Molecular Similarity Based on DOCK-Generated Fingerprints

Hans Briem*,†,‡ and Irwin D. Kuntz‡

Department of Medicinal Chemistry, Boehringer Ingelheim KG, 55216 Ingelheim, Germany, and Department of Pharmaceutical Chemistry and Molecular Design Institute, University of California, San Francisco, California 94143-0446

Received October 31, 1995⊗

An alternative method for defining molecular similarity is presented. By using the docking program DOCK and a reference panel of protein binding sites, fingerprints for a set of molecules have been generated, based on calculated interaction energies. These binding patterns allowed us to calculate matrices of similarity coefficients which subsequently were used for nearest-neighbor searches within the database. Our results indicate that the method is suitable for finding significant similarities of compounds of the same biological activity. Although the overall performance of a traditional 2D similarity method is better in the test systems investigated, our 3D approach can be regarded as complementary since it is able to detect similarities independent of the covalent structure of the compounds. Thus it should be a useful 3D database-searching tool for rational lead discovery.

Introduction

Methods to describe the similarities of molecules have gained great interest in rational drug discovery in recent years.[1] Based on the concept that compounds binding to the same target macromolecule should bear some similarity to each other, such methods could be suitable for searching large molecular databases keyed on a given lead structure. This should enhance the effort to find new leads.

Beyond database searching, there are other applications of similarity metrics that are of particular interest to industrial pharmaceutical research: (1) clustering of in-house compound databases in order to generate structurally representative sets for starting a new biological assay and (2) comparison of in-house databases with databases from external vendors of compounds, to increase the diversity of a corporate database by "rational" acquisition of compounds.

There are numerous ways to assess the similarity of molecules, depending on the choice of molecular properties to compare.[1] The most popular approaches, which have been available in commercial software packages, like MERLIN (DAYLIGHT Chemical Information Systems Inc., Irvine, CA), MACCS (MDL Information Systems Inc., San Leandro, CA), or SYBYL (Tripos Inc., St. Louis, MO) are based on 2D representations of molecules. For measuring the molecular similarity, these methods use the presence or absence of 2D patterns or fragments. Since binding of a ligand to a target receptor is a 3D event, which involves the surfaces of both interacting molecules, a 3D measure might be more appropriate for defining molecular similarity. Some approaches considering 3D properties, including electrostatic potential,[2] shape descriptors,[3] atom distance matrices,[4] or projections of properties on polyhedra,[5] have been published recently. A common problem of these methods is the need for an appropriate procedure to align the molecules prior to calculating their similarity.

A very attractive idea from a completely different perspective has recently been proposed by Terrapin:[6] They use the experimentally determined in-vitro binding potency of compounds against a reference panel of diverse proteins to obtain a so-called affinity fingerprint. Similarity can then be expressed simply by comparing each pair of fingerprints. Their underlying assumption is that compounds which bind similarly to all the proteins in the reference panel are likely to have similar affinity to their target receptor as well. In contrast to the 2D and 3D property-based methods mentioned above, Terrapin's approach takes the binding interaction of a ligand to its receptor as a basis for the molecular similarity metric. Although empirical in nature, such a view from the receptor's perspective might be more appropriate detecting the biological similarity of molecules.

Our idea, described here, is to replace the in-vitro screening of Terrapin's approach by a computer simulation of the docking process. As an equivalent for the reference panel of proteins, we used known 3D structures of proteins. The docking was accomplished with the DOCK suite of programs.[7] Binding affinities were expressed as DOCK scores, using different scoring schemes. In order to test the method, we used a test set of 3D molecular structures with known biological activities, taken from five different activity classes. Our DOCK-based approach should be able to assess significantly higher similarity among compounds of the same activity class as opposed to compounds of different classes. In addition, we compared our method with results obtained using the DAYLIGHT 2D pattern-based approach.

Methods

Selection of the Reference Panel of Proteins. The binding sites used as a reference panel in this study (Table 1) were selected from the Brookhaven Protein Databank (PDB) according to the following criteria: (1) the sites should be dissimilar to each other in order to achieve appropriate diversity of the DOCK scores. Hence, each binding site was selected from a different structural and functional family of proteins. In addition, the correlation coefficients **r** between each pair of proteins were calculated, based on one of the DOCK score matrices (see below). Since the DOCK scores to some extend reflect the shape of the ligands—at least for medium scoring molecules—we have to expect some correlation between different proteins. But as can be seen in Table 2, the

* Corresponding author.
† Boehringer Ingelheim KG.
‡ University of California, San Francisco.
⊗ Abstract published in *Advance ACS Abstracts,* June 1, 1996.

S0022-2623(95)00800-4 CCC: $12.00

Table 1. Reference Panel of Proteins

PDB entry	description	protein class	resolution Å	ref
3DFR	dihydrofolate reductase complex with NADPH and methotrexate	oxidoreductase	1.70	8
3CLA	chloramphenicol acetyltransferase complex with chloramphenicol	acetyltransferase	1.75	9
1ACL	acetylcholinesterase complex with decamethonium	carboxylic esterase	2.80	10
1DWC	α-thrombin complex with modified hirudin[a] and argatroban	serine protease	3.00	11
1EED	endothiapepsin complex with PD125754	aspartic protease	2.00	12
1POP	papain complex with leupeptin	thiol protease	2.10	13
2TSC	thymidilate synthase complex with dUMP and an anti-folate	methyltransferase	1.97	14
na	model of the 5-HT_2 receptor	GPCR	na	15

[a] (Des-amino Asp 55)hirudin (residues 55−65). na, not applicable.

Table 2. Correlation Matrix for DOCK Scores in Reference Panel of Proteins[a]

3DFR	1.000							
3CLA	0.374	1.000						
1ACL	0.137	−0.045	1.000					
1DWC	0.225	0.359	0.244	1.000				
1EED	−0.168	−0.031	0.488	0.335	1.000			
1POP	0.452	0.411	0.231	0.390	0.159	1.000		
2TSC	0.398	0.144	0.443	0.303	0.158	0.373	1.000	
5HT2	0.139	0.029	0.466	0.264	0.297	0.337	0.548	1.000
	3DFR	3CLA	1ACL	1DWC	1EED	1POP	2TSC	5HT2

[a] Correlation coefficients r (N = 972) between each pair of reference binding sites, calculated from a DOCK score matrix (single conformation, force field scoring).

Table 3. Characteristics of Ligand Dataset

activity class	no. of ligands in database	class of target protein
PAF receptor antagonists	136	GPCR
5-HT_3 receptor antagonists	52	ion channel
thromboxane A2 (TXA2) receptor antagonists	49	GPCR
angiotensin-converting enzyme (ACE) inhibitors	40	metallopeptidase
HMG CoA reductase inhibitors	114	reductase
random set	581	

highest correlation coefficient has a value of 0.548 which represents a shared variance of just 30%. Thus we believe that our choice of reference binding sites is well balanced from a statistical point of view.[15] (2) Only targets that are known to strongly bind "druglike" molecules (e.g., compounds with molecular weights < 600) in a well-defined binding pocket were used. (3) To guide the docking experiments, only protein−inhibitor complexes were utilized. As the only exception to this rule, a model of the 5-HT_2 receptor, a member of the G-protein-coupled receptor (GPCR) superfamily, was included in the study. This should check the dependence of the method on experimentally determined binding site structures as opposed to model structures. The 5-HT_2 receptor model, comprising the seven transmembrane-spanning domains only, was generated by G. Vriend,[16] based on the cryoelectronmicroscopy structure of bacteriorhodopsin.[17] To relieve bad steric contacts, we fixed side chain geometries using the Auto-Rotamer option within the INSIGHT II molecular modeling package (Biosym Technologies Inc., San Diego, CA).

Selection of the Ligand Test Set. All ligand structures were extracted from the MACCS Drug Data Report (MDDR), provided by MDL Information Systems Inc. (San Leandro, CA). Compounds were chosen from five different activity classes (Table 3). The target receptors of these compounds belong to completely different families of proteins, and none of the target structures is as yet determined experimentally. All ligands are known to bind in the nanomolar range to their respective target receptor. To serve as a negative control, an additional set of compounds not belonging to any of the five activity classes was selected randomly from the MDDR database.

Preparation of the DOCK Databases. Single-conformer database: The ligand dataset was extracted as 2D structures. Next, 3D structures were generated utilizing the CONCORD program[18] (version 3.0.1). Hydrogens and partial charges, according to the Gasteiger−Marsili method,[19] were added within the SYBYL molecular modeling package (Tripos Inc., St. Louis, MO).

Ten-conformer database: Starting from the CONCORD-generated conformer described above, for each compound 10 conformations were generated using a simulated annealing molecular dynamics protocol described elsewhere[20] utilizing the DISCOVER program (Biosym Technologies Inc., San Diego, CA) and the CVFF force field.[21]

DOCK. Version 3.5 of the DOCK suite of programs was used throughout this study. Details of the DOCK algorithm are described in numerous papers[7] and are not repeated here. In order to prepare the binding pocket for each protein in the reference panel prior to docking the ligands, the following procedure was applied. All crystallographic water molecules and bound inhibitors were removed. For each inhibitor binding pocket a Connolly surface was generated using the MS program.[22] To generate a "negative image" of the site, spheres with a maximal sphere radius of 5 Å were generated by the SPHGEN module of DOCK. Consecutively, each ligand was docked into the binding pockets of every reference protein.

Two different grid-based methods to estimate the ligand−receptor interactions were applied as follows: contact scoring, which simply sums up favorable and unfavorable contacts of ligand atoms with the receptor, regardless of their chemical nature, and force field scoring, which calculates interaction energies utilizing van der Waals and Coulombic terms. For the 10-conformer database, only the best scoring conformation of each compound was used for fingerprint generation.

Generation of DOCK-Based Fingerprints and Similarity Indices. The resulting scores of the DOCK calculations were saved in four different scoring matrices: (1) single conformation/force field scoring, (2) best of 10 conformations/force field scoring, (3) single conformation/contact scoring, and (4) Best of 10 conformations/contact scoring. Each matrix consists of rows representing each ligand and columns representing the protein binding sites of the reference panel. The DOCK scores make up the matrix elements. Consequently, each row can be regarded as a fingerprint of the respective ligand.

There are many ways to obtain a similarity index based on a fingerprint of properties. Fragment-based molecular similarity approaches, such as those developed by DAYLIGHT and MDL, commonly use the Tanimoto coefficient to compare two bitstrings:

$$T_c = B_c/(B_1 + B_2 - B_c) \quad (1)$$

where T_c is the resulting Tanimoto similarity coefficient of two molecules, B_c are the bits in common between two binary fingerprints, and B_1 and B_2 are the bits set in fingerprints 1 and 2, respectively. A problem with this method is that it requires a binary representation of the fingerprint and thus

Table 4. Score Difference Distributions

method	difference			standard deviation	threshold used in eq 3
	max	min	mean		
one conformation, force field scoring	62.3	−60.4	0.4	9.5	19.0
10 conformations, force field scoring	52.2	−69.9	−1.2	8.6	17.2
one conformation, contact scoring	260.0	−258.0	2.5	33.8	67.6
10 conformations, contact scoring	155.0	−164.0	0.3	27.2	54.4

was not immediately applicable to our nonbinary fingerprint of DOCK score values.

Thus we used a slightly modified version that nevertheless lies at the heart of the original Tanimoto method.[23] The term B_c in (1) can be substituted by

$$\sum_{i=1,n}\{\text{threshold} - (|S_{A,i} - S_{B,i}|)\} \quad (2)$$

where $|S_{A,i} - S_{B,i}|$ is the difference of the DOCK scores of two molecules A and B at target i, n is the total number of targets, and threshold is a user-defined value up to which score differences should be taken into account for the similarity calculation. For this study, threshold was set to 2 times the standard deviation of the score difference distribution within one matrix (Table 4). For $|S_{A,i} - S_{B,i}|$ greater than threshold, the term $\{\text{threshold} - (|S_{A,i} - S_{B,i}|)\}$ was set to zero.

Furthermore B_1 and B_2 in eq 1 can be replaced by (n*threshold), respectively, and as n, the number of targets, is the same for both fingerprints, our similarity coefficient is given as

$$S_c = \sum_{i=1,n}\{\text{threshold} - (|S_{A,i} - S_{B,i}|)\}/(2(\text{n*threshold}) - \sum_{i=1,n}\{\text{threshold} - (|S_{A,i} - S_{B,i}|)\}) \quad (3)$$

Applying this equation to each of the scoring matrices, similarity matrices of the size $N(N-1)/2$ were calculated, with N being the number of molecules in the database.

Generation of DAYLIGHT-Based Fingerprints and Similarity Indices. DAYLIGHT uses a fingerprinting algorithm that is based on hashing every possible bond path (up to a certain length limit) through a molecule into a string of bits.[24] The similarity coefficient of two different molecules is calculated by the standard Tanimoto formalism as described in eq 1 above. We used the DAYLIGHT software to generate a similarity matrix for each combination of molecules in our database.

Computing Enrichment Factors for Each Activity Class. Since the absolute values of the similarity coefficients calculated with our DOCK-based method cannot be directly compared to those obtained by DAYLIGHT, we use enrichment factors as a measure of success. For a given activity class, these factors describe the enrichment of compounds of the same action in a nearest-neighbor list.

The following procedure was used to compute enrichment factors: Each ligand in turn was taken as a template for a similarity search. All the other compounds in the database were ranked based upon their similarity with the template ("nearest-neighbor list"). For a given percentage of top scores, the number of compounds which belong to the template's activity class was determined.

The enrichment factor E of a particular compound of activity A for a given percentage p was computed according to the following equation:

$E(p)$ = (no. of hits of activity A/total no. of hits)/((no. of compounds of activity A in database − 1)/(total no. of compounds in database − 1)) (4)

Thus, a compound whose activity class in its nearest-neighbor list is represented by the same portion than in the whole database has an enrichment factor of 1 (i.e., no enrichment at all).

Furthermore, the mean enrichment factors $\bar{E}(p)$ for each activity class at various percent levels were calculated, with n being the total number of compounds in the activity class:

$$\bar{E}(p) = \sum_{i=1,n} E_i(p)/n \quad (5)$$

***k* Nearest-Neighbor (*k*NN) Analysis.** As an alternative method to gauge the success of our approach, we classified our dataset by a well-established nearest-neighbor (nn) approach.[25] Basically, the membership of a compound to a particular activity class is predicted based on the activity class(es) of its k nearest neighbor(s). The analysis involved the following steps: (1) In order to determine the optimal values of k for our given similarity matrices, we first randomly split the five activity classes in half into a training set and a test set. (2) The values for k were varied from 1 to 10, e.g., for $k = 1$ only the most similar compound was used for the prediction. (3) The activity class of each member of the training set was subsequently predicted taking the whole Tanimoto matrix into account. It should be pointed out that the group of compounds randomly taken from the MDDR database was not included in the training set but used in the prediction of the test set, again to serve as a negative control. (4) In cases of equal numbers of nearest neighbors (e.g., a compound has three ACE inhibitors and three PAF antagonists in its nn list), the class with the higher mean Tanimoto coefficient of the nn's was used for the prediction. (5) To cope with the different random expectations (chance predictions) due to the uneven number of members within the activity classes, we judged the success rate for each particular activity class by a prediction factor P, pretty much ressembling the calculation of the enrichment factors in eq 4:

P = (no. of correctly predicted compounds of activity class/total no. of predicted compounds of activity class)/((no. of compounds of activity class in database − 1)/(total no. of compounds in database − 1)) (6)

(6) A mean prediction factor $\bar{P}$ was calculated for each value of k, with m being the number of activity classes:

$$\bar{P} = \sum_{i=1,m} P_i/m \quad (7)$$

(7) The best predictive performance for the training set and thus the optimal value for k was determined by the highest mean prediction factor $\bar{P}_{\text{training}}$. This value of k was then applied for prediction of the test set ($\bar{P}_{\text{test}}$) and the total set ($\bar{P}_{\text{total}}$).

Root Mean Square Difference (rmsd) Distribution. In order to inspect the relative physical orientation of similar compounds in the different binding sites of the reference panel, we first took each compound in the database in turn as a template and generated the respective top 1% hit lists, based on the similarity indices. Subsequently, the template or "mother" structures as well as their corresponding hits or "children" were extracted from the eight different binding sites. To obtain the rmsd spatial distribution of the "children" relative to their "mothers", the templates were superimposed onto each other while retaining the relative orientations of the "children". Then, for each combination of children and binding sites, the rmsd was calculated.

Results and Discussion

Enrichment Factors. The results obtained using eqs 4 and 5 for the different similarity matrices are

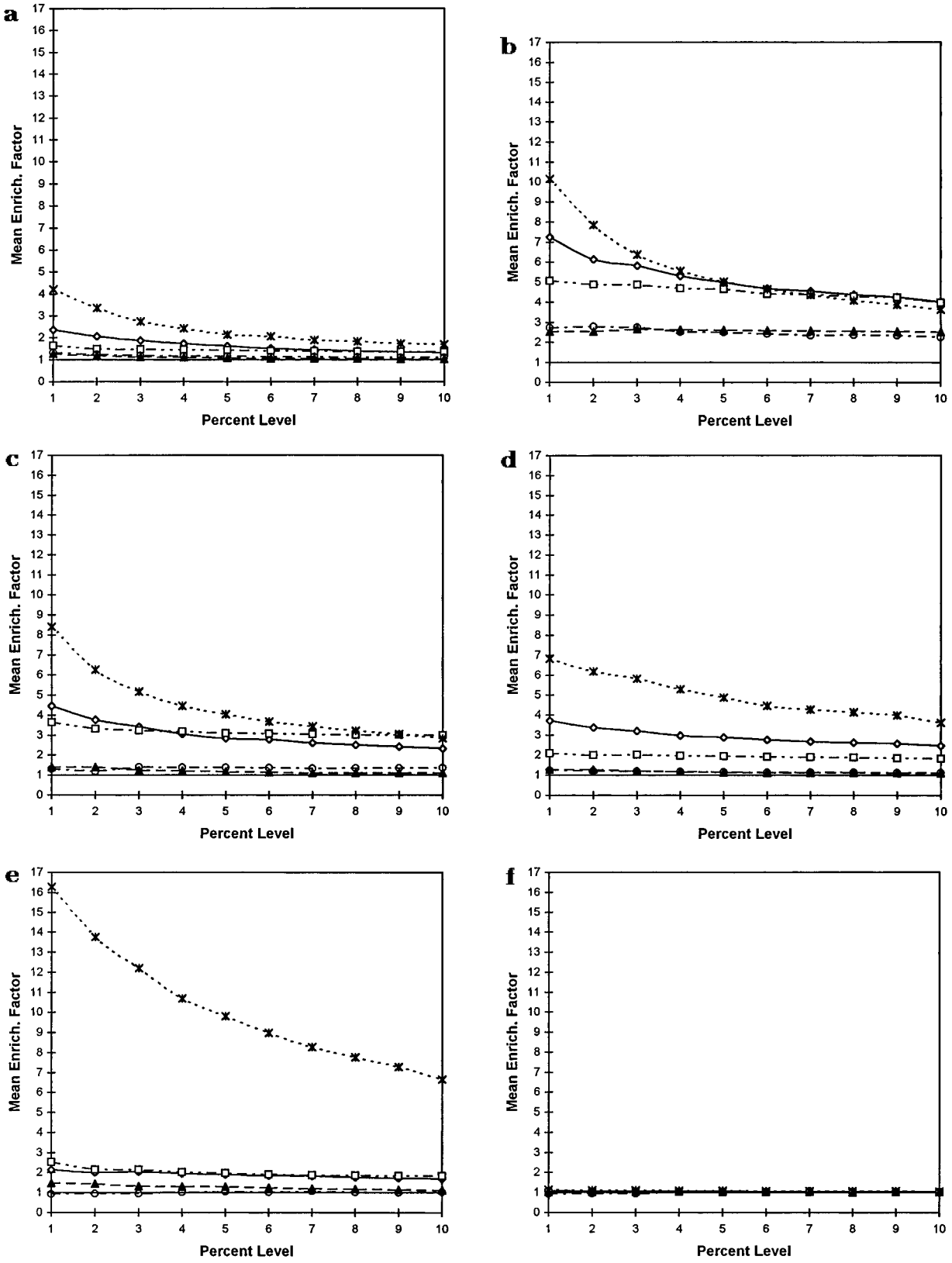

Figure 1. Mean enrichment factors obtained from different DOCK scoring matrices and DAYLIGHT: (a) PAF, (b) 5-HT_3, (c) TXA2, (d) HMG-CoA, (e) ACE, and (f) random. Symbols: (◇) DOCK, one conformation, force field scoring, (□) DOCK, 10 conformations, force field scoring, (▲) DOCK, one conformation, contact scoring, (○) DOCK, 10 conformations, contact scoring, and (*) DAYLIGHT.

shown in Figure 1 and Table 5. From these findings several conclusions may be drawn as follows: (1) all methods applied lead to a mean enrichment of compounds of the same activity class, since all enrichment factors for the top percent levels are greater than one. In other words, our DOCK-based approach is clearly able to cluster compounds having the same biological activity. However, there is great variation in the degree of enrichment, both between different activity classes as well as between different calculation methods. (2) In most instances, the rank order is DAYLIGHT > DOCK, force field scoring, one conformer > DOCK, force field scoring, best of 10 conformers > DOCK, contact scoring, one conformer = DOCK, contact scoring, best

Table 5. Mean Enrichment Factors

percent level (*p*)	method	PAF	5-HT_3	TXA2	HMG-CoA	ACE	random
1	DOCK, 1 conf,[a] ffs[b]	2.4 (2.2)[c]	7.3 (5.0)	4.5 (4.0)	3.7 (2.2)	2.1 (3.1)	1.1 (2.7)
	DOCK, 10 conf, ffs	1.6 (1.1)	5.1 (3.8)	3.7 (2.8)	2.1 (1.1)	2.5 (2.9)	1.1 (2.1)
	DOCK, 1 conf, cs[d]	1.3 (0.9)	2.5 (2.7)	1.4 (1.6)	1.3 (1.2)	1.5 (2.4)	1.0 (0.4)
	DOCK, 10 conf, cs	1.3 (0.8)	2.7 (2.6)	1.3 (1.3)	1.3 (0.9)	0.9 (1.4)	1.1 (0.7)
	DAYLIGHT	4.2 (4.7)	10.2 (5.0)	8.4 (4.9)	6.8 (2.8)	16.3 (9.8)	1.1 (7.3)
2	DOCK, 1 conf, ffs	2.1 (1.9)	6.1 (3.9)	3.8 (3.1)	3.4 (1.7)	2.0 (2.5)	1.1 (2.3)
	DOCK, 10 conf, ffs	1.5 (1.0)	4.9 (3.6)	3.3 (2.2)	2.0 (1.0)	2.1 (2.1)	1.0 (1.9)
	DOCK, 1 conf, cs	1.2 (0.7)	2.6 (2.3)	1.4 (1.3)	1.3 (0.9)	1.4 (1.4)	1.0 (0.4)
	DOCK, 10 conf, cs	1.2 (0.7)	2.8 (2.4)	1.2 (1.1)	1.2 (0.7)	1.0 (0.9)	1.0 (0.5)
	DAYLIGHT	3.4 (3.6)	7.9 (3.7)	6.3 (3.5)	6.2 (2.2)	13.8 (8.7)	1.1 (5.2)

[a] Conformations. [b] Force field scoring. [c] The numbers in parentheses represent the standard deviations. [d] Contact scoring.

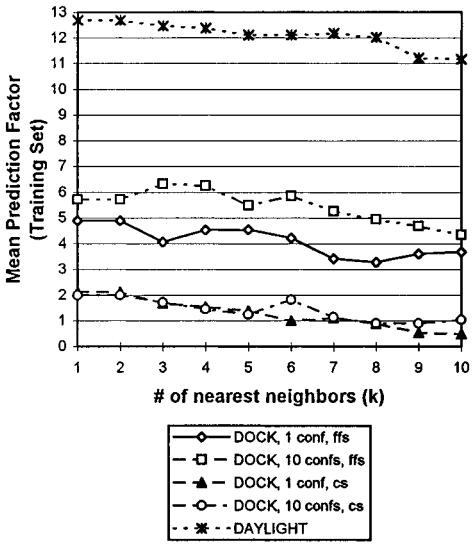

Figure 2. Determination of optimal values for *k*.

of 10 conformers. This observation will be discussed in more detail below. (3) For the random set, as anticipated, none of the methods yields any enrichment. (4) As can be deduced from the rather high standard deviations of the mean enrichment factors, there is a great fluctuation in the individual enrichment factors. Thus, for a given template (or query) structure, there is no guarantee that one can find all compounds of the same activity class in the hit list. Nevertheless the overall chance is clearly higher than random.

***k* Nearest-Neighbor Analysis.** The determination of an optimal value for *k* is shown graphically in Figure 2 by plotting the mean prediction factors $\bar{P}_{\text{training}}$ of the training sets against *k*. An optimal value of $k = 1$ was found for all similarity matrices with the exception of the DOCK score matrix with force field scoring and best of 10 conformers, with $k = 3$ yielding the best predicitions. These values were subsequently used for the predictions of the test sets.

Table 6 summarizes the final results of the *k* nearest-neighbor analysis in detail for the two best performing methods. In addition Figure 3 shows the mean prediction factors $\bar{P}_{\text{total}}$, determined over the whole dataset, for all similarity matrices under investigation.

The rank order of performance obtained by the *k*nn analysis is very similar to the rank order which resulted from calculation of the enrichment factors (DAYLIGHT > DOCK scores with force field scoring and > DOCK scores with contact scoring). It is obvious that our DOCK-based method has the potential to predict the class membership of compounds far better than random, though worse than DAYLIGHT's 2D approach.

DOCK vs DAYLIGHT. As the DAYLIGHT method in all instances performs better than any of our DOCK-based methods, we wanted to gain more insight into which compounds are actually found by each method. We therefore compared the top 1% hit lists from the DAYLIGHT and the DOCK, force field scoring, one-conformation method.

As can be seen in Figure 4, there is only little overlap in these hit lists. On average, about 30% of the DOCK hits (compounds of the same activity class as the template compound) overlap with the respective DAYLIGHT hit lists. Although this is significantly higher than chance probability, one should expect a much higher degree of overlap considering two methods both seeking to find molecular similarity. Thus we believe that our DOCK-based method is complementary to the DAYLIGHT approach since the largest portions of the hit lists are in fact different. Some examples should illustrate this finding (Figure 5).

The difference in the methods is easily understood. Being a fragment-based approach, the DAYLIGHT method is more suited to find common core *substructures* (e.g., the benzodiazepine moiety in many PAF antagonists). Considering the selection criteria for the test set, comprising only compounds already optimized for their respective target, there are clearly only a few such core structures present for each activity class. Thus, given such an ideal case for a fragment-based approach, it is not surprising that DAYLIGHT is superior in overall performance. On the other hand, the merit of our method is finding similarities based on *shape* and *electrostatic complementarity*. Both of these molecular properties seem to be important for obtaining high enrichment factors and good predictive performance in the *k*nn analysis, since the force field-scoring scheme yields generally much better results than pure contact scoring.

Flexible vs Rigid Docking. A puzzling result is the fact that the "best scoring of 10 conformations" method in most instances performs more or less equal to the "one-conformation" method. Taking Terrapin's in-vitro results into account, flexible docking should, in principle, lead to more superior results than rigid docking. Nontheless, there are several possible explanations for our findings:

Ten conformations per compound in most cases is by far not enough to cover the conformational space of a molecule adequately. More advanced methods of flexible docking might therefore lead to better results. The importance of molecular flexibility can be easily deduced from our results: Whereas the 5-HT_3 antagonists, as the most rigid class of compounds in our test set, yield the best results in both analysis methods, the highly flexible ACE inhibitors perform very poorly. Interestingly, the ACE group is the only case where the "best

Table 6. Nearest-Neighbor Classification of Different Similarity Matrices

		training set		test set	
	% random expectation	% correct (no. correct/no. total)	prediction factor (P)	% correct (no. correct/no. total)	prediction factor (P)
		DOCK, 10 Conformations, Force Field Scoring, $k = 3$[a]			
PAF	13.99	33.82 (23/68)	2.42	26.47 (18/68)	1.89
5-HT_3	5.35	46.15 (12/26)	8.63	61.54 (16/26)	11.50
TXA2	5.04	52.00 (13/25)	10.32	45.83 (11/24)	9.09
HMG-CoA	11.73	35.09 (20/57)	2.99	47.37 (27/57)	4.04
ACE	4.12	30.00 (6/20)	7.29	0.00 (0/20)	0.00
random	59.80			79.69 (463/582)	1.33
		DAYLIGHT, $k = 1$[b]			
PAF	13.99	76.47 (52/68)	5.47	82.35 (56/68)	5.89
5-HT_3	5.35	92.31 (24/26)	17.25	65.38 (17/26)	12.22
TXA2	5.04	68.00 (17/25)	13.49	62.50 (15/24)	12.40
HMG-CoA	11.73	91.23 (52/57)	7.78	85.96 (49/57)	7.33
ACE	4.12	80.0 (16/20)	19.44	75.00 (15/20)	18.26
random	59.80	73.49 (427/581)	1.23		

[a] Mean prediction factor $\bar{P}_{total}$ for total set (without random set): 5.82. [b] Mean prediction factor $\bar{P}_{total}$ for total set (without random set): 11.95.

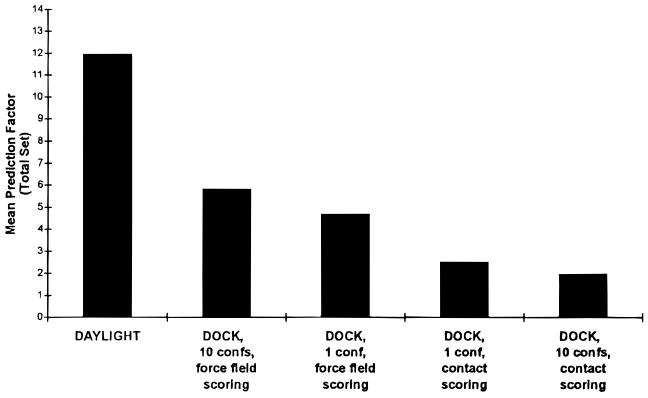

Figure 3. Mean prediction factors for whole datasets.

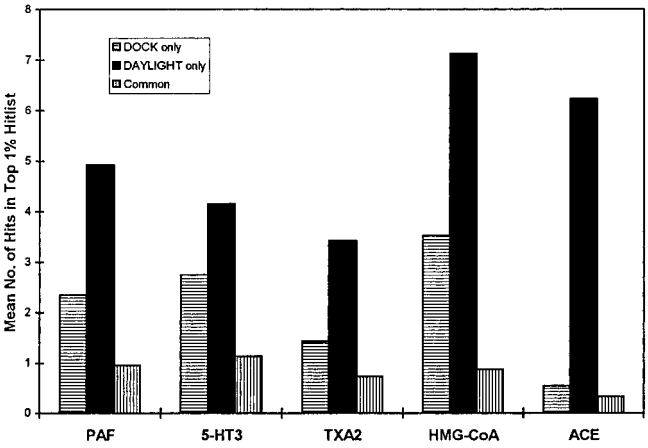

Figure 4. Overlap of top 1% hit lists from DAYLIGHT vs DOCK, single conformer, force field scoring.

of 10 conformations" approach gives slightly better results than the "one-conformation" calculation.

As pointed out above, both shape and electrostatic contributions seem to be of importance in obtaining good results. On the other hand, the shape contribution particularly can vary greatly from one conformation to the other. Since we have used the highest total score of each compound and for each different target, the shape complementarity of compounds of the same activity class might have been lost.

Multiple Binding Modes. Another finding associated with our DOCK-based method is the fact that the relative binding orientation of two molecules with a high similarity index is by no means constant over the reference panel of binding sites. This can be deduced from the rmsd distribution of compounds of the same activity class and belonging to the top 1% hit list (Figure 6). Consequently, compounds with similar DOCK scores throughout the reference panel might gain their similarity by interacting differently with the respective binding pockets. In our opinion, this behavior can be explained by the fact that in many instances we deal with only weak to medium strengths of binding interactions which can be translated to "unspecific" binding in a "real world" setting. We assume that Terrapin's experimental approach has to face the same problem as they deal with a broad range of binding affinities as well. On the other hand, there is more and more evidence from X-ray crystallography data that even compounds with very similar structures and high affinity to their target receptor can bind in totally different relative orientations.[26] Our method seems to reflect this multiple binding mode phenomenon.

Recently it became obvious for some target proteins (e.g., the neurokinin receptors[27]) that there is only very little overlap in the binding sites of competitive, high-affinity antagonists and the respective agonists. In these cases our method might fail to find agonists starting from antagonists and vice-versa, since the DOCK approach relies on a substantial overlap of the binding pockets.

Conclusions and Outlook

By using DOCK scores for a set of molecules docked into a variety of protein binding sites, we can deduce molecular similarity indices which allow us to discriminate between compounds of different activity classes. Furthermore, as opposed to a commonly used 2D approach, we found similarities among compounds of the same activity class possessing completely different chemical scaffolds, yet we see our method as complementary to standard approaches. Since we have only used high-affinity compounds in our test set, the ability to find novel medium-affinity leads in a large compound database has to be demonstrated on new targets.

Opposed to Terrapin's in-vitro study, we have not yet made an attempt to quantitatively predict ligand binding affinities based on the DOCK fingerprints. Encouraged by our results, we will address this ultimate goal in the future on the basis of a "real life" corporate

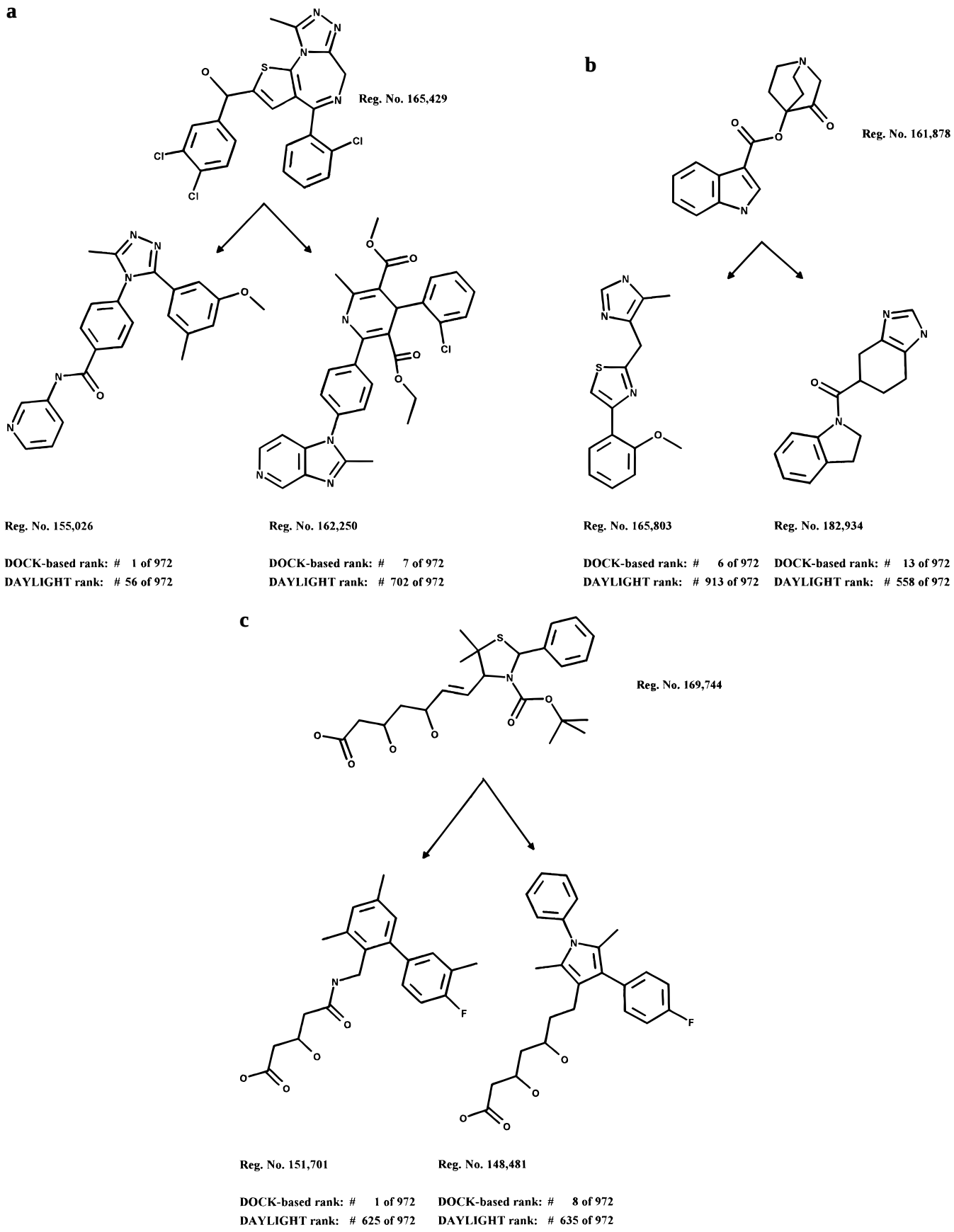

Figure 5. Examples of hits found with the DOCK-based approach (single conformer, force field scoring), which were missed by DAYLIGHT. The top molecules were used as templates for similarity searches, while the bottom molecules show some respective hits. The registry numbers represent the compound numbers internally used in the MDDR database: (a) PAF antagonists, (b) 5-HT_3 antagonists, and (c) HMG-CoA inhibitors.

database. This would allow us to use other tools, like stepwise linear regression, rather than the classification schemes utilized in this study.

The predictive performance of our method clearly relies on the ability of DOCK to simulate the binding event of a ligand to a target protein. We have already

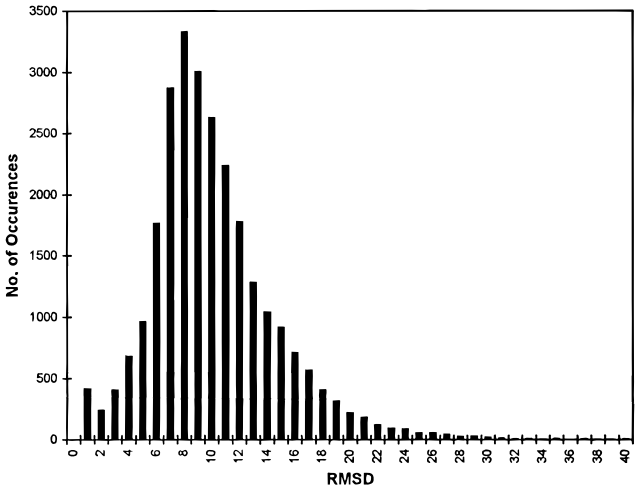

Figure 6. rmsd distribution of relative orientations of similar compounds, taken from the top 1% hit list of the DOCK-based, single-conformation, force field-scoring method.

demonstrated that the force field-scoring method performs much better than the simpler contact scoring, yet considering only van der Waals and Coulombic terms for molecular interactions is surely suboptimal. Thus more elaborate scoring schemes (e.g., based on empirical or quantum mechanical parameters) in conjunction with flexible docking should have the potential to further enhance the results.

Acknowledgment. We wish to thank Peter Brozek, Herbert Köppen, Hubertus Peil, Daniel Gschwend, Connie Oshiro, and Paul McCloskey for their help in various stages of this project. The development of the DOCK programs was supported by NIH.

References

(1) (a) Johnson, M. A., Maggiora, G. M., Eds. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990. (b) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, 1987. Both books contain a number of references to other articles utilizing molecular similarity techniques.

(2) Hodgkin, E. E.; Richards, W. G. Molecular Similarity Based on Electrostatic Potential and Electric Field. *Int. J. Quantum Chem. Quantum Biol. Symp.* **1987**, *14*, 105−110.

(3) (a) Meyer, A. Y.; Richards, W. G. Similarity of Molecular Shape. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 427−439. (b) Good, A. C.; Ewing, T. J. A.; Gschwend, D. A.; Kuntz, I. D. New Molecular Shape Descriptors: Application in Database Screening. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 1−12.

(4) Bemis, G. W.; Kuntz, I. D. A fast and efficient method for 2D and 3D molecular shape description. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 607−628.

(5) Perry, N. C.; Van Geerestein, V. J. Databases Searching on the Basis of Three-Dimensional Molecular Similarity Using the SPERM Program. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 607−616.

(6) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, A.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* **1995**, *2*, 107−118.

(7) (a) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule-Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269−288. (b) Shoichet, B. K.; Bodian, D. L.; Kuntz, I. D. Molecular Docking Using Shape Descriptors. *J. Comput. Chem.* **1992**, *13*, 380−397. (c) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated Docking with Grid-Based Energy Evaluation. *J. Comput. Chem.* **1992**, *13*, 505−524.

(8) Filman, D. J.; Bolin, J. T.; Matthews, D. A.; Kraut, J. Crystal Structure of *Escherichia coli* and *Lactobacillus casei* Dihydrofolate Reductase Refined at 1.7 Å Resolution. *J. Biol. Chem.* **1982**, *257*, 13663−13672.

(9) Leslie, A. G. W. Refined Crystal Structure of Type III Chloramphenicol Acetyltransferase at 1.75 A Resolution. *J. Mol. Biol.* **1990**, *213*, 167−186.

(10) Sussmann, J. L.; Harel, M.; Frolow, F.; Oefner, C.; Goldman, A.; Toker, L.; Silman, I. Atomic Structure of Acetylcholinesterase from *Torpedo californica*: A Prototypic Acetylcholin-Binding Protein. *Science* **1991**, *253*, 872−879.

(11) Banner, D. W.; Hadvary, P. Crystallographic Analysis at 3.0-Å Resolution of the Binding to Human Thrombin of 4 Active Site-Directed Inhibitors. *J. Biol. Chem.* **1991**, *266*, 20085−20093.

(12) Cooper, J. B.; Quail, W.; Frazao, C.; Foundling, S. I.; Blundell, T. L.; Humblet, C.; Lunney, E. A.; Lowther, W. T.; Dunn, B. M. X-Ray Crystallographic Analysis of Inhibition of Endothiapepsin by Cyclohexyl Renin Inhibitors. *Biochemistry* **1992**, *31*, 8142−8150.

(13) Kamphuis, I. G.; Kalk, K. H.; Swarte, M. B. A.; Drenth, J. Structure of Papain Refined at 1.65 Å Resolution. *J. Mol. Biol.* **1984**, *179*, 233−256.

(14) Montfort, W. R.; Perry, K. M.; Fauman, E. B.; Finer-Moore, J. S.; Maley, G. F.; Hardy, L.; Maley, F.; Stroud, R. M. Structure, Multiple Site Binding, and Segmental Accommodation in Thymidylate Synthase on Binding dUMP and an Anti-Folate. *Biochemistry* **1990**, *29*, 6964−6977.

(15) We discarded one of our first choices, the FK506 binding protein (1FKF), as it yielded correlation coefficients of $r > 0.7$ to 3CLA and 1EED.

(16) (a) Vriend, G. Molecular Modeling of GPCRs. *TM7* **1994**, *3*, 1−10. (b) Oliveira, L.; Paiva, A. C. M.; Vriend, G. A common motif in G-protein coupled seven transmembrane helix receptors. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 649−658. The structure of the 5-HT_2 receptor model has been obtained via anonymous ftp from swift.embl-heidelberg.de.

(17) Henderson, R.; Baldwin, J.; Ceska, T. H.; Zemlin, F.; Beckmann, E.; Downing, K. Model of the structure of bacteriorhodopsin based on high resolution electron cryo-microscopy. *J. Mol. Biol.* **1990**, *213*, 899−929.

(18) (a) Rusinko, A., III; Skell, J. M.; Balducci, R.; McGarity, C. M.; Pearlman, R. S. *Concord, A Program for the Rapid Generation of High Quality Approximate 3-Dimensional Molecular Structures*; The University of Texas at Austin and Tripos Associates: St. Louis, MO, 1988. (b) Pearlman, R. S. Rapid Generation of High Quality Approximate 3D Molecular Structures. *Chem. Des. Aut. News* **1987**, *2*, 1−6.

(19) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(20) Esser, F.; Carpy, A.; Briem, H.; Köppen, H.; Pook, K.-H. Synthesis of a new dipeptide template, its X-ray structure, and modeling studies on its conformational features. *Int. J. Pept. Protein Res.* **1995**, *45*, 540−546.

(21) Hagler, A. T.; Huler, E.; Lifson, S. Energy Functions for Peptides and Proteins. I. Derivation of a Consistent Force Field Including the Hydrogen Bond from Amide Crystals. *J. Am. Chem. Soc.* **1974**, *96*, 5319−5327.

(22) (a) Connolly, M. L. Analytical Molecular Surface Calculation. *J. Appl. Crystallogr.* **1983**, *16*, 548−558. (b) Connolly, M. L. Solvent-accessible surfaces of proteins and Nucleic Acids. *Science* **1983**, *221*, 709−713.

(23) We also evaluated the more familiar Euclidian distances and obtained virtually the same results in the nearest-neighbor analyses.

(24) DAYLIGHT Software Manual, release 4.34, DAYLIGHT Inc.,-1994.

(25) Livingstone, D. *Data Analysis for Chemists*; Oxford University Press: Oxford, 1995.

(26) (a) Ringe, D. Binding by design. *Nature* **1991**, *351*, 185−186. (b) Klebe, G.; Abraham, U. On the Prediction of Binding Properties of Drug Molecules by Comparative Molecular Field Analysis. *J. Med. Chem.* **1993**, *36*, 70−80.

(27) (a) Gether, U.; Johansen, T. E.; Snider, R. M.; Lowe, J. A., III; Nakanishi, S.; Schwartz, T. W. Different binding epitopes on the NK_1 receptor for substance P and a non-peptide antagonist. *Nature* **1993**, *362*, 345−348. (b) Huang, R.-R. C.; Yu, H.; Strader, C. D.; Fong, T. M. Localization of the Ligand Binding Site of the Neurokinin-1 Receptor: Interpretation of Chimeric Mutations and Single-Residue Substitutions. *Mol. Pharmacol.* **1994**, *45*, 690−695.

JM950800Y