	pK_i = 4.54 + 3.86 (Ste_1227) − 1.11 (Ste_718) + 6.86 (Ele_1077) − 1.05 (Ste_1131) − 1.25 (Ele_935)
R2	1C4U	10.37	pK_i = 4.85 + 3.56 (Ste_916) + 3.86 (Ele_1481) + 4.54 (Ste_730) + 2.15 (Ste_1306) − 8.59 (Ste_1398)
R3	1D6W	7.96	pK_i = 4.60 − 3.49 (Ele_1755) + 2.46 (Ste_718) + 4.50 (Ste_647) + 2.59 (Ele_779) − 2.70 (Ele_568)
R4	1KTT	5.82	pK_i = 3.58 − 1.18 (Ele_1038) − 1.26 (Ste_1345) + 6.98 (Ste_1086) − 1.16 (Ste_1819) − 1.15 (Ele_754)
R5	1WBG	3.00	pK_i = 3.57 + 2.45 (Ele_624) − 3.59 (Ele_702) − 2.46 (Ste_1190) − 2.24 (Ste_1070) + 5.27 (Ste_886)

[a] Positive coefficients signify favored, while negative indicates disfavored similarity between the target and reference molecules. "Ste" refers to the steric and "Ele" refers to the electrostatic interactions. For e.g., +3.86 (Ste_1227) indicates favored steric similarity at grid point 1277, and −3.49 (Ele_1755) indicates disfavored electrostatic similarity at grid point 1755 between the target and reference molecules.

Table 5. Best LISA Models Developed for the GPB and Thermolysin Datasets[a]

data set	best QSAR equation
GPB	pK_i = 2.89 − 1.48 (Ele_592) − 0.63 (Ste_489) + 1.15 (Ele_584) + 2.16 (Ele_341) + 0.30 (Ste_401)
thermolysin	pK_i = 5.04 − 0.66 (Ele_1916) + 0.83 (Ele_264) + 1.43 (Ste_2082) − 1.03 (Ste_1266) − 1.04 (Ste_1604)

[a] Positive coefficients signify favored, while negative indicates disfavored similarity between the target and reference molecules. "Ste" refers to the steric and "Ele" refers to the electrostatic interactions. For e.g., +1.43 (Ste_2082) indicates favored steric similarity at grid point 2082, and −1.48 (Ele_592) indicates disfavored electrostatic similarity at grid point 592 between the target and reference molecules.

Table 6. Statistical Parameters of LISA, CoMFA, and CoMSIA Models Derived for the Thrombin Dataset[a]

	LISA						
parameter	R1	R2	R3	R4	R5	CoMFA	CoMSIA
mol	1SL3	1C4U	1D6W	1KTT	1WBG	−	−
pK_i	11.85	10.37	7.96	5.82	3.00	−	−
r^2	0.866	0.856	0.848	0.852	0.851	0.875	0.799
N	6	6	6	6	6	4	2
LSE/SEE	0.476	0.514	0.543	0.528	0.524	0.557	0.658
r^2_{bs}	0.850	0.835	0.840	0.845	0.830	0.890	0.802
sd_{bs}	0.028	0.002	0.002	0.019	0.311	0.019	0.051
r^2_{random}	0.320	0.360	0.398	0.328	0.455	0.299	0.310
q^2(L-1-O)	0.765	0.761	0.705	0.809	0.782	0.706	0.510
q^2(L-5-O)	0.768	0.764	0.710	0.810	0.785	0.744	0.512
r^2_{pred}	0.555	0.470	0.428	0.529	0.251	0.488	0.437
p^2	0.540	0.452	0.408	0.513	0.225	0.470	0.417

[a] The r^2 = conventional (non cross-validated) correlation coefficient, N = optimum number of components, LSE = least-square error for LISA models, SEE = standard error of estimate for CoMFA and CoMSIA models, r^2_{bs} and sd_{bs} = mean values of correlation coefficient and standard deviation, respectively, after 100 runs of bootstrapping analysis, r^2_{random} = mean value of r^2 after randomization at 99% confidence interval, q^2(L-1-O) and q^2(L-5-O) = cross-validated correlation coefficient by leave-one-out and leave-five-out, respectively, r^2_{pred} = predictive correlation coefficient of test set, and p^2 = predictive correlation coefficient of test set, as defined by Vedani et al.[54]

in Table 5. The statistical parameters of the LISA, CoMFA, and CoMSIA models developed for the thrombin data set are given in Table 6, while those for the GPB and thermolysin data sets are shown in Table 7. Plots of the experimental vs predicted activity for the best LISA

Table 7. Statistical Parameters of LISA, CoMFA, and CoMSIA Models for the GPB and Thermolysin Datasets[a]

	GPB			thermolysin		
parameter	LISA	CoMFA	CoMSIA	LISA	CoMFA	CoMSIA
r^2	0.831	0.724	0.849	0.899	0.909	0.757
N	6	3	6	6	6	5
LSE/SSE	0.356	0.611	0.465	0.345	0.658	1.065
r^2_{bs}	0.850	0.779	0.893	0.920	0.939	0.829
sd_{bs}	0.299	0.532	0.374	0.305	0.530	0.885
r^2_{random}	0.301	0.299	0.386	0.281	0.398	0.325
q^2(L-1-O)	0.435	0.420	0.363	0.485	0.467	0.475
q^2(L-5-O)	0.405	0.399	0.357	0.450	0.401	0.440
r^2_{pred}	0.650	0.427	−0.512	0.615	0.164	0.149
p^2	0.643	0.417	−0.539	0.608	0.149	0.516

[a] The r^2 = conventional (non cross-validated) correlation coefficient, N = optimum number of components, LSE = least-square error for LISA models, SEE = standard error of estimate for CoMFA and CoMSIA models, r^2_{bs} and sd_{bs} = mean values of correlation coefficient and standard deviation, respectively, after 100 runs of bootstrapping analysis, r^2_{random} = mean value of r^2 after randomization at 99% confidence interval, q^2(L-1-O) and q^2(L-5-O) = cross-validated correlation coefficient by leave-one-out and leave-five-out, respectively, r^2_{pred} = predictive correlation coefficient of test set, and p^2 = predictive correlation coefficient of test set, as defined by Vedani et al.[54]

models of the thrombin data set are shown in Figure 5, with the corresponding plots for the GPB and thermolysin data sets in Figures 6 and 7, respectively. The results for these data sets, as revealed in Table 7, are quite comparable within them (LISA vs CoMFA/CoMSIA) as well as with some earlier studies.[20,43,50]

Though the LISA models derived using different reference molecules are within the requirements of statistical significance, but models derived using references R1 (1SL3) and R4 (1KTT), as evident in Table 6, are the best in terms of overall performance. The models R2 (1C4U) and R3 (1D6W) are also satisfactory, but model R5 (1WBG) has poor external predictive power, though its internal cross-validation results are quite acceptable. The probable reason for this is that the last model has been derived from a reference molecule with the lowest biological activity in the series. From Table 6, it can be rationalized that the choice of the reference molecule does not have a significant influence on the statistical parameters of the model or the internal predictive power; however, the choice of the reference does affect the external predictive power of the model. This can be judged from a significant decrease in the r^2_{pred} and p^2 values of the LISA

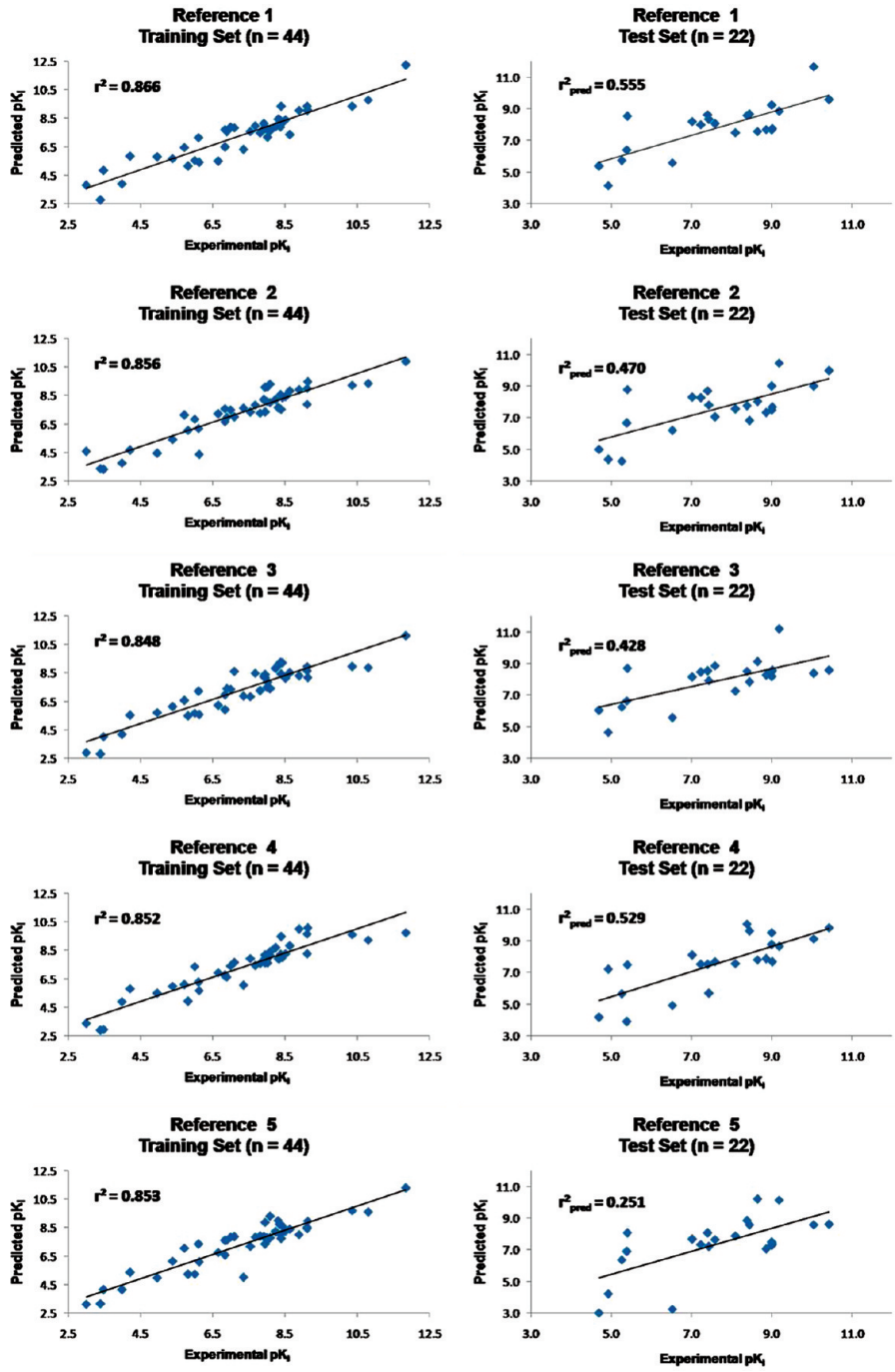

Figure 5. Plots of the experimental vs predicted pK_i values of the best LISA models for the thrombin data set.

models built using reference molecules with lower biological activities (e.g., R2, R3 and R5). Thus, it can be concluded that it is beneficial to develop LISA models using the most active molecule in the series as the reference.

The CoMFA model shows much better results than that of CoMSIA. On the other hand, the best LISA models (R1 and R4) and the CoMFA and CoMSIA models are quite comparable as far as the basic statistical parameters (r^2 and LSE/SEE) are concerned, except the number of components used in model building. Also, in the bootstrap analysis, the best LISA models performed unvaryingly compared to those of the CoMFA and CoMSIA models. However, the best LISA models outperform the models derived from CoMFA and CoMSIA in terms of the internal cross-validation q^2 (both L-1-O and L-5-O) as well as in the external prediction (r^2_{pred} and p^2).[49]

The thrombin data set compiled here is being analyzed by QSAR for the first time. There are literature reports of 3D-QSAR models obtained by various methods on different series of thrombin inhibitors,[51,52] many of which are also included in the present data set. Since, the training sets used in the construction of the models and the overall data set are not the same, a direct comparison of the models (in terms of statistics) with those obtained by LISA is not possible. However, the outcome of the models, i.e., the interpretation of the models in terms of chemistry can be compared. Hence, the LISA models are being compared with our own CoMFA and CoMSIA studies carried out on the same data set of thrombin inhibitors.

The LISA models developed from different reference molecules are given in Table 4. As mentioned earlier, the

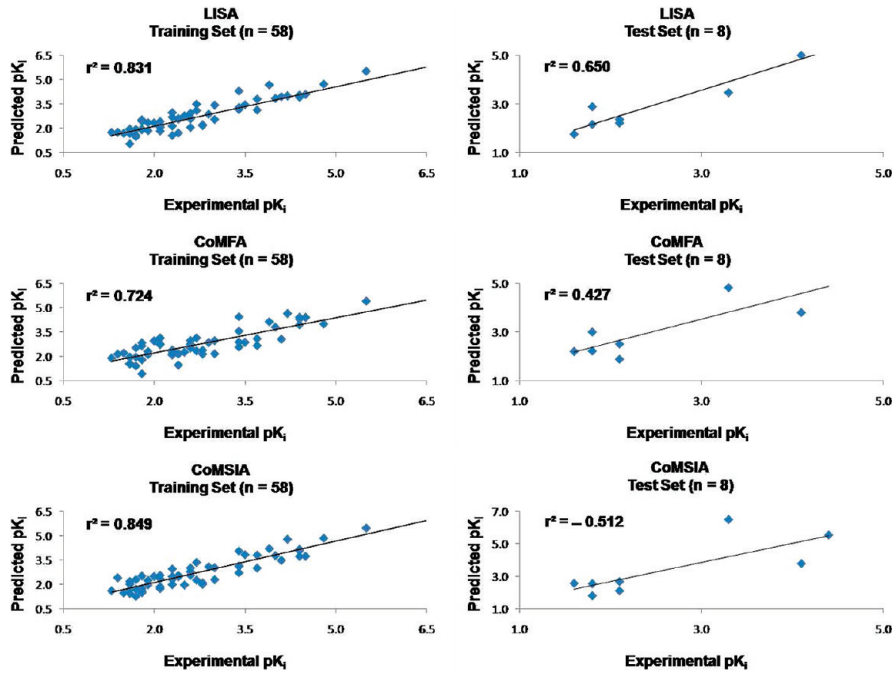

Figure 6. Plots of the experimental vs predicted pK_i values of the best LISA, CoMFA, and CoMSIA models for the GPB data set.

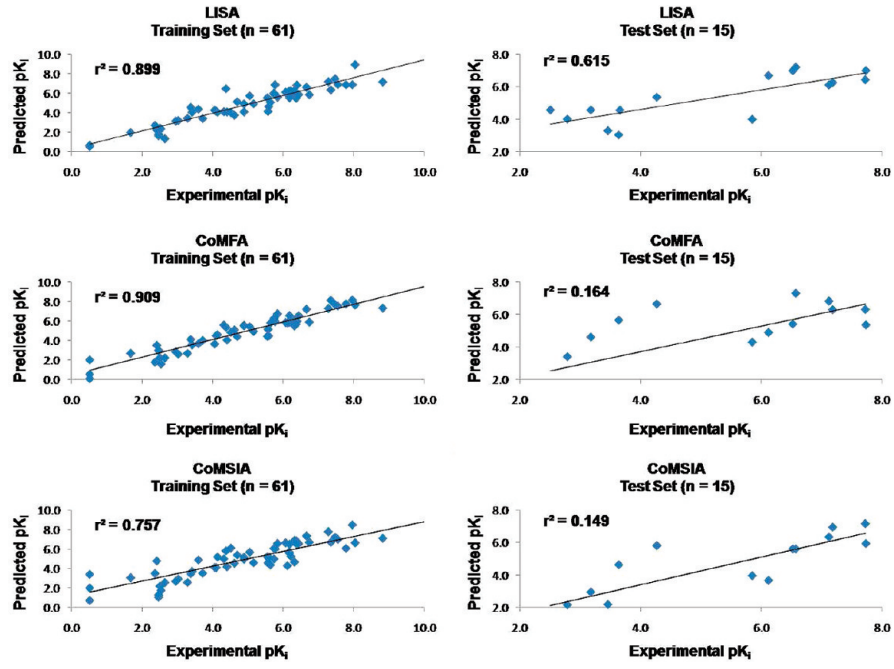

Figure 7. Plots of the experimental vs predicted pK_i values of the best LISA, CoMFA, and CoMSIA models for the thermolysin data set.

sign of the coefficient of the LSI descriptors in the QSAR equations will dictate whether an increase (favor) or decrease (disfavor) in the similarity of the target molecule with the reference compound will enhance activity. The descriptor with a positive coefficient indicates that similarity with the reference molecule should be preserved at that particular grid point in space. Likewise, a negative coefficient of a descriptor suggests that the property of the target molecule should be made "dissimilar" to the reference molecule, explicitly at that grid point. For the electrostatic probe (H^+), a positive coefficient of the LSI descriptor specifies "favored" electrostatic similarity with the reference molecule, while negative coefficient symbolizes "disfavored" electrostatic similarity. These regions are shown graphically as blue and red spheres, respectively, in Figure 2a. For the steric probe (CH_3), a positive coefficient of the descriptor represents "favored" steric similarity (green sphere), while a negative coefficient indicates "disfavored" steric similarity (yellow sphere) with the reference molecule (Figure 2a). For example, an analysis of the best LISA model R1 (Table 4) reveals that an enhancement in the activity of the target molecule can be obtained by increasing its steric similarity at grid point 1227 (positive coefficient of this descriptor in the equations) and by reducing its steric similarity at grid points 718 and 1131 (negative coefficient of these descriptors in the QSAR equations) with the reference molecule. The activity of target molecule can also be augmented by increasing its electrostatic similarity at grid point 1077 (positive coefficient in the

equations) and by decreasing its electrostatic similarity at grid point 935 (negative coefficient in the QSAR equations) with the reference molecule. Similarly, according to the second best LISA model R4, increasing steric similarity of the target molecule with the reference at grid point 1086 will improve its biological activity. Likewise, on account of the negative coefficients in the QSAR equations, rendering the target molecule dissimilar with respect to the reference molecule in terms of steric properties at grid points 1345 and 1819 and electrostatic properties at grid points 754 and 1038 will enhance activity.

The thrombin active site is divided into three distinct regions: the S1 pocket (Asp189, Ala190, Cys191, Glu192, Gly216 and Gly219, and Cys220), the S2 pocket (Tyr60B, Trp60E, and His57) and the S3 pocket (Glu97B, Asn98, Leu99, Ile174, and Glu217). Suffixes B and E in Tyr60B, Trp60E, Glu97B, etc., refers to the residual insertion loop residues in the thrombin chains.[52,53] As shown in Figure 2a, analysis of the best LISA models (R1 and R4) for the thrombin data set reveals four regions where steric similarity between the target and reference molecules is disfavored (depicted by the yellow spheres), two regions where steric similarity between them is favored (represented by green spheres), three regions where electrostatic similarity between them is disfavored (displayed by the red spheres), and a single region where electrostatic similarity between the target and reference molecules is favored (demonstrated by the blue sphere). The LISA spheres are in agreement with the corresponding CoMFA and CoMSIA contours, as revealed in Figure 2b and c, respectively. One yellow sphere is seen around the planar tetrazole moiety of molecule 39 (Figure 2a), which occupies the S1 pocket of the thrombin active site. Given that all the structures have been derived from X-ray crystal complexes, when the model is superimposed onto the thrombin active site (Figure 4), this yellow sphere is observed near the diminutive-sized S1 pocket, limiting the addition of bulk at this position. Thus, the constitution of the S1 pocket supports the disfavored steric similarity between the target and reference molecules around the tetrazole moiety. This is borne out by the fact that most of the active molecules possess a planar-protonated moiety, like guanidine, benzamidine, amidine, etc., in the vicinity of this yellow sphere. However, molecule 46 with a pyrazole ring and a bulky nonplanar piperidine ring in this region produces strong steric clashes with residues in the S1 pocket, leading to its very low activity. Similarly, the low activity of molecule 47 stems from the fact that its triazole ring with a $-SCH_3$ substituent creates steric hindrance with the groups in this region, and this is also indicated by the yellow sphere in the LISA analysis. Two yellow spheres are observed near the difluoromethyl group of molecule 39. In the active site, these spheres occupy a region near the S3 pocket (Figure 4) which does not permit any bulk to be added to the molecule due to probable steric clashes with the voluminous side chains of Ile174 and Glu217. One yellow sphere is seen near the meta position of the pyridyl ring of molecule 39. This area of the molecule falls between the S2 and S3 pockets, which is occupied by the bulky side chains of Tyr60B, Trp60E, Leu99, and Asn98 (Figure 4), that forbid any mass to be added to the molecule around this position. However, there is enough space between the S1 and S3 pockets of the active site (Figure 4) where two green spheres have been observed in the vicinity of the amide bond linking the piperidyl ring with the benzyl ring in molecule 39. Thus, bulky groups incorporated at these sites of the molecule can be well accommodated by the active site, which ensures tighter binding. A blue sphere near the NH group, linking the piperidyl ring with the difluoromethyl group in molecule 39, suggests that a electropositive substituent in this region is favored and could form hydrogen bonds with the backbone amide carbonyl of the nearby active-site residue Gly216 (Figure 4). Similarly, the three red spheres seen around the difluoromethyl group, the tetrazolyl moiety, and the ring-containing tetrazolyl group suggests substitutions with more electronegative groups at these positions that can form hydrogen bonds with the backbone atoms of the surrounding active-site residues Ile174, Cys191, and Ala190, respectively (Figure 4). Thus, the features of the LISA models are in agreement with that contained in the thrombin active site. In an exhaustive QSAR review of structurally diverse thrombin inhibitors, the authors concluded that steric features and collinear hydrophobic parameters govern the potency of the molecules.[51] A majority of the descriptors contained in the LISA models are predominantly steric parameters. This observation is, thus, in accordance with the literature report.

Rationalization of LISA Approach. As stated previously, a positive coefficient for a LISA descriptor in the QSAR equation indicates that similarity between the target and reference molecule should be preserved (or preferably enhanced) at that grid point. On the other hand, a negative coefficient recommends making the target molecule dissimilar to the reference molecule at that point, in order to increase its activity. According to the best LISA model R1 (Table 4), the activity of the target molecule can be augmented by increasing its steric similarity with the reference molecule at grid point 1227 and by reducing it at points 718 and 1131 as well as by enhancing its electrostatic similarity with the reference molecule at grid point 1077 and by reducing it at point 935. Similarly, the next best LISA model R4 (Table 4) suggests increasing the steric similarity between the target and reference molecules at grid point 1086 and by reducing it at points 1345 and 1819 as well as by decreasing the electrostatic similarity of the target with the reference molecule at grid points 754 and 1038 will increase its biological activity.

Table 8 shows the corresponding steric/electrostatic interaction energy values of the important LSI descriptors extracted by the best LISA models, for some selected molecules. The steric field is the van der Waals energy, and the electrostatic field is the Coulombic interaction energy between the probe and the molecule. It is to be noted that the more negative the value of the steric (van der Waals) and electrostatic (Coulombic) interaction energies, the stronger the interaction is between the ligand and the probe (receptor). Similarly, less negative or more positive values of these interaction energies imply weaker interaction between the respective groups of the ligand and the probe (receptor). The final goal is to improve the biological activity of the target molecule by manipulating the LISA descriptors extracted by the QSAR equations. This is accomplished by reducing the magnitude of those descriptors, which are negatively correlated with the activity (so as to change their "−" sign to "+"), while at the same time increasing the magnitude of those descriptors correlated positively with the

Table 8. Corresponding Steric/Electrostatic Interaction Energy Values of the Important LSI Descriptors Extracted by the Best LISA Models for the Thrombin Dataset, for Some Selected Molecules[a]

		Ste_718	Ste_1086	Ste_1227	Ste_1819	Ele_754	Ele_935	Ele_1038	Ele_1077
mol ID	pK_i	(−)[a]	(+)	(+)	(−)	(−)	(−)	(−)	(+)
39 (R1)	11.85	−0.024	−0.054	3.511	−0.032	0.221	0.405	1.028	0.883
31	10.05	−0.027	−0.054	2.939	−0.030	0.085	0.184	0.439	0.394
46	3.40	−0.053	−0.016	0.773	0.000	0.002	0.026	−0.041	0.050
49	8.00	−0.036	−0.055	−0.452	−0.090	1.070	1.217	2.790	1.868
55	6.00	−0.016	−0.026	−0.177	−0.015	0.055	−0.038	1.189	−0.050
60	4.00	−0.012	−0.029	−0.084	−0.042	−1.122	−1.226	−1.144	−1.960

[a] Signs within the parenthesis refer to the signs of the coefficients of the respective LSI descriptors in the QSAR equations.

Table 9. Possible Modifications Required to Increase the Activity of the Target Molecule under Various Situations

		sign of steric/electrostatic interaction energies		
situation no.	sign of LISA descriptor in QSAR equation	reference molecule	target molecule	required modification in target molecule to increase activity (or render it more active)
1	+	+	+	↑+/(+)
2	+	+	−	↓−/(+)
3	+	−	+	↓+/(−)
4	−	+	+	↓+/(−)
5	−	−	−	↓−/(+)
6	−	−	+	↑+/(+)
7	−	+	−	↑−/(−)
8	+	−	−	↑−/(−)

activity (so as to intensify their "positive" contributions). Equation 1 contains the product of interaction energies of the probe (steric/electrostatic) with the reference (P_{Ai}) and target (P_{Bi}) molecules in the numerator and the normalization function containing the squares of these values in the denominator. Since the magnitude/sign of the interaction energy of the reference molecule is fixed, the magnitude/sign of the LISA descriptors in the QSAR equations can be modulated only indirectly by manipulating the (sign of) interaction energy values of the target molecule so as to make it more or as potent as the reference molecule. This can be done in the manner described below.

Table 9 shows permissible modifications in the target molecule in terms of the signs of its (steric/electrostatic) interaction energy values. This enables design of more active compounds under various situations, depending on the signs of the coefficients of the LISA descriptors in the QSAR equations and on the signs of the interaction energy values of the reference molecule. Consider the first situation in Table 9 where the sign of the coefficient of a LISA descriptor in the QSAR equation and the steric or electrostatic interaction energy values of both the reference and target molecules also are positive; the activity of the target molecule can be increased by intensifying the magnitude of the interaction in the same direction. For example, if an ethyl group is positioned near grid point Ste_1227 (Table 8) in the reference molecule, then replacing/substituting a relatively less bulky group (e.g., methyl group) at this position in the target molecule will shift its steric interaction to a less negative or a more positive value by reducing its van der Waals interaction with the receptor. This less negative or more positive value, upon multiplication with the corresponding positive value of the reference molecule, would subsequently return a higher positive magnitude, thus, enhancing the biological activity. However, in a related scenario (situation four in Table 9), if the sign of the same LISA descriptor is negative in the QSAR equations, then replacing/substituting a relatively more bulkier group (like a propyl group) at the same position will render the steric interaction of the target molecule to a negative (or less positive) value by increasing its van der Waals interaction with the receptor. This negative (or less positive) value after multiplication with the corresponding positive value of the reference molecule will render a negative (or less positive) value, which in turn upon multiplication with a negative coefficient would result in an increased contribution to the activity. In this manner, the signs (and/or magnitude) of interaction energies of the target molecule can be modulated under all the situations mentioned in Table 9, that could result in improved potency.

In a similar manner, the steric/electrostatic interaction energies associated with the important LISA descriptors shown in Table 8 are related to the variation in the biological activities of the molecules. Molecule 39, having the highest pK_i value in the data set, was used as the reference molecule for constructing the best LISA model R1. All other molecules in Table 8 were compared with this reference molecule in terms of their steric/electrostatic interaction energy values.

According to the best LISA models, the steric similarity between the target and reference molecules is "favored" at the grid point 1086 due to its positive coefficient in the QSAR equations of the best LISA models (Table 4). Since the van der Waals interaction energy of the reference molecule is "negative" (−0.054, Table 8) at this grid point, strengthening the steric interaction (rendering it more "negative") at this grid point for the target molecule will result in its increased activity. But the van der Waals interaction energies of molecules 46, 55, and 60 are less "negative" (−0.016, −0.026, and −0.029, respectively, Table 8) than that of the reference molecule (−0.054, Table 8), thus, accounting for their reduced thrombin inhibitory activity. Similarly, the steric similarity of the target molecule with the reference molecule is also "favored" at grid point 1227

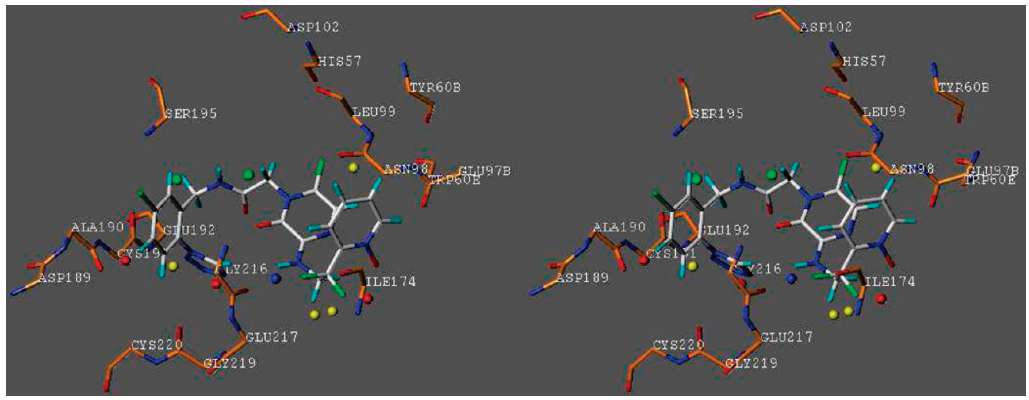

Figure 8. A stereoview of the active site of thrombin showing molecule 39 with important receptor residues (heavy atoms only).

because of its positive coefficient in the QSAR equations (Table 4). Thus, to have improved activity, the target molecule should have a "positive" value of the steric (van der Waals) interaction energy at grid point 1227, similar to the reference molecule which also has a "positive" steric interaction energy at this grid point (3.511, Table 8). Since at this grid point, the van der Waals interaction energy of all other molecules in Table 8 (31, 46, 49, 55, and 60) are less "positive" (or "more negative") compared to the reference molecule (3.511, Table 8), these molecules have relatively lower biological activities.

The best LISA models also indicate that the electrostatic (Coulombic) similarity between the target and reference molecules is "disfavored" at grid points 754, 935, and 1038, due to their negative coefficients in the QSAR equations (Table 4). The electrostatic (Coulombic) interaction energies of the reference molecule at these grid points have "positive" values (0.221, 0.405, and 1.028, respectively, Table 8), thus, the target molecule should have "negative" (or less "positive") values of the Coulombic interaction energies at these points to have improved activity. However, the Coulombic interaction energies of molecule 49 at grid points 754, 935, and 1038 are more "positive" (1.070, 1.217, and 2.790, respectively, Table 8) than that of the reference molecule (0.221, 0.405, and 1.028 respectively, Table 8); thus, causing a reduction in the activity of molecule 49. Molecule 55 also has lower biological activity due to its more "positive" (1.189, Table 8) Coulombic interaction energy at grid point 1038, compared to that of the reference molecule (1.028, Table 8).

Likewise, the relationship between the variation in biological activity of the molecules and interaction energies related to the other important LSI descriptors (shown in Table 8) can be rationalized on the basis of LISA models.

CONCLUSIONS

In conclusion, we report here a simple procedure for the formulation of accurate and easily interpretable QSARs. Molecular similarity has been explored by dissecting the overall molecular global similarity into local values, and these are used as descriptors in the QSAR formalism. For calculating the local similarity at a given point on the grid surrounding the molecule, the potential at that grid point for a molecule is compared to that of a reference (generally the most active) molecule in the data set. Since, the most active (reference) molecule usually exhibits optimal potentials on the surrounding grid, the similarity/dissimilarity of the potentials at grid points for all molecules in the data set can be used to explain the variation in the biological activity. Depending on the nature of the probe, different types of potentials (e.g., electrostatic, steric, lipophilic, etc.) can be used to calculate the local similarity index (LSI). The method can also be used to build a pseudoreceptor in the absence of any information on the binding protein. The local indices for similarity analysis (LISA) approach has been applied and validated on three large and diverse data sets—thrombin, glycogen phosphorylase *b*, and thermolysin inhibitors. The LISA models are comparable (or even better with respect to some parameters) to the models obtained by the standard comparative molecular field analysis (CoMFA) and comparative molecular similarity index analysis (CoMSIA) methods. Thus, LISA is a unique, alternative method based on a simple concept of local molecular similarity. The outcome of LISA can be graphically displayed, which gives insights into the ligand−protein binding mechanisms and provides clues for modification of the structure of the molecules to improve their potency.

ACKNOWLEDGMENT

The computational facilities used in this work were built through grants from the Department of Science and Technology (SR/FST/LSI-163/2003), the Council of Scientific and Industrial Research (01(1986)/05/EMR-II), and the All India Council of Technical Education, New Delhi. Jitender Verma and Alpeshkumar Malde thank CSIR, New Delhi for financial support.

Supporting Information Available: Structures of thrombin, GPB, and thermolysin inhibitors. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* **1962**, *194*, 178–180.
(2) Hansch, C.; Muir, R. M.; Fujita, T.; Maloney, P. P.; Geiger, F.; Streich, M. The Correlation of Biological Activity of Plant Growth Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients. *J. Am. Chem. Soc.* **1963**, *85*, 2817–2824.
(3) Hansch, C.; Fujita, T. ε-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626.
(4) Hansch, C. A. Quantitative Approach to Biochemical Structure-Activity Relationships. *Acc. Chem. Res.* **1969**, *2*, 232–239.

(5) Free, S. M. J.; Wilson, J. W. A Mathematical Contribution to Structure Activity Studies. *J. Med. Chem.* **1964**, *7*, 395–399.
(6) Dean, P. M. *Defining molecular similarity and complementarity for drug design. In Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Blackie Academic & Professional: Glasgow, UK, 1995; pp 1−23.
(7) Good, A. C. *3D Molecular similarity indices and their application in QSAR studies. In Molecular Similarity in Drug Design,* Dean, P. M., Ed.; Blackie Academic & Professional: Glasgow, UK, 1995; pp 24−56.
(8) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today* **2007**, *12*, 225–233.
(9) Maldonado, A. G.; Doucet, J. P.; Petitjean, M.; Fan, B. T. Molecular similarity and diversity in chemoinformatics: from theory to applications. *Mol. Diversity* **2006**, *10*, 39–79.
(10) Kubinyi, H. Molecular similarity. 1. Chemical structure and biological action. *Pharm. Unserer Zeit.* **1998**, *27*, 92–106.
(11) Kubinyi, H. Molecular similarity. 2. The structural basis of drug design. *Pharm. Unserer Zeit.* **1998**, *27*, 158–172.
(12) Barbosa, F.; Horvath, D. Molecular similarity and property similarity. *Curr. Top. Med. Chem.* **2004**, *4*, 589–600.
(13) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.
(14) Willett, P. Chemoinformatics - similarity and diversity in chemical libraries. *Curr. Opin. Biotechnol.* **2000**, *11*, 85–88.
(15) Hodgkin, E. E.; Richards, W. G. Molecular similarity based on electrostatic potential and electric field. *Int. J. Quantum Chem.* **1987**, *14*, 105–110.
(16) Kubinyi, H.; Hamprecht, F. A.; Mietzner, T. Three-Dimensional Quantitative Similarity−Activity Relationships (3D-QSiAR) from SEAL Similarity Matrices. *J. Med. Chem.* **1998**, *41*, 2553–2564.
(17) Gallegos, A.; Robert, D.; Gironés, X.; Carbó-Dorca, R. Structure-toxicity relationships of polycyclic aromatic hydrocarbons using molecular quantum similarity. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 67–80.
(18) Carbó-Dorca, R.; Leyda, L.; Arnau, M. How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem.* **1980**, *17*, 1185–1189.
(19) Cramer III, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). I. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
(20) 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.
(21) Tokarski, J. S.; Hopfinger, A. J. Three-dimensional molecular shape analysis-quantitative structure-activity relationship of a series of cholecystokinin-A receptor antagonists. *J. Med. Chem.* **1994**, *37*, 3639–3654.
(22) Silverman, B. D.; Platt, D. E. Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. *J. Med. Chem.* **1996**, *39*, 2129–2140.
(23) Carbo, R.; Leyda, L.; Arnau, M. How Similar is a Molecule to Another? An Electron Density Measure of Similarity Between two Molecular Structures. *Int. J. Quantum Chem.* **1980**, *17*, 1185–1189.
(24) Duca, J. S.; Hopfinger, A. J. Estimation of Molecular Similarity Based on 4D-QSAR Analysis: Formalism and Validation. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1367–1387.
(25) Datar, P. A.; Khedkar, S. A.; Malde, A. K.; Coutinho, E. C. Comparative residue interaction analysis (CoRIA): a 3D-QSAR approach to explore the binding contributions of active site residues with ligands. *J. Comput. Aided Mol. Des.* **2006**, *20*, 343–360.
(26) Verma, J.; Khedkar, V. M.; Prabhu, A. S.; Khedkar, S. A.; Malde, A. K.; Coutinho, E. C. A comprehensive analysis of the thermodynamic events involved in ligand-receptor binding using CoRIA and its variants. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 91–104.
(27) Dhaked, D. K.; Verma, J.; Saran, A.; Coutinho, E. C. Exploring the binding of HIV-1 integrase inhibitors by comparative residue interaction analysis (CoRIA). *J. Mol. Model.* **2009**, *15*, 233–245.
(28) Khedkar, S. A.; Malde, A. K.; Coutinho, E. C. Design of Inhibitors of the MurF Enzyme of Streptococcus pneumoniae Using Docking, 3D-QSAR, and de Novo Design. *J. Chem. Inf. Model.* **2007**, *47*, 1839–1846.
(29) Pissurlenkar, R. R. S.; Coutinho, E. C. HomoSAR: An Integrated Approach Using Homology Modeling and Quantitative Structure-Activity Relationship for Activity Prediction of Peptides. *Scholarly Research Exchange* **2008**, 2008.
(30) Cocchi, M.; Benedetti, P. G. D. Use of the Supermolecule Approach to Derive Molecular Similarity Descriptors for QSAR Analysis. *J. Mol. Model.* **1998**, *4*, 113–131.
(31) Petke, J. D. Cumulative and discrete similarity analysis of electrostatic potentials and fields. *J. Comput. Chem.* **1993**, *14*, 928–933.
(32) Good, A. C. The calculation of molecular similarity: alternative formulas, data manipulation and graphical display. *J. Mol. Graph.* **1992**, *10*, 144–151.
(33) FitzGerald, G. A. The human pharmacology of thrombin inhibition. *Coron. Artery Dis.* **1996**, *7*, 911–918.
(34) Chu, A. J. Biochemical strategies to anticoagulation: a comparative overview. *Curr. Vasc. Pharmacol.* **2004**, *2*, 199–228.
(35) Becker, R. C. Novel constructs for thrombin inhibition. *Am. Heart J.* **2005**, *149*, S61–72.
(36) Haas, S.; Schellong, S. New anticoagulants: from bench to bedside. *Hamostaseologie* **2007**, *27*, 41–47.
(37) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(38) *Sybyl*, version 7.1; Tripos Associates Inc.: St. Louis, MO, 2005.
(39) Halgren, T. A. Merck Molecular Force Field V. Extension of MMFF94 using Experimental Data, Additional Computational Data and Empirical Rules. *J. Comput. Chem.* **1996**, *17*, 616–641.
(40) *Cerius2*, version 4.8; Accelrys Inc.: San Diego, CA, 1998.
(41) Oikonomakos, N. G. Glycogen phosphorylase as a molecular target for type 2 diabetes therapy. *Curr. Protein Pept. Sci.* **2002**, *3*, 561–586.
(42) Greenberg, C. C.; Jurczak, M. J.; Danos, A. M.; Brady, M. J. Glycogen branches out: new perspectives on the role of glycogen metabolism in the integration of metabolic pathways. *Am. J. Physiol. Endocrinol. Metab.* **2006**, *291*, E1–8.
(43) Gohlke, H.; Klebe, G. DrugScore meets CoMFA: adaptation of fields for molecular comparison (AFMoC) or how to tailor knowledge-based pair-potentials to a particular protein. *J. Med. Chem.* **2002**, *45*, 4153–4170.
(44) Gerber, P. R.; Muller, K. MAB. a generally applicable molecular force field for structure modelling in medicinal chemistry. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 251–268.
(45) Adekoya, O. A.; Sylte, I. The thermolysin family (M4) of enzymes: therapeutic and biotechnological potential. *Chem. Biol. Drug Des.* **2009**, *73*, 7–16.
(46) Reddy, A. V. Thermolysin: a peptide forming enzyme. *Indian J. Biochem. Biophys.* **1991**, *28*, 10–15.
(47) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.
(48) Wold, S.; Johansson, E.; Cocchi, M. *PLS: Partial Least Squares Projections to Latent Structures. In 3D QSAR in Drug Design: Theory, Methods and Applications,* Kubinyi, H., Ed.; ESCOM Science Publishers: Leiden, The Netherlands, 1993; pp 523−550.
(49) Richard, D.; Cramer, R. D., III; Bunce, J. D.; Patterson, D. E.; Frank, I. E. Crossvalidation, Bootstrapping, and Partial Least Squares Compared with Multiple Regression in Conventional QSAR Studies. *Quant. Struct.-Act. Relat.* **1988**, *7*, 18–25.
(50) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A comparison of methods for modeling quantitative structure-activity relationships. *J. Med. Chem.* **2004**, *47*, 5541–5554.
(51) Kontogiorgis, C. A.; Hadjipavlou-Litina, D. Quantitative Structure - Activity Relationships (QSARs) of Thrombin Inhibitors: Review, Evaluation and Comparative Analysis. *Curr. Med. Chem.* **2003**, *10*, 525–577.
(52) 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.
(53) Bode, W.; Turk, D.; Karshikov, A. The refined 1.9-Å X-ray crystal structure of D-Phe-Pro-Arg chloromethylketone-inhibited human a-thrombin: Structure analysis, overall structure, electrostatic properties, detailed active-site geometry, and structure-function relationships. *Protein Sci.* **1992**, *1*, 426–471.
(54) Vedani, A.; Dobler, M.; Lill, M. A. Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. *J. Med. Chem.* **2005**, *48*, 3700–3703.

CI900224U