Evaluation of a Novel Shape-Based Computational Filter for Lead Evolution: Application to Thrombin Inhibitors

Jayashree Srinivasan,* Angelo Castellino, Erin K. Bradley, John E. Eksterowicz, Peter D. J. Grootenhuis, Santosh Putta, and Robert V. Stanton

Deltagen Research Laboratories, 4570 Executive Drive, Suite 400, San Diego, California 92121

Received October 23, 2001

A novel shape−feature-based computational method is described and used to rapidly filter compound libraries. The computational model, built using three-dimensional conformations of active and inactive molecules, consists of a collection of whole molecule shapes and chemical feature positions that are ranked according to their correlation with activity. A small ensemble of these shapes and features is used to filter virtual compound libraries. The method is applied to two thrombin data sets and is shown to be efficient in identifying novel scaffolds with enhanced hit rates.

Introduction

Lead Evolution: A Definition. Analysis of (pre)-clinical drug development candidates has shown that problems with pharmacokinetics and safety are responsible for nearly 50% of all failures.[1] To some extent, these problems are coupled to the chemical class of the candidate. To increase the chances for clinical success and to navigate around chemical space patented by competitors, developing backup compounds based on chemically diverse series is a necessary step in the drug discovery pathway. This process of converting knowledge of one active chemical class into the discovery of a new class of compounds, while maintaining the desired activity, is what we refer to as lead evolution.

Previous Lead Evolution Strategies. Screening large chemically diverse corporate or combinatorial libraries to identify new chemical classes is a widely used strategy. This approach, typically carried out at an early stage of a project, has proven to be successful but costly. A more cost-effective means to identifying a new chemical class would involve the integration of computational and experimental methods to use information about the biological target and its known ligands. Previous computational methods that have been successfully applied toward lead evolution are discussed below.

(A) Methods Based on Receptor or Ligand Structure. In cases where a high-resolution three-dimensional (3D) structure of the biological target is available, 3D database docking and structure-based design have been applied for lead evolution.[2−4] An obvious limitation to this approach is that structures of a majority of pharmacologically interesting receptors are unknown. In cases where the receptor structure is unknown, structural features that contribute to biological activity must be inferred from the ligands. To that effect, ligand-based pharmacophore methods have been applied in conjunction with database searching to demonstrate lead evolution.[5−8]

(B) Methods Based on Similarity Searching. The hypothesis for similarity searching in chemical space is the similarity property principle, which states that compounds that are chemically alike in some way will have similar biological activities.[9] Therefore, given a target molecule, one that has been shown to exhibit some biological activity, a similarity search of a chemical database aids the procedure of finding a new lead. Searching chemical databases using this method can be very fast and therefore used to quickly scan through large collections of molecules. However, these methods typically rely on two-dimensional (2D) chemical connectivity or other 2D descriptors such as substructural fragment information. The emphasis of these methods on bond connectivity can be a limitation for lead evolution as illustrated by the study of Zheng and coworkers[10] who applied a Focus-2D method to search a combinatorial N-substituted glycine virtual library. Some successes have been reported with 3D database searches using a single 3D structure of the ligand as a query.[11−18]

Shape-Based Method: A New Approach. In this paper, we describe a novel ligand-based computational method for rapid lead evolution that is independent of chemical class and is an effective discriminator of activity. It involves the generation of ensembles of 3D shape-based descriptors. A collection, referred to here as the ensemble model, of these shape-based descriptors is constructed from a training set of compounds of known activity. This ensemble is used to filter source pools of compounds in new chemical classes resulting in compound sets enriched in activity against the target. To further elucidate the robustness of this method, its ability to identify novel active scaffolds relative to a commonly used 2D similarity searching technique is presented.

Several implementations of molecular shape-based descriptors, as well as excluded volume models, have been presented in the literature.[19−22] In general, these approaches involve either creating a "shrink wrap" around a pharmacophore description of a conformer to a molecule[21] or using an ensemble of 3D shapes based on pseudo-receptors[22] or comparing shapes generated from fragments within a molecule.[20] As compared to the published shape-based studies, the new method uses an

* To whom correspondence should be addressed. Tel: (858)625-6409. Fax: (858)625-9293. E-mail: jsrinivasan@deltagen.com.

10.1021/jm010494q CCC: $22.00
Published on Web 05/07/2002

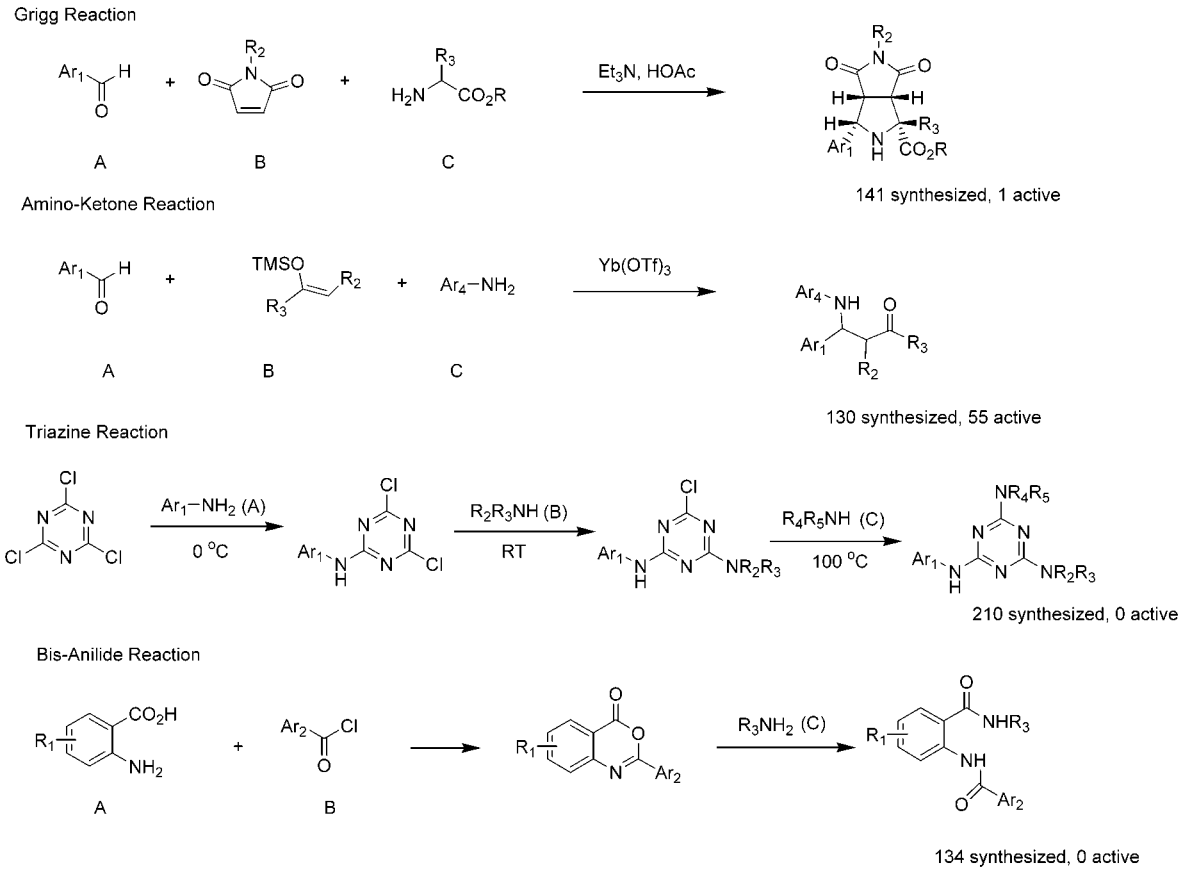

Figure 1. Reaction sequences for combinatorial library syntheses. The letter identifiers refer to the components described in the General Synthetic Procedures section. The number of compounds synthesized and the number of actives are indicated below each scaffold.

all atom representation for each of the training set compound conformers. Like the published methods, the current approach is demonstrated in the context of database searching. However, the speed of this method allows for searching large virtual libraries holding hundreds of thousands of compounds.

Research Strategy

Retrospective Analysis. One of the main advantages of the present shape-based method is that it can be applied to problems where little prior information about the target exists. To demonstrate the effectiveness and advantages of the new technique, it is necessary to apply it to problems that are widely known and that are rich in data. This proof-of-concept study on known data is referred to as retrospective analysis.

The present study describes two retrospective analyses of the shape-based technique on a well-characterized problem: the search for novel thrombin inhibitors. A shape-based ensemble model is built using data gleaned from literature. The model is applied first to a data-mining problem on a subset of the MDDR collection[23] and second to a more challenging and realistic (though still retrospective) problem of searching an in-house library that was synthesized in the course of internal research on thrombin inhibitors.

Thrombin Data Set: Literature. Thrombin inhibitors have a therapeutic potential for the treatment and prevention of thrombosis-related disorders.[24] The structure and activity of several potent thrombin inhibitors are published and provide a basis for a good benchmark data set.[25,26] An ensemble model was generated using a training set active and inactive molecules where the active molecules were taken from literature.

Thrombin Data Set: Synthetic Library. A library of 634 compounds was synthesized in the course of an internal research for thrombin inhibitors. The compounds synthesized were based on five different chemistries of which four are shown in Figure 1.[27] The chemistries were chosen based on the following criteria: (i) a dissimilar display of the key positively charged feature within the reaction products, (ii) the scope of the reaction as defined from synthetic development knowledge,[28] (iii) the commercial availability of the reactants as given in the ACD,[29] (iv) the ability to isolate products from their crude reaction mixtures despite the variability in reaction yields, and (v) the ability to achieve rapid synthetic access. To address the last criterion, these chemistries were either multicomponent (Grigg and amino−ketone) or multistep, one pot (Bis-anilide and triazine) reaction sequences.

For each reaction sequence represented in Figure 1, one of the reaction components contained the positively charged key feature. Specifically, component A of the triazine reaction and component C of the amino−ketone reaction contain 3- and 4-aminobenzamidine. Component C of the Grigg reaction contains the aminobenzamidines and guanidine-containing amines, and component C of the Bis-anilide reaction consists only of amines with a guanidine substituent.

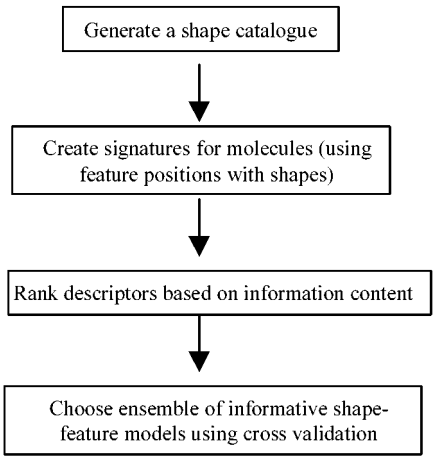

Figure 2. Flowchart of the computational model building process.

Computational Methods

Model Generation and Shape−Feature Method. The computational model consists of a collection of shape−feature descriptors. Each descriptor combines a 3D shape with a single chemical feature type located at a particular position within the shape. The chemical features selected reflect important ligand−receptor interactions and include hydrogen bond donors and acceptors, positive and negative charges, aromatic rings, and hydrophobes. A concise description of the method is given below while the algorithmic details are described elsewhere.[30]

Generation and cross-validation of the computational model is a four step process and is schematically depicted in Figure 2. First, a conformational model is generated for all active ligands using the in-house program CONAN.[31−33] For the active molecules, each conformer is placed onto a 3D grid and aligned with a user-defined key feature in combination with the principle moments. In this study, the defined key feature is a positive charge present in each of the thrombin inhibitors. The grid points that fall within the van der Waals surface of the molecule in a particular conformation define the molecular shape. Henceforth, the word shape refers to this key feature-centered molecular shape. Closely related shapes are eliminated during this process based on a user-defined volume Tanimoto index for comparison. The collection of all resulting 3D shapes is called the shape catalog. The combination of a shape, a feature type, and a feature position is defined as a shape−feature descriptor, and the combinations of all feature types, in all grid positions, of all shapes in the shape catalog are defined as the descriptor space.

The second step is to create a signature for every molecule in the training set. This is done by generating a conformational model for the molecule and then representing the presence or absence of all the descriptors that can be matched by a molecule's conformers. This binary description of all conformations of a molecule in terms of the descriptor space is defined as its signature.

In the third step, a subset of these descriptors is selected that is best able to differentiate between active and inactive molecules. This is done by ranking each shape−feature descriptor by its calculated information content.[8,34,35] The descriptors with the highest information content are retained and used together as an ensemble model. Finally, in the fourth step, a cross-validation is carried out in order to characterize the predictive value of the model.

Library Filtering. The ensemble model is developed based on information from the biological activity of the training set molecules. Because the ensemble model no longer retains chemical connectivity information, it can be used to search for new chemical classes with similar biological activity as the training set. To that end, virtual libraries are filtered by first generating the signature of the compounds using the previously defined shape catalog.[36] Next, the virtual compounds are given a score based on the number of shape−feature descriptors that are matched in the ensemble model. Molecules from the virtual libraries whose scores are greater than a defined threshold are proposed to have activity similar to the training set.

Comparison to Similarity Search. Two-dimensional similarity searches were carried out, using the training set actives, to compare with the performance of the shape−feature method. Queries were calculated for the actives with MACCS keys[37] as implemented within MOE.[38] The query set consisted of the most active compound and the four remaining most structurally diverse active molecules. All source pool compounds were then ranked based on similarity to each of these five query molecules. An equal number of molecules were retained for each query molecule. This procedure was repeated three times, each time starting with a different active molecule from the training data set and the four most structurally diverse molecules (from this new starting molecule) as the remainder of the five query molecules.

Results and Discussion

Ensemble Model Building. A thrombin data set used for building the ensemble model consisted of 38 thrombin inhibitors taken from the literature[25,26,39] and 2418 chemically diverse inactive compounds. The average Daylight[40] pairwise Tanimoto similarity of the 38 literature actives is calculated to be 0.32. These Daylight fingerprints are created with the default parameters. These default parameters allow for a maximum (creation) fingerprint size of 2048 bits and repeatedly fold the fingerprint to a minimum size of 64 bits. The calculated similarity of 0.32 suggests that, as a collection, the active molecules are topologically different from each other. The measured activities of these inhibitors range in their K_i values from micromolar to subnanomolar. The set of 2418 inactives came from a chemically diverse internal screening library.

A shape catalog containing 327 shapes was created from conformers of the active compounds. The shape catalog was generated using a grid with 31.5 Å × 24 Å × 18 Å dimensions, a grid spacing of 1.5 Å, and a positive charge as the key feature. Signatures were then generated for each of the training set compounds using the shape catalog and the defined features (acceptors, donors, positive and negative charge, aromatic rings, and hydrophobes). The combination of shapes, features, and grid size defines the descriptor space and yields a signature of ∼7.9 million bits.

An ensemble of the top 500 shape−feature descriptors (containing 50 unique shapes), as ranked by information

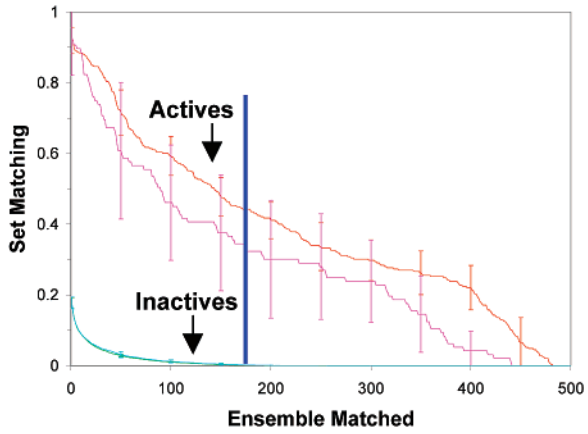

Figure 3. Cross validation and ensemble performance plot. The initial set of thrombin data is split into active and inactive subsets. Each subset is further split with 75% of the compounds used for training and 25% reserved as the test set. The active training set is represented with a red line, and the active test set is represented with a magenta line. The inactive training set is represented in green with the inactive test set in cyan (note that the two lines coincide and appear as one). The performance of the top 500 models (individual shape−feature descriptors) is evaluated in the following way. The *x*-axis indicates the minimum number of models matched by a given set of compounds. The *y*-axis then shows the fraction of compounds for a given set that match that threshold. For example, approximately 45% of the test set actives and 2% of the test set inactives match at least 100 of the top 500 models. For this study, the cutoff was set at 175 models, as indicated by the vertical blue line. Applying this cutoff to the database searches, a compound that matches any 175 out of the 500 models in the ensemble is considered to have passed the selection filter. Ten trials of cross validation experiments were done, and the error bars are shown. For each experiment, the training (75%) and test (25%) sets were chosen randomly.

content, was selected using this training set. The cross-validated ensemble performance plot is shown in Figure 3. From this plot, a threshold for the computational filter was set to 175 based on retaining the maximum number of active molecules and the minimum number of inactive molecules. Compounds that matched any 175 or more descriptors in the 500 shape−feature ensemble model were considered to have passed the selection filter. This threshold filter was applied when selecting compounds from the two libraries discussed below.

MDDR Database Filtering. To evaluate the performance of our ensemble model as a database filtering tool, we used it to filter the MDDR database,[23] which holds over 100 000 druglike compounds. Because our model requires positively charged molecules, it was necessary to prefilter the MDDR for molecules having at least one positively charged group. Applying this prefilter yielded a set of 35 462 molecules, and this subset of the MDDR is defined as the source pool for the search. Within this subset, matching the key words "thrombin inhibitor" identified thrombin actives. After duplicates from the training set were removed, 540 actives were found in this set. The average pairwise Tanimoto similarity among these actives was 0.33 as calculated from their Daylight fingerprints. Additionally, the average pairwise similarity of the MDDR actives to the literature actives was calculated to be 0.31. On the basis of these and additional similarity comparisons (see Table 1), we infer that the two sets (literature and MDDR actives) differ substantially.

Table 1. Average Pairwise Tanimoto Similarity Comparisons of Active Molecules in the MDDR to Those in the Literature Set

MDDR actives	percent with similarity < 0.75[a]	percent with similarity < 0.50[a]
source pool[b]	82	33
shape−feature filter[c]	80	21
MACSS keys similarity filter[d]	76	0

[a] Similarity comparison to all 38 literature active molecules. [b] The MDDR source pool contains 540 actives. [c] A total of 181 actives were selected by the shape−feature filtering. [d] A total of 67 actives were filtered by the MACCS key similarity search method.

When the shape−feature ensemble filter was applied, 507 compounds out of the 35 462 molecules were selected. Of the 507 compounds that passed the selection criterion, 181 were thrombin actives (33%) and 326 were not listed as thrombin active molecules. This resulted in an enrichment ratio (defined as a footnote in Table 2) of 23 (see Table 2).

The MACCS keys-based 2D similarity search was applied to the same source pool by retaining the top 100 molecules for each query molecule, for a total of 500 molecules (see Table 2). This method selected 67 out of the 540 thrombin actives (12% of the actives). The average enrichment was 8.6, which is 2.7 times smaller than that obtained from the shape−feature method.

Synthetic Library Filtering. The shape−feature ensemble model was evaluated as a tool for lead evolution by searching the set of compounds from a synthesized collection around five separate templates (see Figure 1). Overall, this synthetic library consisted of 634 compounds. Sixty-four of those molecules, on three of the five templates, were identified as active thrombin inhibitors (where 50% inhibition or greater at 25 μM inhibitor concentration was considered active).

Filtering the synthetic library using the shape−feature ensemble model resulted in the selection of 109 compounds out of the 634 total (Table 2). This selection process recovered 15 out of the 64 actives (23%) resulting in a library enrichment ratio of 1.4. Two of the three possible active templates, compounds within the amino−ketone and Grigg templates, were represented among the set of molecules that passed the filter, and an example from each template is represented in Figure 4. Note that the average pairwise Tanimoto similarity of Daylight fingerprints between the selected active compounds from the synthetic library and the training set actives was only 0.24. It is very encouraging, however, that the shape−feature method is predictive even when the literature set and source pool actives are topologically very dissimilar. Of particular interest is the fact that the source pool of synthetic actives uniformly displays weak thrombin inhibition, with IC_{50} values in the micromolar range, unlike that of the MDDR actives, which display potent thrombin inhibition.

Results from the search using the MACCS key similarity method retaining the top 20 molecules for each query molecule, for a total of 100 molecules, are also summarized in Table 2. An enrichment ratio of 0.6 was obtained by the similarity method as compared to 1.4 for the shape−feature method. Considering that a random selection should on average result in an enrich-

Table 2. Summary of Database (MDDR and Combinatorial Library) Searches

	MDDR		combinatorial library	
	shape−feature filtering	MACCS key similarity searching	shape−feature filtering	MACCS key similarity searching
no. of molecules selected	507	500	109	101
no. of actives selected	181	67	15	6
percent of actives selected	33	12	23	9
enrichment ratio[a]	23	8.6	1.4	0.6
no. of active scaffolds selected[b]			2	1

[a] The enrichment ratio is calculated as the ratio between the fraction of actives in the selected set and the fraction of actives in the source pool. For example, in the MDDR set, the fraction of actives in the source pool is (540/35 462) = 0.0152. While the fraction of actives in the selected set for the shape−feature method is (181/507) = 0.357. On the basis of the above definition, the enrichment ratio for the shape−feature method is (0.357/0.0152) = 23. [b] The combinatorial library consists of five different scaffolds. The number of scaffolds represented in the selected set of active molecules is indicated. The number of scaffolds was not determined for the MDDR searches.

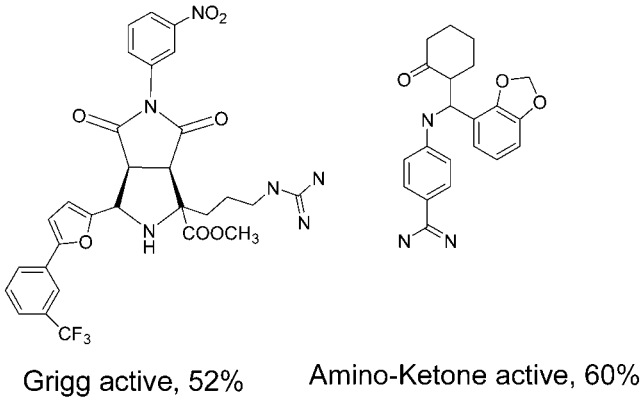

Figure 4. Examples of synthetic actives (human thrombin percent inhibition at 25 μM inhibitor concentration).

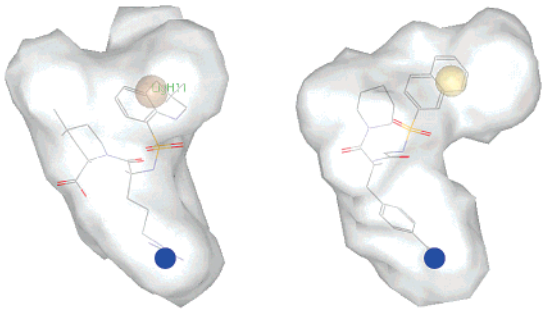

Figure 5. Fit of the cocrystal ligand conformations of NAPAP (left) and argatroban (right) in two of the 500 shape−feature models used in the ensemble. The key feature (positive charge) is depicted as a blue sphere for both structures. A hydrophobic feature is shown as the brown sphere for NAPAP while the aromatic centroid feature is shown as the gold sphere for argatroban.

ment of 1.0, this study emphasizes the difficulties of using a topological descriptor to search novel chemical space.

Comparison of the Shape Catalog with Inhibitor Cocrystal Structures. We investigated the relationship between shapes in the shape catalog and the shapes derived from X-ray ligands. This comparison is interesting since no knowledge of the ligand cocrystal structure was included in the construction of the shape catalog. In Figure 5, the cocrystal ligand structures for NAPAP and argatroban, crystallized with thrombin,[41] are displayed within two informative shapes from the ensemble model. The two shapes were picked after visual inspection of the top-ranking descriptors and do not necessarily present the most informative ones. Overall, the cocrystal conformer shape and the two informative shapes are in qualitative agreement. The chemical features are also well-placed to pick up the hydrophobic feature in NAPAP and aromatic ring for argatroban.

Conclusions

In this paper, we have demonstrated the use of a novel shape−feature-based searching method, which because of its abstraction from chemical connectivity and 3D nature is useful for lead evolution. Searches of the MDDR and a synthetic library show that the method can detect thrombin actives with reasonable enrichment even when the training set is chemically dissimilar to the library that is searched. When compared with the performance of a more conventional 2D similarity searching method, the shape−feature-based method performs better in terms of enrichment for activity and ability to select novel scaffolds. Interestingly, there seems to be a qualitative agreement between some of the shapes generated and the shapes of experimentally observed inhibitors in complex with thrombin.

Experimental Section

Abbreviations used as follows: DME, ethylene glycol dimethyl ether; MeCN, acetonitrile; DIEA, *N*,*N*-di-isopropylethylamine; DMF, *N*,*N*-dimethylformamide; DMSO, dimethyl sulfoxide; NAPAP, *N*-α-((2-naphthylsulfinyl)glycyl)-DL-*p*-amidinophenylalanylpiperidine.

Combinatorial Synthesis. Reactions were conducted in a polystyrene 96 well reactor with delivery of components and accessory reagents by a Tecan liquid-handling system. Technology for execution of the solution phase chemistries has been described in a preliminary account of the syntheses of these libraries.[42] After the reaction sequence was completed, volatiles were removed with a Genevac HT-12 and the residues were dissolved into 200 μL of DMSO. Purification was then effected on a YMC-Pack ODS-A column with acetonitrile−water gradients buffered with 0.03% trifluoroacetic acid (TFA). Collection was effected by mass-triggering[43] using a PE-SIEX API150EX single quadrupole mass spectrometer with a Gilson 204 fraction collector into microtiter plates to maintain a collection-well to synthesis-well correspondence. Orthogonal detection allowed for "on-the-fly" quantification of the collected material after construction of the appropriate calibration curves.[44]

After the solvent was removed, the residues were diluted to 10 mM with DMSO with a Tecan RSP 150/8 Genesis liquid handler to provide master plates that were reformatted into daughter plates for biological screening. Database integration of synthetic and analytical data allowed for sample tracking and volume calculations for all liquid-handling steps.[45]

Grigg Reaction.[46] To a mixture of 40 μL of 0.4 M aldehyde (component A), 40 μL of 0.4 M *N*-substituted maleimide (component B), and 160 μL of 0.1 M amino-ester (component C), 16 μL of 2 M triethylamine followed by 20 μL of 0.4 M acetic acid was added. All components and accessory reagent solu-

tions were prepared in DMF with the exception of some amino-esters where ethanol was the more appropriate solvent for dissolution. The resulting mixture was agitated for 5 h at 60 °C. The crude reaction mixtures were then subjected to high-performance liquid chromatography mass spectrometry (HPLC-MS) purification.

Amino−Ketone Reaction.[47] To a mixture of 20 μL of 1.0 M aldehyde (component A) and 24 μL of 1.0 M silyloxyether (component B), 20 μL of an approximately 1 M solution of the amino-benzamidine (component C) was added. Component solutions were prepared in acetonitrile. To the component mixture, 20 μL of 0.3 M $Yb(OTf)_3$ in MeCN was added. After the mixture was agitated for 16 h at room temperature, the crude reaction mixture was subjected to HPLC-MS purification.

Triazine Reaction.[48] For execution of a single-pot sequence, inputs were segregated in order of nucleophilicity. Thus, component A contains the anilines, 3- and 4-aminobenzamidines, component B contains both primary and secondary amines while component C contains only the more nucleophilic secondary amines.[28] Prior to HPLC-MS purification, the reconstituted reaction residues were centrifuged to allow for injection of the supernatants.

Bis-Anilide Reaction.[49] For execution of a single-pot sequence, the 2-substituted-4*H*-3,1-benzoxazin-4-one intermediates were formed by reacting the anthranillic acids with an excess of the acid chlorides. Subsequent addition of the suitably protected guanidine-containing amine serves to quench the excess acid chloride and to ring open the benzoxazinone intermediates to give the final products.[28] Prior to HPLC-MS purification, random wells were checked to ensure deprotection was complete.

Screening Protocol. A fluorescence-based assay was performed by Chromagen, Inc., San Diego, CA, based on competition of a high-affinity Glu-Pro-Arg peptide substrate and the synthetic compounds to the thrombin active site. The assay was carried out in a 96 well plate format with each column containing a synthetic compound. Internal standard for the assay involved the use of NAPAP in the place of the synthetic compound. The control for every plate contained a 10% DMSO stock solution placed in the first two rows and columns of the plate. To the prepared wells containing 10 μL of the synthetic compounds, 95 μL of the buffer solution containing the peptide substrate was added and the fluorescence at 405 nm was measured. To these equilibrated wells, buffered human α-thrombin solution was added. After a short equilibration, the fluorescence of each well was measured every 5 min for 20 min in order to follow the enzyme kinetics. The percent inhibition of the synthetic compound was calculated from the measured fluorescence after correcting for the measurements made from the control wells.

Acknowledgment. We thank Klaus Gubernator, Xu Bai, and Xin Gu for their contributions on the synthesis of the in-house library, Leslie Robinson for productive conversations regarding the use of a key-feature with the shape-based method, the molecular design group for support, and the original programmers of the software used here, Jonathan Greene and Paul Beroza.

References

(1) Kennedy, T. Managing the drug discovery process. *Drug Discovery Today* **1997**, *2*, 436−444.
(2) DesJarlais, R. L.; Sheridan, R. P.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. Docking flexible ligands to macromolecular receptors by molecular shape. *J. Med. Chem.* **1986**, *29*, 2149.
(3) Leach, A. R. Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.* **1994**, *235*, 345−356.
(4) Makino, S.; Ewing, T. J.; Kuntz, I. D. DREAM++: flexible docking program for virtual combinatorial libraries. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 513−532.
(5) Van Drie, J. H.; Nugent, R. A. Addressing the challenges of combinatorial chemistry: 3D databases, pharmacophore recognition and beyond. *SAR QSAR Environ. Res.* **1998**, *9*, 1−21.
(6) Mason, J. S.; Morize, I.; Menard, I. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: Overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251−3264.
(7) Davies, K.; Briant, C. Combinatorial chemistry library design using pharmacophore diversity. *Network Sci.* **1996**, *1*, http://www.netsci.org/Science/Combichem/feature05.html.
(8) Bradley, E. K.; Beroza, P.; Penzotti, J. E.; Grootenhuis, P. D. J.; Spellmeyer, D. C.; Miller, J. L. A rapid computational method for lead evolution: description and application to α1-adrenergic antagonists. *J. Med. Chem.* **2000**, *43*, 2770−2774.
(9) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley: New York, 1990.
(10) Zheng, W. F.; Cho, S. J.; Tropsha, A. Rational combinatorial library design. 1 Focus-2D: A new approach to the design of targeted combinatorial chemical libraries. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 251−258.
(11) Beavers, M. P. Discovery of a small molecule motilin antagonist through electronic database screening. Innovative Computational Applications, San Francisco, CA, Oct 25−26, 1999.
(12) Bures, M. G. Recent Techniques and Applications in Pharmacophore Mapping. In *Practical Application of Computer-Aided Design*; Charifson, P. S., Ed.; Marcel-Dekker: New York, 1997; pp 39−72.
(13) Van Drie, J. H. Strategies for the determination of pharmacophoric 3D database queries. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 39.
(14) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity profiling and design using 3D pharmacophores: Pharmacophore-derived queries (PDQ). *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214−1223.
(15) Finn, P. W.; Mason, J. S. Computer-based screening of compound databases for the identification of new leads. *Drug Discovery Today* **1996**, *1*, 363−370.
(16) Ashton, M. J.; Jaye, M. C.; Mason, J. S. New perspectives in lead generation II: Evaluating molecular diversity. *Drug Discovery Today* **1996**, *1*, 71−78.
(17) Lam, P. Y. S.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; L. T., B.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N.; Chang, C.; Weber, P. C.; Jackson, D. A.; Sharpe, T. R.; Erickson-Viitanen, S. Rational design of potent, bioavailable, nonpeptide cyclic ureeas as HIV protease inhibitors. *Science* **1994**, *263*, 380−384.
(18) Martin, Y. C. 3D database searching in drug design. *J. Med. Chem.* **1992**, *35*, 2145−2154.
(19) Kenan, D. J.; Tsai, D. E.; Keene, J. D. Exploring molecular diversity with combinatorial shape libraries. *Trends Biochem. Sci.* **1994**, *19*, 57−64.
(20) Cramer, R. D.; Poss, M. A.; Hermsmeier, M. A.; Caulfield, T. J.; Kowala, M. C.; Valentine, M. T. Prospective identification of biologically active structures by topomer shape similarity searching. *J. Med. Chem.* **1999**, *42*, 3919−3933.
(21) Van Drie, J. H. "Shrink-wrap" surfaces: A new method for incorporating shape into pharmacophoric 3D database searching. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 38.
(22) Walters, E. D.; Hinds, M. R. Genetically evolved receptor models: A computational approach to construction of receptor models. *J. Med. Chem.* **1994**, *37*, 2527−2536.
(23) *MDDR*; Molecular Design Ltd.: San Leandro, CA, 2000.
(24) Berliner, L. J. *Thrombin Structure and Function*; Plenum Press: New York, 1992.
(25) Vacca, J. P. Thrombosis and Coagulation. *Ann. Rep. Med. Chem.* **1998**, *33*, 81−90.
(26) Malley, M. F.; Tabernero, L.; Chang, C. Y.; Ohringer, S. L.; Roberts, D. G. M.; Das, J.; Sack, J. S. Crystallographic determination of the structures of human α-thrombin complexed with BMS-186282 and BMS-189090. *Protein Sci.* **1996**, *5*, 221−228.
(27) One of the five templates is proprietary, and its structure cannot be revealed.
(28) A full description for evaluating and characterizing the combinatorial libraries described in the present work will be presented in the appropriate forum.
(29) *Available Chemicals Directory (ACD)*; Molecular Design Ltd.: San Leandro, CA, 1998.
(30) Putta, S.; Lemmen, C.; Beroza, P.; Greene, J. A Novel Shape-Feature Based Approach to Virtual Library Screening. *J. Chem. Inf. Comput. Sci.*, submitted for publication, 2001.
(31) Teig, S. L.; Smellie, A. S. Method and apparatus for conformationally analyzing molecular fragments. W09859306, U.S.A., 1998.
(32) Smellie, A.; Teig, S. Conformational analysis by intersection: ring conformations. Presented at the 217th National Meeting of the American Chemical Society, Anaheim, CA, Mar 21−26, 1999.

(33) The software generates a full set of conformations, and a maximum of 100 conformers (per stereocenter) for each active molecule is selected for this study. These selected conformers represent sampling of the complete conformational space.

(34) Information content, I, is calculated with the following equation:

$$I = -\frac{1}{N}\log\left(\left(\frac{N_a}{N}\right)^{N_a}\left(\frac{N_i}{N}\right)^{N_i}\right) + \frac{1}{N}\log\left(\left(\frac{N_{ap}}{N_p}\right)^{N_{ap}}\left(\frac{N_{ip}}{N_p}\right)^{N_{ip}}\left(\frac{N_{an}}{N_n}\right)^{N_{an}}\left(\frac{N_{in}}{N_n}\right)^{N_{in}}\right)$$

The terms are N, total number of compounds (actives and inactives); N_a, total number of actives; N_{ap}, number of actives matched (true positive); N_{an}, number of actives not matched (false negative); N_i, total number of inactives; N_{ip}, number of inactives matched (false positive); N_{in}, number of inactives not matched (true negative); and N_i (number of inactives matched). There are two terms in information content equation, the first represents the uncertainty as to whether a molecule is active, and the second accounts for the uncertainty as to whether a molecule is active given whether it fits the hypothesis. Note that (i) the information content is a function of both active and inactive molecules and (ii) the descriptors used to represent the molecules (actives and inactives) are not required to be orthogonal.

(35) Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of common functional configurations among molecules. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 563−571.

(36) The same number of conformers (100 per stereocenter) was generated (as the training set) for all of the molecules in the virtual library filtering studies.

(37) Hazen, G.; et al. *MACCSII, Facilities Guide and Reference*; Molecular Design Ltd.: San Leandro, CA.

(38) *Molecular Operating Environment (MOE, V2000.2)*; Chemical Computing Group: Montreal, Quebec, 2000.

(39) Engh, R. A.; Brandstetter, H.; Sucher, G.; Eichinger, A.; Baumann, U.; Bode, W.; Huber, R.; Poll, T.; Rudolph, R.; van der Saal, W. Enzyme flexibility, solvent and weak interactions characterize thrombin-ligand interactions: implications for drug design. *Structure* **1996**, *15*, 1353−1362.

(40) Daylight Toolkit, www.daylight.com, Santa Fe, NM, 1995.

(41) Banner, D. W.; Hadvary, P. Crystallographic analysis at 3.0A resolution of the binding to human thrombin of four active site-directed inhibitors. *J. Biol. Chem.* **1991**, *266*, 200 085−200 093.

(42) Baldino, C. M. Perspective articles on the utility and application of solution-phase combinatorial chemistry. *J. Comb. Chem.* **2000**, *2*, 97−99.

(43) Kassel, D. B. Combinatorial chemistry and mass spectrometry in the 21st Century drug discovery Laboratory. *Chem. Rev.* (Washington, DC) **2001**, *101*, 255−267.

(44) Zambias, R. A.; Kassel, D. B. Automated on-line evaporation light scattering detection to quantify isolated fluid sample compounds in microtiter plate format. W6077438, U.S.A., 2000.

(45) Cohen, J. Integrating molecular modelling, high throughput purification and biology for lead generation. Presented at the International Symposium on Laboratory Automation and Robotics, Boston, MA, 1999.

(46) Grigg, R.; Gunaratne, N.; Sridharan, V. X = Y − ZH systems as potential 1,3-dipoles. Part 14. Bronstead and Lewis acid catalysis of cycloadditions of arylidene imines of α-amino acid esters. *Tetrahedron* **1987**, *43*, 5887−5898.

(47) Kobayashi, S.; Araki, M.; Ishitani, H.; Nagayama, S.; Hachiya, I. Activation of imines by use of rare earth metal triflates. *Kidorui* **1995**, *26*, 310−311.

(48) Hajduk, P. J.; Dinges, J.; Schkeryantz, J. M.; Janowick, D.; Kaminski, M.; Tufano, M.; Augeri, D. J.; Petros, A.; Nienaber, V.; Zhong, P.; Hammond, R.; Coen, M.; Beutel, B.; Katz, L.; Fesik, S. W. Novel inhibitors of Erm methyltransferases from NMR and parallel synthesis. *J. Med. Chem.* **1999**, *42*, 3852−3859.

(49) Pavlidis, V. H.; Perry, P. J. The synthesis of a novel series of substituted 2-phenyl-4H-3,1-benzoxazin-4-ones. *Synth. Commun.* **1994**, *24*, 533−548.

JM010494Q