Articles

SQ: A Program for Rapidly Producing Pharmacophorically Relevent Molecular Superpositions

Michael D. Miller,*,† Robert P. Sheridan,‡ and Simon K. Kearsley‡

Department of Molecular Systems, Merck Research Laboratories, Sumneytown Pike, West Point, Pennsylvania 19486, and Department of Molecular Systems, Merck Research Laboratories, Rahway, New Jersey 07065

Received November 2, 1998

A new method SQ has been developed to provide fast, automatic, and objective pairwise three-dimensional molecular alignments. SQ uses an atom-based clique-matching step followed by an alignment scoring function that has been parametrized to recognize pharmacologically relevant atomic properties. Molecular alignments from SQ are consistent with known drug−receptor interactions. We demonstrate this with six pairs of receptor−ligand complexes from the Brookhaven Protein Data Bank. The SQ-generated alignment of one isolated ligand onto another is shown to approximate the alignment of the ligands when the receptors are superimposed. SQ appears to be better than its predecessor SEAL (Kearsley and Smith, *Tetrahedron Comput. Methodol.* **1990**, *3*, 615−633) in this regard. SQ has been tailored so that, given one molecule as a probe, it can be used to routinely scan large chemical databases for which precomputed conformations have been stored. The SQ score, a measure of 3D similarity of each candidate molecule to the probe, can be used to rank compounds for the purposes of chemical screening. We demonstrate this with three probes (a thrombin inhibitor, an HIV protease inhibitor, and a model for angiotensin II). In each case SQ can preferentially select from the MDDR database other compounds with the same activity as the probe. We further show, using the angiotensin example, how SQ can identify topologically diverse compounds with the same activity.

Introduction

While the structures for some pharmaceutically interesting receptors are known at atomic detail, little or no structural information is available for most of them. Chemists must therefore resort to an approach wherein the structural requirements for biological activity are deduced by comparing small molecules. One of the most common types of operations in molecular modeling is to superimpose two or more molecules onto each other. Many superposition methods have been described.[1−15] There are two major applications of superposition. One might be called “pharmacophore elucidation”. A chemist might superimpose two or more molecules and then observe which properties are common to them all at certain locations in space. Presumably the common properties are required for activity. Another application is 3D database searching, wherein a chemist uses an interesting molecule as a probe and searches a large database for other compounds which are similar in 3D structure. Not all types of 3D searches use superpositions, but the ones that do have the advantage that they generate atomic correspondences between the probe and molecules from the database.

Generally superposition methods fall into two categories: field-based [1,4,6,15] and atom-based.[11,12] In the first category are methods which attempt to quantify similarity by projecting one or more properties of the molecules into space or onto a surface. These methods have the desirable features of being unbiased by atom positions. However these techniques suffer from a need to carefully sample orientation space to avoid being trapped in local minima. They usually make comparisons based on the fit of the molecules as a whole, which makes this approach unsuitable for comparing molecules of very different sizes. Finally these methods tend to be CPU-intensive, which makes them less suitable for searching large databases. In the second category are methods that generate sets of atom−atom pairings to specify alignments. These approaches have the desirable features of being fast and capable of handling molecules of disparate size. Most current methods suffer, however, from two drawbacks. They are very sensitive to atom positions, and they have only rudimentary scoring of alignments, so that visual inspection of the alignments is often necessary.

Most current methods, of either category, assume that the molecules to be superimposed are rigid. The fact that most druglike molecules are flexible adds another layer of complexity. Typically, flexibility is addressed by using many explicit conformations for each molecule[1,4−7,11] or allowing on-the-fly flexing,[3,8−10,13,14] although only the former is usually fast enough for database searches.

An ideally useful superposition method would com-

[*] To whom correspondence should be addressed.
[†] Merck Research Laboratories, PA.
[‡] Merck Research Laboratories, NJ.

10.1021/jm9806143 CCC: $18.00 © 1999 American Chemical Society
Published on Web 04/21/1999

bine the best features of the methods discussed above and have the following characteristics:

1. It should be automatic and unbiased.

2. It should produce all relevant superpositions, including those which are nonobvious or nonintuitive.

3. It should generate a meaningful measure of the goodness of superposition.

4. It should work on very similar as well as very dissimilar molecules.

5. It should allow for local comparisons, so it will work on molecules of very different sizes.

6. It should recognize those aspects of molecules that are important for receptor recognition.

7. It should be very rapid, so database searches are practical.

In this paper we will present SQ, a new method of producing molecular alignments. SQ combines in a self-consistent manner several lines of research in our laboratory: pharmaceutically relevant atomic descriptors,[16,17] technology for producing representative conformations,[18] clique-searching strategies,[19] and a robust and objective superposition scoring function. We demonstrate with specific examples how SQ possesses most of the features that make for a useful superposition method.

Methods

Atomic Representation. In SQ we typically consider only non-hydrogens. Each atom has a composite "SQ type" which includes information about atomic number, hybridization, and physiochemical types. Physiochemical types are the same used in our work in FLOG.[19] These types are 1 = cations, 2 = anions, 3 = neutral H-bond donors, 4 = neutral H-bond acceptors, 5 = polar (unspecified H-bonding group), 6 = hydrophobic, 7 = other. The physiochemical types are meant to represent ionization states at physiological pH. How assignments are made is given in Bush and Sheridan.[20] A square matrix **P** stores the user-assigned similarity of any of the 43 SQ types with any other. During the development of SQ, the original elements of **P** were assigned by our intuitive ideas about how much the physiochemical types should resemble each other. These were modified for element type and hybridization. Some values of **P** were adjusted manually so better superpositions could be obtained for a set of druglike molecules. Our final values of **P** are available as Supporting Information.

Overview of Superposition Algorithm. For any SQ superpostion, one molecule is the "probe" and one is the "candidate". The candidate is to be optimally superimposed on the probe. Both are held rigid. In the case of database searches, each candidate may be a different conformation of the same molecule. SQ superposition is a two-step process. Initial superpositions of the candidate onto the probe are generated by clique-finding algorithm. The second step optimizes the superposition so that the superposition score as measured by our SQuEAL function (*S*teric and *Qu*alitative *E*lectronic *AL*ignment) is maximum. The overall strategy is given in Scheme 1.

Clique-Finding. As in the earlier descriptions of a similar algorithm for FLOG,[19] initial superpositions of each conformation onto the probe (steps 2.2−2.3) are generated by a method of systematic distance matching wherein a clique-finding algorithm looks for sets of candidate atom−probe atom pairs such that all the candidate atom−atom distances are the same as the corresponding distances in the probe within a given tolerance. The minimum number of pairs *nodlim* and the distance tolerance *dislim* are under user control. The default values, 4 atoms and 1.5 Å, have been found to work well in a variety of situations.

Scheme 1

```
1. Read in the probe. Assign the atom properties. Calculate the atom-atom distances.
2. For each candidate conformation:
        2.1 Read the coordinates.  Assign the atom properties.
            Calculate the atom-atom distances.
        2.2 Find all sets of candidate atom- probe atom pairs that constitute a
            colored clique (see text).
        2.3 For each clique:
                2.3.1 Produce an initial superposition from the pairs .
                2.3.2 Optimize the superposition with a simplex algorithm until
                      the SQuEAL function gives a maximum value.
        End 2.3
        2.4  Record the superposition with the highest SQuEAL score from all the cliques.
End 2.
```

An innovation relative to our previous work is the development of "colored cliques". That is, atom *i* in the probe and atom *j* in the candidate are allowed to match only if $W(i,j) > 0$ (see the section on the SQuEAL function), that is, if they have similar character. The inclusion of this simple match constraint, by keeping only the most promising matches, greatly reduces the number of cliques that must be optimized. This is similar in spirit to the "labeled matching" method described by Schoichet and Kuntz.[21] Details of the clique-finding algorithm are in the Appendix.

As before we also use the concept of "essential atoms" in the probe. Given N_{ess} essential atoms, the user may specify that at least N_{req} essential atoms appear in every clique where $1 \le \mathrm{N}_{\mathrm{req}} \le N_{\mathrm{ess}}$. Clique building starts at essential atoms, and cliques with less than N_{req} essential atoms are eliminated. This adds an additional speed-up. Essential atoms are one way of including SAR information. In SQ we allow for three distinct types of essential atoms: +, $, and %. The essential atom + must be matched by some atom in the candidate, but the candidate atom can be any type. The essential atom $ must be matched by an atom in the candidate with the same atomic number (e.g., nitrogen with nitrogen). The essential atom % must be matched by an atom in the candidate of the same physiochemical type (e.g., cation with cation). If no essential atoms are assigned by the user, the atom closest to the centroid of the probe is assigned as +.

SQuEAL Scoring Function. This superposition scoring function is invoked in step 2.3.2 in Scheme 1. It is highly modified from the SEAL function[4] which may be considered its ancestor. The score (eq 1) is a function of the particular relative orientation of candidate and probe:

$$\mathrm{score} = \sum_{j=1}^{\mathrm{candidate}} \sum_{i=1}^{\mathrm{probe}} W(i,j)O(i,j) + \text{cavity term} + \text{restraint term} \quad (1)$$

The more positive this number, the better the superposition. The term $W(i,j)$ is a measure of the overall similarity of atoms *i* and *j* in the context of a molecule and is of the form:

$$W(i,j) = \beta\,[\text{atomic property similarity}] + (1-\beta)\,[\text{steric environment similarity}] \quad (2)$$

where "atomic property similarity" is the similarity of the character of the atoms *i* and *j* and is a function of $P(i,j)$, while "steric environment similarity" measures the similarity of the "exposure" of the two atoms. Detailed definition of the two components is given in the Appendix. The parameter β controls the balance between the two components and $0 \le \beta \le 1$. The default value is $\beta = 0.5$. The atomic overlap term $O(i,j)$ is of the form:

$$O(i,j) = \exp(-\alpha r(i,j)^2) \quad (3)$$

where $r(i,j)$ is the distance between *i* and *j*. The parameter α controls whether the alignment hypersurface will be more steep around the atoms ($\alpha \sim 0.5$) or more flat ($\alpha \sim 0.1$). Values between 0.3 and 0.4 seem the most useful, and the default value is $\alpha = 0.3$ ($O(i,j)$ reaches half its maximal value at 1.5 Å—about a bond length). The cavity term is of the form:

$$\text{cavity term} = \max(0, r_c - r(i,j)) \tag{4}$$

It penalizes situations where atoms in the candidate extend more than a distance of r_c, the cavity radius, from the nearest probe atom. The default value is $r_c = 6.0$; that is, the candidate can be somewhat bigger than the probe. The penalty restraint is of the form:

$$\text{restraint term} = -10r(i,j) \tag{5}$$

It applies only to the $i-j$ pairs where i is an essential atom and j is a candidate atom that is matched to i. This restraint keeps the matched atoms from drifting too far from each other during simplex optimization.

Simplex Optimization. As in our previous work, the spatial orientation of each molecule is characterized as a quaternion rotation matrix, plus three components for translation. The quaternion is well-behaved with the simplex optimizer. The optimizer maximizes the score by moving the candidate as a rigid body until the orientation differs from the previous step by less than 0.01 Å translation and 0.5° rotation. This typically takes 10−15 iterations. The optimization can move the initial orientation by as much as 12 Å and 180°.

Results

There are two types of experiments where we will demonstrate the utility of SQ. The first is to show that SQ produces good alignments of one rigid molecule onto another. The second is to show that SQ can select active compounds out of large databases.

Pairwise Superposition. How do we objectively and unambiguously evaluate whether a method is producing pharmaceutically optimal superpositions? One approach is to examine the receptor−ligand complexes for which we have two different ligands on the same receptor. In the first phase of the experiment, one superimposes the receptor structures. The ligands are carried along, and the orientation of one ligand relative to the other can be taken as the “observed superposition” of those ligands. In the second phase, one superimposes one isolated ligand onto the other using the superposition method to be tested. We can call the orientation from this phase the ”predicted superposition”. The predicted superposition is then compared with the observed superposition. This approach was used by Klebe and coworkers[5] when they investigated the effects on adjustable parameters when adding extra terms to the SEAL function. For this work we chose six pairs of crystal structures from the Brookhaven Protein Data Bank[22] with ligands possessing a wide variety of chemical functionality. Chemical structures for these are shown in Figure 1.

The observed superpositions were produced by superposing the equivalent α carbons of the two proteins and then extracting the ligands. The calculated root-mean-square residual (RMS) from the alignment of one protein on another averaged about 0.3 Å. This reflects the precision of the experimental data as well as protein conformational changes which take place upon ligand binding and provides a lower limit to the RMS for this type of experiment. Predicted superpositions were generated by running SQ on the ligand coordinates directly from the crystal structures. The default parameters were used except that $r_c = 20.0$ to allow for differences in sizes. Four values of α were tried. The top two highest-scoring predicted superpositions S1 and S2 were kept. The difference of these superpostions from the observed superpositions was measured by the RMS over non-hydrogen atoms. The SEAL method[4] is used for comparison. The results of these trials are summarized in Table 1. An example of a predicted versus observed superposition is shown in Figure 2.

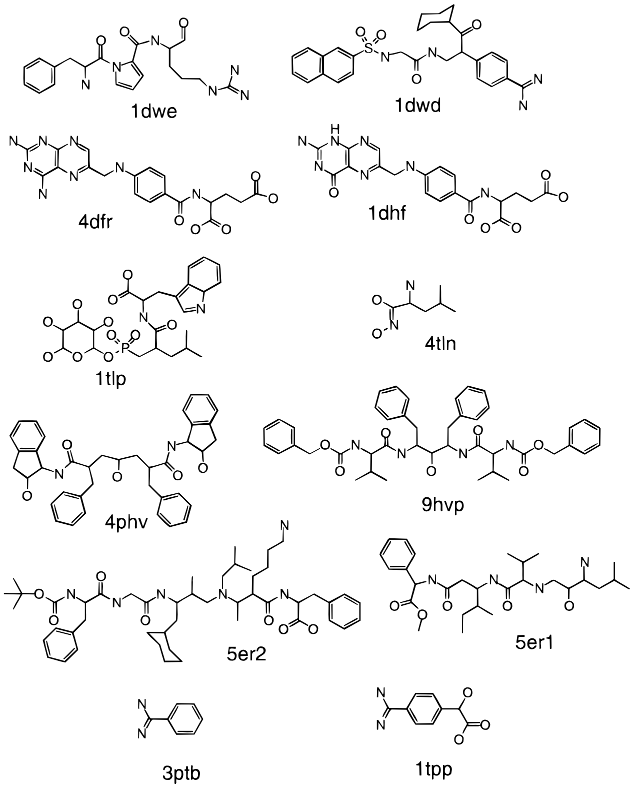


Figure 1. Ligands from six pairs of ligand−enzyme crystal complexes used in the evaluation of the SQuEAL alignment function. The receptors are as follows: 1dwe and 1dwd, human thrombin; 4dfr and 1dhf, *E. coli* dihydrofolate reductase; 1tlp and 4tln, *Bacillus* thermolysin; 4phv and 9hvp, HIV-1 protease; 5er2 and 5er1, endothiapepsin; 3ptb and 1ttp, bovine trypsin.

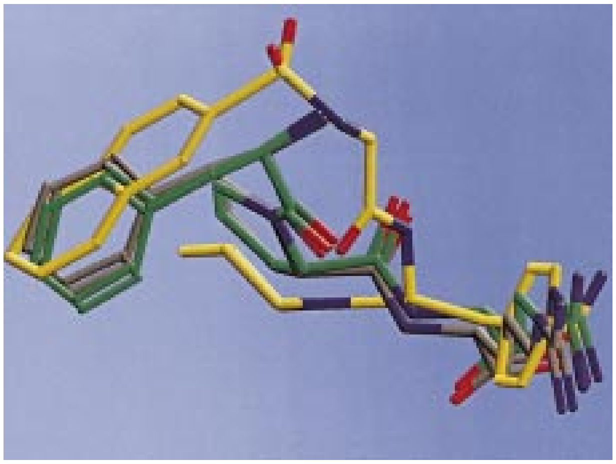


Figure 2. Representative superposition obtained with SQ. In yellow is the probe molecule from 1dwd. The structure with green carbons is the ligand from 1dwe acting as the candidate. It is in the “observed” orientation obtained by superimposing the α carbons of the 1dwe protein onto the 1dwd protein. The structure with gray carbons is the candidate oriented by SQ with $\alpha = 0.3$.

Table 1. Results Obtained for Aligning Six Pairs of Protein Ligands

			$\alpha = 0.2$		$\alpha = 0.3$		$\alpha = 0.4$		$\alpha = 0.5$	
ligand pair[a]	time[b]	superposition	score	RMS[c]	score	RMS	score	RMS	score	RMS
			SEAL							
ldwe/ldwd	2.96	S1	140.08	0.87	139.48	0.26	82.8	0.82	67.88	0.82
		S2	122.5	1.36	106.15	0.83	68.33	1.36	55.51	1.37
4dfr/1dhf	3.55	S1	180.55	0.19	141.75	0.19	120.26	0.19	106.4	0.19
		S2	178.37	0.23	139.49	0.23	74.83	1.45	62.81	1.44
1tlp/4tln	1.00	S1	61.03	0.77	43.25	0.8	33.55	0.79	28.26	0.89
		S2	56.98	1.45	38.19	1.46	32.07	0.84	24.98	1.07
4pvh/9hvp	14.89	S1	187.15	0.12	132.55	0.12	103.94	1.72	85.92	1.72
		S2	186.14	1.72	132.19	1.72	103.61	0.12	85.17	0.13
5er2/5er1	3.84	S1	141.96	1.65	101.79	1.66	81.97	1.66	70.	1.66
		S2	126.01	1.8	89.6	1.81	72.36	1.81	60.77	1.81
3ptb/1tpp	0.20	S1	87.79	0.7	73.72	0.67	64.05	0.7	59.47	0.67
		S2	77.52	0.66	71.63	0.81	56.91	0.65	45.71	0.85
RMSmean[d]				0.88		0.81		1.12		1.13
DELTAmean[e]				0.09		0.12		0.18		0.22
			SQ							
ldwe/dwd	0.52	S1	226.93	0.08	172.85	0.07	141.04	0.06	117.5	0.09
		S2	124.09	1.3	86.42	1.3	80.54	1.36	53.53	1.3
4dfr/1dhf	0.88	S1	230.75	0.19	187.33	0.19	162.51	0.2	145.3	0.21
		S2	113.73	1.29	184.28	0.24	100.09	0.48	67.22	0.48
1tlp/4tln	0.53	S1	48.21	0.18	41.39	0.17	37.04	0.18	33.93	0.19
		S2	32.82	1.76	25.55	1.72	21.1	1.71	18.98	1.69
4pvh/9hvp	3.68	S1	250.83	0.11	186.6	0.11	149.73	0.11	124.4	0.11
		S2	234.97	1.72	180.97	1.72	144.32	1.72	120.4	1.72
5er2/5er1	1.40	S1	136.14	0.61	87.58	0.73	68.77	0.68	49.73	0.83
		S2	135.15	2.06	72.58	1.48	63.	2.07	48.49	2.11
3ptb/1tpp	0.11	S1	108.25	0.17	91.42	0.12	80.99	0.13	74.18	0.14
		S2	108.09	0.67	91.17	0.67	80.44	0.68	73.63	0.69
RMSmean				0.29		0.32		0.31		0.37
DELTAmean				0.31		0.27		0.30		0.36

[a] The first ligand is taken as the probe and the second as the candidate. [b] Seconds on a R4400 Silicon Graphics workstation. The time is insensitive to the value of α. [c] RMS differences (Å) are those computed between the ligand as bound to the protein versus the ligand as superposed onto the probe. [d] Root mean square over six examples. [e] Root mean square score |S1 - S2|/S1 over six examples.

Table 2. Database Screening Experiments Against MDDR Flexibase

3D probe	source of probe	keywords used to determine activity	no. of actives
NAPAP	crystal structure (leps)	thrombin inhibitor	300
indinavir	gas-phase energy minimization	HIV-1 protease inhibitor	328
AII peptide	model-building	angiotensin II blocker	2031
–	–	PAF antagonist (control)	1311
–	–	renin inhibitor (control)	1144
–	–	antidiabetic (control)	993

Here we use two measures to assess SQ versus SEAL. The first, taken directly from Klebe et al., is the RMSmean, the RMS displacement for S1 over our six examples. This measures how close the superpositions are to experiment. The smaller the RMSmean the better the method. The second, modified from Klebe et al., is DELTAmean which is the difference in score between S1 and S2 expressed as a fraction of S1. This is one measure of the spread in scores. The larger the DELTAmean the better the method. For most values of α, RMSmean is about the same or lower for SQ than for SEAL and DELTAmean is higher. Thus SQ is shown to get better results than SEAL. Moreover, SQ is seen to take about one-third the time of SEAL.

Overall the highest-scoring SQ superposition approximates the experimental superposition well with the RMSmean being better than 0.37, with the results being fairly insensitive to α. The fact that the SQ-predicted superpositions are about as good as the lower limit of 0.3 Å is an impressive result, as none of the external forces imposed by the receptor are included in the SQ method. This implies that the superposition of isolated ligands can contain implicit information about their binding mode to the same receptor.

DataBase Screening. In the previous section we have demonstrated the ability of SQ to produce accurate molecular superpositions of single molecules. In this section we will investigate whether the SQ score can discriminate between molecules. To assess the suitability of the SQ function for database searching we have constructed a flexibase[18] from the MACCS Drug Data Report (MDDR),[23] a licensed database of druglike compounds compiled from the patent literature. The MDDR flexibase is composed of ~70 000 molecules represented by ~1.1 million explicit conformations.

MDDR compounds have associated with them one or more keywords in the “activity field”. The keywords we use here are listed in Table 2. We will assume, for the purposes of this study, that a compound with the keyword “thrombin inhibitor”, for example, is an active and that a compound without this keyword is an inactive.

We will conduct a series of virtual screening experiments. In each we will select a 3D probe and compare

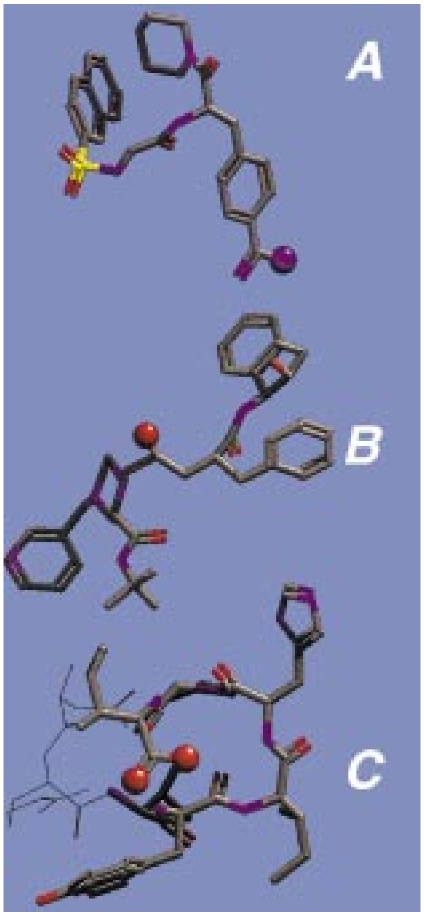


Figure 3. SQ probes used in the database searching experiments. (A) The crystal structure of NAPAP as cocrystallized with human thrombin (1DWD). One of the guanidine nitrogens was declared essential type %, given that many thrombin inhibitors have cations at that position. (B) The global energy minimum conformation of the HIV-1 protease inhibitor indinavir. The hydroxyl oxygen was declared essential type +, given the importance of having an atom to hydrogen bond to the active site aspartate, usually the hydroxide in many HIV protease inhibitors. Nondefault SQ parameters: *nodlim* = 7, $\alpha = 0.4$. (C) A model of angiotensin II. All atoms in residues 3−8 are match centers (stick). Other atoms (wire) are ignored. Atoms shown as spheres were declared essential type %, given the importance of having an anion at that position. One was required for a match. Nondefault SQ parameters: *nodlim* = 7, $\alpha = 0.4$.

it to every conformation in the MDDR flexibase. The best-scoring conformer is retained per compound, and then the compounds are ranked by decreasing SQuEAL score. Compounds are "assayed" in this order, and we monitor how fast actives accumulate. (The same method was used in Kearsley et al.[16] to measure the effectiveness of topological descriptors.) For instance, if the probe is a thrombin inhibitor, we will monitor the accumulation of MDDR molecules with the keyword "thrombin inhibitor". We can also monitor the accumulation of actives in an unrelated activity, for instance "renin inhibitor", as a control.

3D probes can be obtained in a number of ways depending on the level of information available to the modeler. The SQ probes for our experiments are shown in Figure 3. The first is the structure of NAPAP cocrystallized with thrombin.[24] It is an example of a probe taken from experimental data. The second probe is the HIV-1 protease inhibitor indinavir[25] in the conformation of lowest energy by the MMFF force field.[26] It is an example of a modeled structure. The third probe is a model proposed for the receptor-bound conformation of angiotensin II.[27] This is an especially interesting probe because it is "discontinuous"; that is, we expect the candidates to match parts of the probe that are not nearby in bond topology.

If we use one or two essential points, screening against the MDDR takes roughly 84 h on a single processor Silicon Graphics Indigo2 with R10000 processor, for a screening rate of 224 conformations/min. These searches are trivially parallelizable, and we usually run searches on several processors simultaneously, giving us a completed search overnight.

The results of the virtual assays are summarized in Figure 4. A biological assay with typical throughput will accommodate at least a few hundred samples. We therefore show the results of the top 500 rank-ordered compounds. If the SQ score had no utility in selecting actives, actives would appear at a rate proportional to their frequency in the entire database. This is indicated by the line marked "random". In every case there is a significant increase in the selection of the appropriate actives over random, sometimes as much as 20-fold. That is, nearly one out of three or four compounds would be an active. Most of the control sets of actives perform similarly to random or worse; i.e., they are selected against. The only control set that is significantly above random is the renin inhibitors in Figure 4B. This is not surprising, since renin inhibitors and HIV protease inhibitors can be very similar. However, the renin inhibitor curve is still consistently below the HIV-1 protease inhibitor curve. Thus the SQ score demonstrates a real ability to select for compounds of a given activity. Generally, we find that the ranks in any given search are not very sensitive to the SQ parameters (data not shown).

Structural Diversity. Many methods, both 2D and 3D, exist for selecting active compounds from large databases. While it is important that a method select active compounds at a much greater rate than expected by chance, it is also important that the method select compounds in more than one structural class. 3D similarity methods, which ignore bonding, can usually find molecules that are topologically different yet present similar spatial character. To demonstrate this specifically for SQ, the angiotensin II results will be examined in more detail. There a few thousand angiotensin II blockers, which fall into distinct structural classes. Table 3 lists one possible classification of the nonpeptide blockers and their rank in the 250 highest-scoring compounds. Table 3 shows a very good enrichment in angiotensin blockers; nearly one out of every five compounds tested is active. It is especially significant that all these nonpeptide blockers were selected using a peptide probe. Out of the seven major structural classes, all have at least one representative with rank ≤ 250, and four have at least one representative with rank ≤ 100. Figure 5 shows an example of a small molecule, with the external registry number 193559 in the MDDR database, docked onto the angiotensin II model.

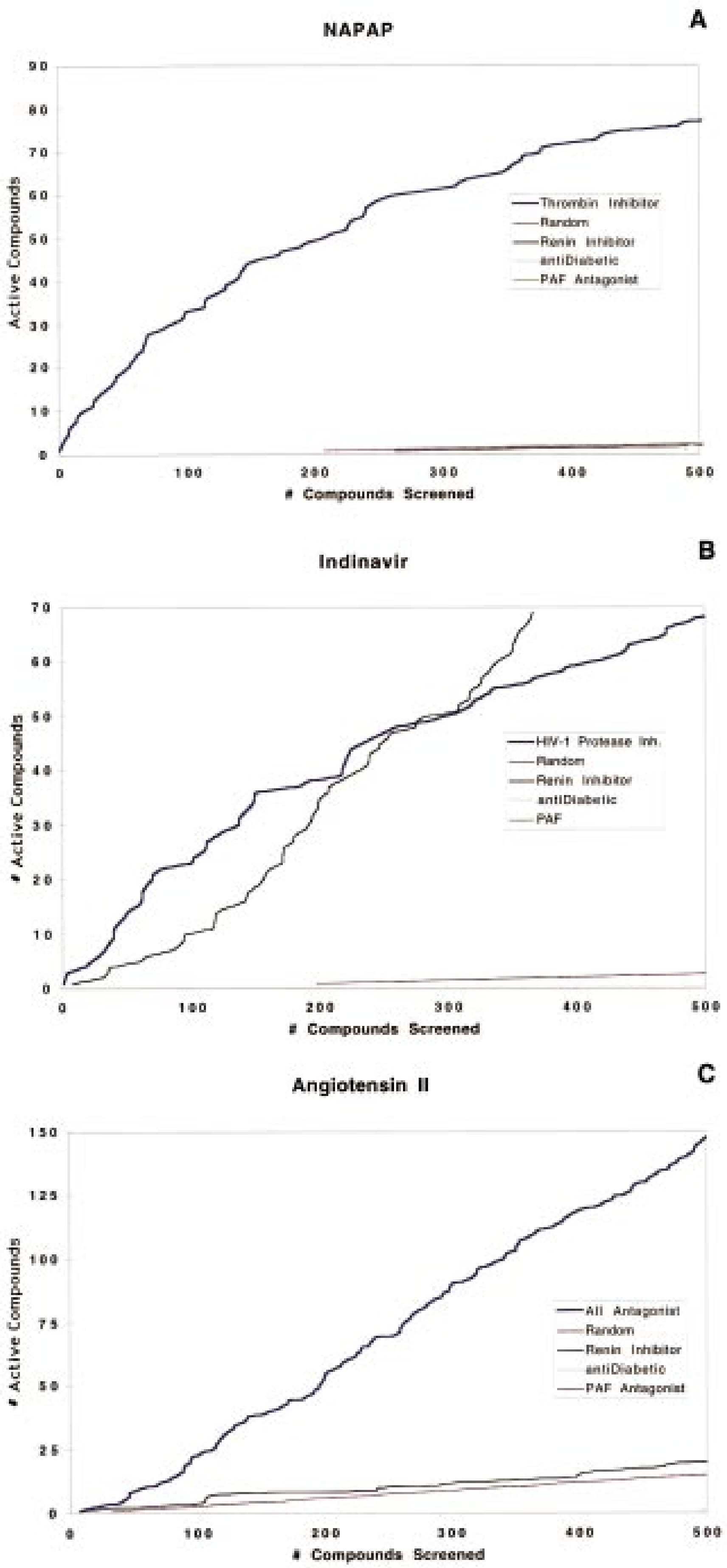
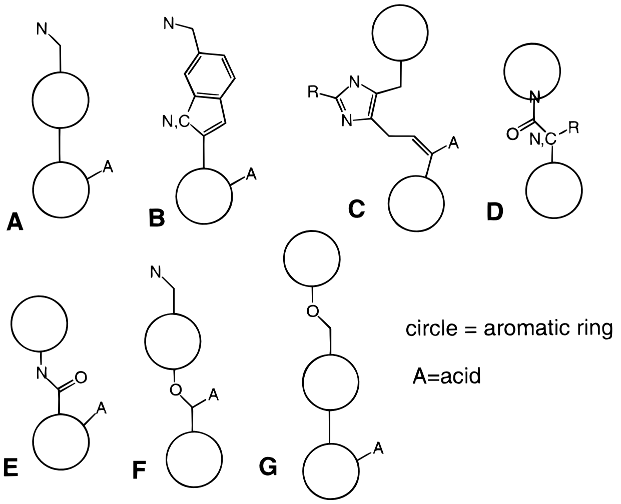

NAPAP
A
Active Compounds
90
80
70
60
50
40
30
20
10
0
0
100
200
300
400
500
# Compounds Screened
Thrombin Inhibitor
Random
Renin Inhibitor
antiDiabetic
PAF Antagonist
Indinavir
B
# Active Compounds
70
60
50
40
30
20
10
0
0
100
200
300
400
500
# Compounds Screened
HIV-1 Protease Inh.
Random
Renin Inhibitor
antiDiabetic
PAF
Angiotensin II
C
# Active Compounds
150
125
100
75
50
25
0
0
100
200
300
400
500
# Compounds Screened
AII Antagonist
Random
Renin Inhibitor
antiDiabetic
PAF Antagonist

Table 3. Angiotensin Actives in MDDR Classified by Structural Families and Their Ranks in the SQ Search



class A	compound	193196	214158	213662	193258	210449	215497			
	rank	44	52	97	160	203	204			
class B	compound	193400	193399	193061						
	rank	77	112	113						
class C	compound	211727	173764	220277	220273	179243	191960	211682	179014	193061
	rank	8	24	40	46	54	57	99	109	113
	compound	220274	212944	210457	218116	211637	175599	220275	220276	188900
	rank	115	120	129	135	153	159	161	163	167
	compound	187010	213683	191141	189952					
	rank	178	185	192	218					
class D	compound	211908	218282	211907	179025	179029	216455	210825	211910	213196
	rank	14	60	74	90	128	140	197	227	230
	compound	179024								
	rank	246								
class E	compound	193557								
	rank	174								
class F	compound	215549	179767							
	rank	211	217							
class G	compound	191064	190365							
	rank	145	221							
other	compound	215418	193559	215419	215550					
	rank	45	61	103	173					

Discussion

Two major applications for alignment algorithms are generating the alignments of a small number of molecules and searching databases. We have demonstrated that SQ produces pharmacologically meaningful alignments of individual molecules even if the molecules are not particularly similar topologically. Also we have shown that SQ is fast enough to be routinely used to search large 3D databases. The SQuEAL function is discriminating enough that the highest-scoring compounds from the databases are greatly enriched in actives and general enough that many of the actives are from diverse chemical classes. Merck scientists have discovered many novel active compounds by doing SQ searches.

Most of the robustness of SQ comes from the fact that it is a two-layer procedure. The clique-matching step is equivalent to atom-based matching methods such as DISCO,[11] and the simplex optimization is equivalent to many field-based methods such as SEAL.[4] Thus the best features of atom-based and field-based methods are combined. The clique-matching rapidly provides a small diverse sampling of promising initial orientations that the slower simplex optimization step can work on. This helps keep the optimizer from spending large amounts of time investigating less promising superpositions. Also the optional use of essential points in the clique-matching step permits the user to put severe constraints on the type of superpositions that are allowed. This adds some of the flavor of substructure searching to SQ, so that unwanted superpositions can be quickly eliminatedfrom consideration. Conversely, the simplex optimization step allows the superpositions to drift from the original orientation dictated from the clique-matching. Also the SQuEAL score quantifies the overlap of all the atoms, not just the ones in the clique. Adjustable parameters allow the user to change the relative influence of clique-matching and simplex optimization on the final superpositions.

Many field-based superposition methods (see, for example, refs 4 and 7) use the overlap of electrostatic potentials or similarity between partial charges as a

Figure 4. Plots of number of active molecules found versus number of molecules tested for searches over the MDDR database with the three probes in Figure 3. Some of the curves are not visible because the first active occurs after the first 500 molecules tested. (A) NAPAP as a probe. The corresponding actives are "thrombin inhibitors". (B) Indinavir as a probe. The corresponding actives are "HIV-1 protease inhibitors". (C) Angiotensin II as a probe. The corresponding actives are "angiotensin II blockers".

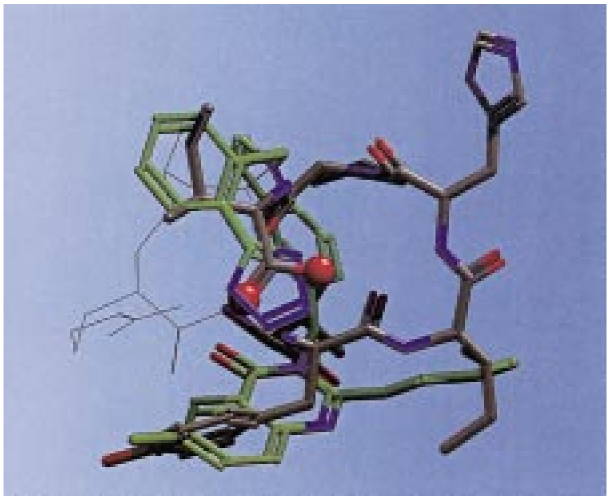


Figure 5. Superposition of 193559 (green) with the model of angiotension II in Figure 5 (gray).

scoring function. This requires having at least polar hydrogens in the models and partial charges on all atoms. These types of scoring functions tend to be dominated by the atoms with the most extreme charges, and many local details around individual atoms are lost in favor of global properties such as the dipole moment. The SQuEAL function, in contrast, uses only non-hydrogen atoms. This greatly smooths the superposition scoring surface. That pharmaceutically relevant physiochemical types are used instead of partial charges prevents a few atoms from dominating the score. Thus the SQuEAL function attains a good balance between local and long-range effects.

SQ treats both candidate and probe molecules as rigid, and one potential limit of our current implementation is that it depends on having several precomputed conformations of candidate molecules. It is possible to miss some important superpositions and/or miss compounds in a database search because key conformations are missing. Some methods, for instance genetic algorithms (see ref 10), have been implemented wherein molecules can adjust their conformations on-the-fly. However, these are generally so time-intensive that only a few pairs of molecules can be superimposed in a reasonable time, and thus the methods are not usually applicable to database searches. We have found that a reasonably complete sampling of conformational space can be precomputed at not too large a cost, so that missing conformations are usually not a problem for us.[18] Moreover, we also argue that it does not matter whether some compounds are missed as long as the set of selected compounds is enriched in actives.

The cost effectiveness of 2D versus 3D methods is an important issue in database searches. Certainly, 3D methods are more computationally and informationally expensive. A lot of structural and/or SAR information is usually needed to specify a 3D query. Also, for us the flexibase approach incurs the computational cost of precomputing conformers. Moreover, SQ, while relatively rapid for a 3D similarity method, is slow compared to most 2D similarity methods. We have compiled a lot of experience running comparable probes with SQ and TOPOSIM,[16] our in-house descriptor-based 2D similarity method. A straighforward comparison of TOPOSIM and SQ is difficult. TOPOSIM is descriptor-based and all parts of the probe molecule are treated uniformly, while SQ is based on superposition and parts of the molecule can be emphasized with essential points. The atomic descriptors are quite different. Finally, the relative sizes of the probe and candidate are handled differently. Still, one can make some generalizations. For druglike probes the number of actives found in the top-scoring 500 molecules is roughly the same for TOPOSIM and SQ. However, SQ usually finds very different actives than TOPOSIM. (Our descriptor-based 3D similarity search system GEOSIM[17] finds sets of actives much more like TOPOSIM than like SQ.) On the other hand, when the probe is a folded peptide, TOPOSIM usually cannot find any nonpeptide actives in the top 500, whereas SQ has no difficulty doing so, as seen in our angiotensin II example. This is a clear advantage of considering positions in space, rather than through-bond distances. How many actives are found is not the only criterion of cost effectiveness. Another consideration is the diversity of the top 500 molecules. Typically SQ generates a much more diverse set than TOPOSIM. Again, this is due to SQ seeing only atoms positions and ignoring bonds. Finally, a very important feature of a method like SQ is that during the search it suggests the best superposition of each of the candidates onto the probe, something TOPOSIM and GEOSIM, which are descriptor-based, cannot do. Superpositions are essential for the chemist to suggest pharmacophores or propose chemical modifications.

SQ has proved very versatile, and many useful extensions, beyond those presented here, have been implemented at Merck. For instance, SQ is trivially modified so that the probes and candidates can include various “dummy atoms” with unique match properties. Examples are ring centroids, ring normals, and hydrogen-bond extenders. These can be used to further constrain the desired type of superpositions for specific problems. Also, the SQ program has been written so that the SQuEAL function can be replaced by another scoring function. For instance, one can take the score from a precalculated regular grid. If the grid is generated from a receptor site, SQ becomes equivalent to our docking program FLOG.[19] The original use for SQ at Merck was for database searching. However, we have implemented a method (MEGA-SQ) for pharmacophore elucidation (similar in concept to DISCO[11]) wherein precalculated conformations of many molecules are compared pairwise with SQ. Ensembles of one conformation from each molecule are then constructed. The ensembles with the highest sum of pairwise scores represent the best superposition of the molecules. MEGA-SQ will be presented separately at a future date.

Conclusion

We have developed a program (SQ) that performs automatic objective pairwise alignment of molecules. It produces alignments that are pharmaceutically relevant, and it can select diverse active compounds from large databases, making it an extremely useful tool in a medicinal chemistry laboratory.

Acknowledgment. The development of SQ would not be possible without the feedback of many members of Molecular Systems in the past 6 years. The authors

Scheme 2

```
2.2.1 For pairs of probe atoms i and candidate atom j=1 to $N_{candidate}$ where W(i,j) > 0

    2.2.1.1 Reset atoms that can be matched: MP(all)=1; MC(all)=1;
        Pair i-j is the first pair in the clique:
            npair=1; CP(npair)=i; CC(npair)=j ; MP(i)=0; MC(j)=0; i'=i, j'=j

    2.2.2.2 For pairs of probe atom k and candidate atom m where the
            the following is true:
            1) MP(k)=1 and MC(m)=1 and
            2) W(k,m) > 0
            3) The distances between k and all the atoms in CP match,
               within dislim, all the corresponding distances between m and
               all the atoms in CC.
        Calculate the residuals R(i',j',k,m) (eq 2)
        If R(i',j',k,m) < MinR then MinR=R(i',j',k,m); k'=k; m'=m;
    End 2.2.2.2
    2.2.2.3 If there is no such pair, then go to 2.2.2.5
    2.2.2.4 Add the k' - m' pair to the clique:
                npair=npair+1; CP(npair)=k'; CC(npair)=m' ; MP(k')=0; MC(m')=0
                i'=k' ; j'=m' ; Reset
        Go to 2.2.2.2
    2.2.2.5 The clique is finished. If npair $\geq$ nodlim, record the clique.
End 2.2.1.
```

especially wish to thank the members of the Applications Group who applied SQ to practical problems.

Appendix

Clique-Building. The pseudocode for the clique-building step is given below**.** A clique is a paired set of atoms in arrays **CP** and **CC.** A "data mask" in the form of temporary arrays **MP** and **MC** is used to track which probe or candidate atoms are available for matching (1) or have already been matched (0) (Scheme 2).

This algorithm is "comprehensive" in that all probe−candidate atom combinations are tried. Our approach is that at each step in the search, the algorithm selects the atom−atom pair for which the weighted distance difference $R(i',j',k,m)$ (eq A1) relative to the previously selected pair is a minimum; that is, it traverses the path of "minimum weighted residuals". Once a path is complete, the root atom pairs are examined for alternate allowed paths until none are identified. We refer to this multipass approach as "breath first−depth second". We tried a number of weighting schemes for computing $R(i',j',k,m)$. In our earlier work,[19] $R(i',j',k,m) = |D(i',k) - D(j',m)|$.

We find that a more complicated scheme improves the results. In addition to the first term for $R(k,m)$, we include a second term that downweights small distances and a third term that emphasizes good atom pair matches.

$$R(k,m) = |D(i',k) - D(j',m)| + [100/D(i',k)]\exp(-2(D(i',k) - 2)) + 0.5/W(k,m)^2 \quad \text{(A1)}$$

SQuEAL Function. The individual terms for $W(i,j)$ in eq 2 in the main text are defined in the following way:

$$\text{atomic property similarity} = c_p P(i,j) + c_s(0.5 - |P(i,j)|) \quad \text{(A2)}$$

The second term upweights near-neutral interactions which would otherwise be swamped out by the first term. The default are $c_p = 1.5$ and $c_s = 0.20$. The values of **P** are available as Supporting Information.

$$\text{steric environment similarity} = c_o|E(i) - E(j)| \quad \text{(A3)}$$

The term $E(i)$ is a rapid way to approximate the exposed surface area of atom i by looking at two other atoms (k and m) in the same molecule.

$$E(i) = 0.5\sum_{k\neq i}\sum_{m\neq i}(1 - r_{dot})\exp(-0.3(r_{ave} - 3.0)^2) \quad \text{(A4)}$$

If **a** and **b** are the vectors $i \rightarrow k$ and $i \rightarrow m$, r_{dot} is the dot product of the normalized **a** and **b** and r_{ave} is the mean length of **a** and **b**. The default is $c_o = 0.33$. We find that this equation gave significantly better results than the original SEAL volume expression.

Supporting Information Available: Matrix **P** that specifies the similarity between SQ types. This information is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Lemmen, C.; Hiller, C.; Lengauer, T. RigFit: a new approach to superimposing ligand molecules. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 491−502.

(2) Masek, B. B.; Merchant, A.; Matthew, J. B. Molecular skins, a new concept for qualitative shape matching of a protein with its small molecule mimics. *Proteins* **1993**, *17*, 193−202.

(3) Perkins, T. D. J.; Dean, P. M. An exploration of a novel strategy for superposing several flexible molecules. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 155−172.

(4) Kearsley, S. K.; Smith, G. M. An alternative method for the alignment of molecular structures: maximizing electrostatic and steric overlap. *Tetrahedron Comput. Methodol.* **1990**, *3*, 615−633.

(5) Klebe, G.; Meitzner, T.; Weber, F. Different approaches toward an automatic structural alignment of drug molecules: Applications to sterol mimics, thrombin and thermolysin inhibitors. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 751−778.

(6) Perry, N. C.; van Geerstein, V. J. Database searching on the basis of three-dimensional molecular similarity using the SPERM program. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 607−616.

(7) Thorner, D. A.; Willett, P.; Wright, P. M.; Taylor, R. Similarity searching in files of three-dimensional chemical structures: representation and searching of molecular electrostatic potentials using field-graphs. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 163−174.

(8) Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Drug Des.* **1995**, *9*, 532−549.

(9) McMartin, C.; Bohacek, R. S. QXP: Powerfull, rapid computer algorithms for structure-based drug design. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 333−334.

(10) Handschuh, S.; Wagener, M.; Gasteiger, J. "Superposition of three-dimensional chemical structures allowing for conformational flexibility by a hybrid method. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 220−232.

(11) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 83−102.

(12) Artymiuk, P. J.; Bath, P. A.; Grindley, H. M.; Pepperrell, C. A.; Poirrette, A. R.; Rice, D. W.; Thorner, D. A.; Wild, D. J.; Willett, P.; Allen, F. H.; Taylor, R. Similarity searching in databases of three-dimensional molecules and macromolecules. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 617−630.

(13) Lemmen, C.; Lengauer, T.; Klebe, G. FLEXS: a method for fast flexible ligand superposition. *J. Med. Chem.* **1998**, *41*, 4502−4520.

(14) Jain, A. N.; Deitterich, T. G.; Lathrop, R. H.; Chapman, D.; Critchlow, R. E., Jr.; Bauer, B. E.; Webster, T. A.; Lozano-Perez, T. Compass: a shape-base machine learning tool for drug design. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 635−652.

(15) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. A molecular field-based similarity approach to pharmacophoric pattern recognition. *J. Mol. Graph. Modelling* **1997**, *15*, 114−121.

(16) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical similarity using physiochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.

(17) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical similarity using geometric atom pair descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128−136.

(18) Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P.; Miller, M. D. Flexibases: A way to enhance the use of molecular docking methods. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 565−582.

(19) Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P. FLOG: a system to select quasi-flexlible ligands complementary to a receptor of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 153−174.

(20) Bush, B. L.; Sheridan, R. P. PATTY: a programmable atom typer and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756−762.

(21) Shoichet, B. K.; Kuntz, I. D. Matching chemistry and shape in molecular docking. *Protein Eng.* **1993**, *6*, 723−732.

(22) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **1977**, *112*, 535−542.

(23) *MACCS−II Drug Data Report*, MDDR V97.2; Molecular Design Ltd.: San Leandro, CA, 1997.

(24) Banner, D. W.; Hadvary, P. Crystallographic analysis at 3.0 Angstrom resolution of the binding to human thrombin of four active site-directed inhibitors. *J. Biol. Chem.* **1991**, *266*, 20085−20093.

(25) Chen, Z.; Li, Y.; Chen, E.; Hall, D. L.; Darke, P. L.; Culberson, C.; Shafer, J. A.; Kuo, L. C. Crystal structure at 1.9 Angstrom resolution of human immunodeficiency virus (HIV) protease complexed with L-735,524, an orally bioavailable inhibitor of HIV proteases. *J. Biol. Chem.* **1994**, *269*, 26344−26348.

(26) Beach, M. D.; Chasman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A. Accurate ab initio quantum chemical determination of the relative energetics of peptide conformations and assessment of empirical force fields. *J. Am. Chem. Soc.* **1997**, *119*, 5908−5920.

(27) Prendergast, K.; Adams, K.; Greenlee, W. J.; Nachbar, R. B.; Patchett, A. A.; Underwood, D. J. Derivation of a 3D pharmacophore model for the angiotensin-II site one receptor. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 491−512.

JM9806143