Training a Scoring Function for the Alignment of Small Molecules

Shek Ling Chan* and Paul Labute

Chemical Computing Group, Inc., 1010 Sherbrooke Street West, Suite 910, Montreal, H3A 2R7, Canada

Received June 11, 2010

A comprehensive data set of aligned ligands with highly similar binding pockets from the Protein Data Bank has been built. Based on this data set, a scoring function for recognizing good alignment poses for small molecules has been developed. This function is based on atoms and hydrogen-bond projected features. The concept is simply that atoms and features of a similar type (hydrogen-bond acceptors/donors and hydrophobic) tend to occupy the same space in a binding pocket and atoms of incompatible types often tend to avoid the same space. Comparison with some recently published results of small molecule alignments shows that the current scoring function can lead to performance better than those of several existing methods.

INTRODUCTION

In drug discovery projects, the three-dimensional (3D) structure of the receptor is not always available. In such cases, ligand optimization often depends on a meaningful alignment of the active compounds. Pharmacophore model building, 3D quantitative structure−activity relationship (QSAR), comparative molecular field analysis (CoMFA),[1] and ligand-based virtual screening all depend on a good algorithm to flexibly align small molecules. In light of the importance of the subject, many methods have been developed to perform the task. On top of the methods mentioned in the comprehensive review of Lemmen and Lengauer,[2] new efforts have continued to appear.[3−16]

Similar to docking, the small molecule alignment problem can be divided into two parts. First, the conformational/pose space has to be thoroughly searched. Second, a scoring function needs to be able to distinguish good alignment poses from other possibilities. The current work is about obtaining a good scoring function for this purpose.

In principle, a scoring function that can distinguish the correct alignment pose from incorrect ones is not necessarily the same as a scoring function that upon minimization gives poses that resemble the correct one as much as possible. In practice, a good scoring function probably performs well for both purposes. In this work, we are primarily interested in a scoring function for distinguishing good poses. In any case, we believe that the *exact* answer to the molecular alignment problem may not be simple. It depends on how the crystal structures are superposed. For example, using only α carbons in the binding pocket for superposition may give a slightly different answer from that using all the pocket atoms or using the pharmacophore elements involving the ligand and the pocket. And then there is the resolution of the crystal structure itself. Moreover, the ligand may have some freedom of movement inside a pocket. Hence trying to nudge down the geometric difference between a proposed alignment hypothesis and the "correct" alignment all the way to zero may not be necessary, or even meaningful, as long as the correct alignment mode is obtained.

For scoring functions used in docking, one can use molecular mechanics force fields derived from first principles. The same approach is not feasible for scoring functions for small molecule alignment. Such functions need to be derived from some statistical analysis of a training set of known molecular alignments. The Protein Data Bank[17] (PDB) is an obvious source to obtain such a training set. As technology progresses, the number of entries in the PDB has grown exponentially. Just in the five years since 2003, the number of entries has more than doubled (see http://www.pdb.org). Moreover, many old entries have been revisited and cleaned up, and many low-resolution structures have been superseded by better ones. By now, there are hundreds of systems of entries with identical proteins but different ligands. We believe that now is a good time to make use of the growth in the PDB to revisit the derivation of a scoring function for molecular alignment.

METHODS

Scoring Function. For the problem of small molecule alignment, there are two main types of scoring functions. The first type is atom based: When two molecules are being aligned, the score consists of a sum of terms that are based on intermolecular atom pairs (i.e., each pair has one atom coming from each molecule). The second type of score is field based: The electrostatic or steric fields of the molecules or their surfaces are compared to arrive at a score. This would be somewhat slower to compute since the fields or the molecular surfaces need to be calculated from the atomic coordinates. Hence it would be difficult to flexibly refine an ensemble of aligned molecules or to deal with a huge number of conformers. Indeed several of the field-based alignment methods rely on an independent conformational search engine to generate good conformers (e.g., Shapelets,[6] BRUTUS,[13] the "Molecular Field Extrema" method,[11] and MIMIC[18]). Since our scoring function will be used to distinguish good binding poses among a huge number of possibilities, it needs to be calculated quickly. An atom-based function would hence be more appealing than a molecular field-based function.

* Corresponding author. E-mail: slchan3@yahoo.com.

10.1021/ci100227h
Published on Web 09/10/2010

When creating a hypothesis for an alignment of molecules, it is desirable for atoms (from different molecules that are being aligned) with similar properties to be near to each other. Intermolecular atom pairs that are far away should have little influence on the goodness of an alignment. This can be achieved by associating each intermolecular atom pair with similar properties with a Gaussian scoring term: $w \exp(-\alpha r^2)$, where r is the distance between the two atoms, α is a parameter controlling the range of the interaction, and w is the weight of this term. It is also desirable for the alignment hypothesis for atoms with incompatible properties (e.g., hydrophobic vs hydrophilic) to not superpose. This can be achieved using the same functional form but with a negative value for w. Besides atoms, hydrogen-bond projected features can also be considered. These features represent the expected positions of the hydrogen-bonding partner on the binding pocket based on the ligand geometry. The overlap of hydrogen-bond projected features of the same type should be favorable for an alignment hypothesis. Again, the same Gaussian score can work, with r being the distance between the intermolecular features.

While the Gaussian function is widely used in molecular alignment methods (e.g., FAP,[4] Pharao,[5] MOE,[16] SQ,[19] and SEAL[20]), there are other functional forms that can provide a similar effect. These include piecewise linear functions (e.g., FLAME),[12] and piecewise linear and quadratic functions (e.g., TFIT).[21] It can be construed that indeed any bell-shaped function can work in a similar fashion. In their pioneering work 20 years ago, Kearsley and Smith[20] noticed that the Lorentzian function ($1/(1 + \alpha r^2)$) yielded similar results to the Gaussian but was significantly faster to compute. However, the advance in computational technology to calculate exponents has eliminated this advantage of the Lorentzian function. Simple tests showed that the calculation speeds are now similar. For this work we chose the Gaussian because it involves a minimal number of tunable parameters and has a simple, continuous shape. Another consideration is that function optimization procedures often require the first and second derivatives of the function. In the case of the Gaussian, once the function has been calculated, its first and second derivatives can be obtained without much extra work.

A scoring function S could thus be constructed based on a sum of Gaussian terms:

$$S = \sum_k \mathrm{T}_k = \sum_k w_k \sum_{i \in \mathrm{A}_k} \sum_{j \in \mathrm{B}_k} \exp(-\alpha r_{ij}^2)$$

In the formula, T_k is a term encouraging/discouraging the overlap of a particular pair of types (A_k, B_k) of atoms/features; w_k is the weight of the term; i and j are atoms/features from the two molecules; r_{ij} is the distance between the intermolecular atom/feature pair; and α is a parameter controlling the range of this type of interaction. Details of the various terms, T_k's, are given in Table 1. Ligands binding in the same mode are expected to have substantial volume overlap. The first term in the scoring function, T_0, rewards overlap between the heavy atoms of the different ligands. Next, hydrogen bonding is an important interaction determining ligand binding. The terms T_1 and T_2 encourage the overlap of pairs of intermolecular donors and acceptors. Following this is the hydrophobic term T_3 rewarding the overlap of hydrophobic atoms. After these "attractive" terms come the "repulsive" terms that discourage the overlap of intermolecular atom pairs of incompatible types. First, hydrophobic and hydrophilic atoms (atoms capable of hydrogen bonding) are not expected to occupy the same part of the binding pocket, as reflected by T_4 and T_5. Second, hydrogen-bond acceptors that are not also donors are expected to avoid positions of donors that are not also acceptors. This is reflected by terms T_6 and T_7. The last two terms, T_8 and T_9, are based on hydrogen-bond projected features. They encourage the overlap of hydrogen-bond projected features of the same type.

Table 1. Details of the Scoring Function Terms, T_k's[a]

	atom/feature type			
T_k	A_k (molecule 1)	B_k (molecule 2)	w_k	α
T_0	heavy	heavy	1	α_a
T_1	donor	donor	w_{DA}	α_a
T_2	acceptor	acceptor	w_{DA}	α_a
T_3	hydrophobic	hydrophobic	w_{HH}	α_a
T_4	hydrophobic	acceptor/donor	w_{R1}	α_a
T_5	acceptor/donor	hydrophobic	w_{R1}	α_a
T_6	acceptor-not-donor	donor-not-acceptor	w_{R2}	α_a
T_7	donor-not-acceptor	acceptor-not-donor	w_{R2}	α_a
T_8	donor projected feature	donor projected feature	w_{PF}	α_p
T_9	acceptor projected feature	acceptor projected feature	w_{PF}	α_p

[a] The scoring function consists of a summation of terms, T_k's, each of which has the form:

$$\mathrm{T}_k = w_k \sum_{i \in \mathrm{A}_k} \sum_{j \in \mathrm{B}_k} \exp(-\alpha r_{ij}^2)$$

Each row of the table corresponds to one term, T_k, and i and j denote atoms (for T_0–T_7) or features (for T_8 and T_9) coming from the first and second molecule, respectively. These atoms/features have to be of certain types, A_k and B_k, as given in the second and third columns of the table, and r_{ij} is the distance between the atoms/features. The names of the various constants for the different terms, w_k and α, are given in the fourth and fifth columns of the table. All the atom-based terms have the same value for α, namely α_a, while all hydrogen-bond projected feature-based terms have another value for α, namely α_p.

To keep the number of parameters low, the same value of α, α_a, was used for all atom-based terms. The projected feature terms used a different value of α, α_p, allowing these terms to be more diffuse than the atom-based terms if necessary.

Note that we have ordered the terms in the scoring function. Atom-based terms were followed by projected feature terms. For the atom-based terms, attractive terms were followed by repulsive terms. And within the attractive terms, the ordering was by the expected importance and intensity of that type of interaction. Based on this order, the scoring function was built up term by term. At each stage, the one or two parameter(s) of the newly added term was/were optimized. A series of scoring functions were created using different values for the newly added parameter(s). The goodness of a scoring function was measured, as described below. The parameter value(s) corresponding to the best scoring function in the series would be picked. Then we proceeded to the next stage to optimize the next parameter(s).

Test Sets. After the scoring function was optimized, it was incorporated into the flexible alignment functionality in molecular operating environment (MOE).[22] Molecular alignment results were compared with two recent publications.[6,23]

Proschak et al.[6] performed rigid-body alignments for eight thermolysin ligands using their Shapelets method and

compared their result with two previous alignment methods.[24,25] We performed validation runs on these thermolysin ligands to see how our scoring function fared.

Chen et al.[23] studied the geometric accuracy of aligning small molecules using ROCS[26] and FLEXS.[27] They used eight protein targets as test sets: HIV protease, cyclin-dependent kinase 2 (CDK2), estrogen receptor 1 (ESR1), mitogen-activated protein kinase 14 (p38), thermolysin, human rhinovirus capsid, elastase, and trypsin. Six of these (all but thermolysin and elastase) consist exclusively of structures from the publicly accessible PDB.[17] We used these six systems for validation runs. Both rigid-body alignments and flexible alignments were compared.

For training a scoring function for the validation runs, all systems that contain any PDB entries involved in the test sets were naturally excluded.

Compiling the Training Set. Ligands from identical or almost identical pockets (see below for details) from the PDB were used for training the scoring function. Only PDB entries with a resolution of better than 2.5 Å were used. Ligands were defined as freestanding molecules with no more than 300 heavy atoms and 10 residues. In order to exclude waters and very small molecules, each ligand must have at least one pair of heavy atoms separated by three bonds. Ligands were also required to be reasonably drug-like. Ligands without carbon atoms or with exotic elements (elements other than C, H, N, O, S, P, F, Cl, Br, and I) were discarded. Hemes and sugars were excluded. ATP analogs were excluded because their large population in the PDB would likely skew the data. Ligands without any rings were also excluded so as to remove the lipids without excluding any drug-like compounds.

We first obtained a list of protein domains from Structural Classification of Proteins (SCOP).[28] For each domain, all PDB entries containing that domain were superposed using the protein backbone of that domain. Ligands were then detected around the domain. Overlapping ligands were clustered into groups so that each group would correspond to one pocket. For each pocket, a fine superposition was performed based on the α carbons common to all pockets.

Since we started from SCOP domains, it was possible that pockets within a group could involve different residues or have different geometries. For our purpose of scoring function training, it may not be very meaningful to study the alignment of ligands in different pockets. Therefore we divided each pocket into subgroups in the following way. We first collected all α carbons within 7.5 Å of any ligand. These were then clustered using single linkage clustering with a cutoff of 1.0 Å. Clusters with no α carbon within 6.0 Å of any ligand were discarded. A table was then constructed with each row corresponding to one protein and each column corresponding to one α carbon cluster. A cell of the table has the residue type as the entry if the protein contributes to that cluster. The pocket similarity between a pair of proteins was defined to be the fraction of table columns that are identical (i.e., same residue type, or empty cell, for all cells in the column). After the pocket similarities between all protein pairs were determined, the proteins were clustered, using complete linkage and pocket similarity as the metric, with a cutoff of 0.8. In other words, any two pockets in the same group had at least 80% of the table columns identical. Each original pocket group might thus be divided into multiple groups. For each resulting group, a fine superposition on the 3D structures was performed again using α carbons common to all pockets, since the definition of the pocket might have changed due to the change in the content of the ligand set. This whole procedure was repeated until there was no new division of the pocket groups.

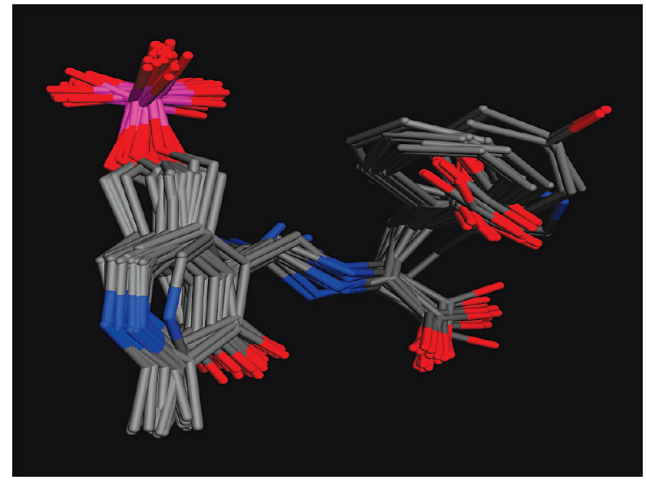

Figure 1. Aspartate aminotransferase ligands from their aligned PDB structures.

Upon inspection, some systems were found to be uninteresting in the sense that the aligned ligands only displayed one type of structure at most regions of the pocket. An example is given in Figure 1. Studying this type of system would not give us more information than studying self-alignments. Hence they were discarded from the training set.

It is desirable to have the conformational space reasonably covered when dummy poses were generated. Therefore the training ligands were restricted to those having no more than 12 rotatable bonds. Finally, to ensure diversity, a set of representative ligands was chosen for each system such that no pair of representatives had a Tanimoto similarity of over 0.8 on their MACCS fingerprints. Systems with fewer than five representatives were discarded.

Only 100 systems remained after these final filters, and 85 systems had no more than 20 representative ligands. The highest number of representatives in a system was 53. This is our comprehensive aligned small molecules data set.

For training a scoring function for the validation runs, 11 of the 100 systems in the comprehensive data set were excluded because they contained entries that would be used in the runs. This was more than the number of test systems (7) because some test systems (e.g., the 57 ligands in the CDK2 system) had variations in the pocket residue composition or geometry and were hence subdivided into more than one pocket group when our comprehensive data set was constructed.

Dummy Pose Generation. To keep things simple, the training of the scoring function was based on trying to align a flexible source ligand onto a rigid target ligand. The input conformation of the source ligand was randomized from the crystal conformation. Dummy poses were generated by making small adaptations to the MOE docking engine,[22] as described below.

For each system of n aligned ligands, all $n \times n$ pairs of ligands were considered with the following requirement. The fraction of heavy atoms in the target that lie within 2 Å of

any heavy atom of the source must exceed 0.5, and the fraction of heavy atoms in the source that lie within 2 Å of any heavy atom of the target must exceed 0.75. This excluded ligands with different binding modes and ensured that no source ligand was substantially larger than the target.

Conformers of the source ligand were systematically generated by assigning favorable torsion angles. If this resulted in more than 5000 conformers, a random subset of the systematically generated conformers would be used. Only conformers with severe atomic overlaps were discarded. This is because we envisioned that when the scoring function is used to distinguish poses with the correct binding mode, any poses that are fished out can be refined to improve the pose geometry and any atomic overlap within the molecule will be relieved. For each source−target pair, 15 000 dummy alignment poses were generated, one at a time. Each time, a random set of three heavy atoms on the target ligand was chosen, together with a random conformer of the source ligand. Triplets of heavy atoms on the source ligand, which formed a triangle similar to that formed by the three chosen target atoms, were used to align the source ligand onto the target ligand. Poses whose root-mean-square deviation (rmsd) for the heavy atoms was under 1.5 Å were considered duplicated, and only one copy was kept.

RESULTS AND DISCUSSION

Scoring Function Training. For each scoring function, the scores for all dummy poses were calculated. For each set of alignments involving a particular pair of source and target ligands, we counted the number of "good" poses, n25, among the top 25 scoring poses. As for whether a pose is "good" or not, we used fuzzy logic. Poses with an rmsd (heavy atoms only) from the correct pose of under 2.0 Å were considered good. Poses with an rmsd of over 3.0 Å were considered not good. Poses with an rmsd between 2.0 and 3.0 Å were considered partial good poses, with the "partiality" decreasing linearly from 1 to 0. For each system (i.e., same pocket), we averaged the number of good poses for all source−target pairs in the system. We then averaged these averaged numbers over all 89 training systems. The higher this final averaged number was, the better the scoring function was.

The receiver operating characteristic (ROC) curve is often used to measure the success of scoring functions for virtual screening.[29] In short, the ROC curve plots the fraction of true positives recovered (out of all active compounds) against the fraction of false positives obtained (out of all inactive compounds), as the cutoff value varies. It is generally considered that the bigger the area under the curve, the better is the scoring function. We felt that, for our purpose, using the area under the ROC curve was not as suitable as using "the number of good poses amongst the top 25", since we generally can only afford to consider a limited number of top scoring poses. Hence all but the very beginning part of the ROC curve would be of significance to us.

As for why 25? We think that human inspection can typically handle 10 or so poses. On the other hand, if a thorough refinement is desired as part of an automatic process, then it would be reasonable to intensively process up to about 100 poses per ligand. Hence here we used the "top 25 poses". As a precaution, we repeated all the calculations and validation runs using the "top 100 poses" instead of the "top 25". It turned out that only one out of the six optimized parameters for the scoring function would come out different, and the quality of the of validation runs remained very similar. The details are given in Section A of the Supporting Information.

Table 2a. Parameter Optimization for α_a[a]

α_a (Å^{-2})	0.06	0.125	0.175	0.25	0.35	**0.5**	0.7
n25	3.10	4.11	4.44	4.67	4.79	**4.83**	4.76

[a] The scoring function at this stage is T_0, the volume overlap term. Refer to Table 1 for details of T_0. This table gives the average number of good poses, n25, within the top 25 scoring poses for various values of α_a.

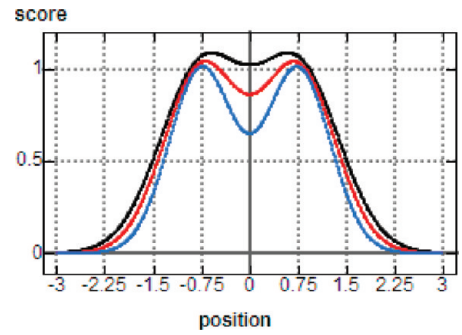

Figure 2. How the scoring function is controlled by the parameter α, for the case of a target molecule consisting of two atoms, located at positions −0.75 and 0.75. For a source atom to be superimposed onto the target molecule, the scoring function is $\exp(-\alpha r_1^2) + \exp(-\alpha r_2^2)$, where r_1 and r_2 are the distances from the source atom to the two target atoms. The scoring functions corresponding to α values of 2.0, 1.5, and 1.2 are plotted in blue, red, and black, respectively.

Table 2b. Parameter Optimization for w_{DA}[a]

w_{DA}	0	0.5	1	2	**4**	8	16	32
n25	4.83	5.28	5.61	5.98	**6.10**	5.85	5.40	4.99

[a] The scoring function at this stage is ($T_0 + T_1 + T_2$). The new terms are T_1 and T_2, which correspond to the attraction between hydrogen-bond donor/acceptor atoms. Refer to Table 1 for the details of these terms. This table gives the average number of good poses, n25, within the top 25 scoring poses for various values of w_{DA}.

As mentioned previously, the constituent terms of the scoring function are given in Table 1. The scoring function was built up term by term. At each stage, only one (or two) parameter was introduced. This parameter was varied, and the most favorable value for it was picked. We then proceeded to optimize the next parameter. The optimization procedure is captured in Tables 2a−2h. Each table corresponds to one stage of optimization where one (or two) parameter was optimized. The resulting numbers of good poses among the top 25, n25, corresponding to the various values of the parameter being optimized are given in the tables. The optimal value of the parameter and its corresponding value of n25 are highlighted in bold.

We started by including only the volume overlap term (T_0 in Table 1) in the scoring function. In other words, to a first approximation, the bigger the volume overlap between two molecules, the better was the score. In order to make the scoring function quick to calculate, hydrogen atoms were

Table 2c. Parameter Optimization for w_{HH} using the MOE Definition for Hydrophobic Atoms[a]

w_{HH}	0	0.5	**1**	2	4	8
n25	6.10	6.33	**6.36**	6.23	5.90	5.44

[a] The scoring function at this stage is ($T_0 + T_1 + T_2 + T_3$). The new term is T_3 and corresponds to the attraction between hydrophobic atoms. Refer to Table 1 for the details of this term. This table gives the average number of good poses, n25, within the top 25 scoring poses for various values of w_{HH}.

Table 2d. Parameter Optimization for w_{HH}, where Hydrophobic Atoms Are Defined as Atoms That Are at Least Two Bonds Away from any Hydrogen-Bonding Atom[a]

w_{HH}	0	0.5	**1**	2	4	8
n25	6.10	6.30	**6.32**	6.17	5.81	5.29

[a] The scoring function at this stage is ($T_0 + T_1 + T_2 + T_3$). The new term is T_3 and corresponds to the attraction between hydrophobic atoms. Refer to Table 1 for the details of this term. This table gives the average number of good poses, n25, within the top 25 scoring poses for various values of w_{HH}.

Table 2e. Parameter Optimization for w_{R1} using the MOE Definition for Hydrophobic Atoms[a]

w_{R1}	0	−0.125	−0.25	−0.5	**−1**	−2	−4	−8
n25	6.36	6.38	6.40	6.44	**6.45**	6.36	5.92	4.56

[a] The scoring function at this stage is ($T_0 + T_1 + T_2 + T_3 + T_4 + T_5$). The new terms T_4 and T_5 correspond to the repulsion between hydrophobic and hydrogen-bonding atoms. Refer to Table 1 for the details of these terms. This table gives the average number of good poses, n25, within the top 25 scoring poses for various values of w_{R1}.

Table 2f. Parameter Optimization for w_{R1}, where Hydrophobic Atoms Are Defined as Atoms That Are at Least Two Bonds Away from any Hydrogen-Bonding Atom[a]

w_{R1}	0	−0.125	−0.25	**−0.5**	−1	−2	−4	−8
n25	6.32	6.33	6.34	**6.35**	6.34	6.23	5.89	5.22

[a] The scoring function at this stage is ($T_0 + T_1 + T_2 + T_3 + T_4 + T_5$). The new terms T_4 and T_5 correspond to the repulsion between hydrophobic and hydrogen-bonding atoms. Refer to Table 1 for the details of these terms. This table gives the average number of good poses, n25, within the top 25 scoring poses for various values of w_{R1}.

Table 2g. Parameter Optimization for w_{R2}[a]

w_{R2}	**0**	−0.125	−0.25	−0.5	−1	−2
n25	**6.45**	6.45	6.45	6.44	6.43	6.40

[a] The scoring function at this stage is ($T_0 + T_1 + T_2 + T_3 + T_4 + T_5 + T_6 + T_7$). The new terms T_6 and T_7 correspond to the repulsion between hydrogen-bond donor atoms that are not acceptors and hydrogen-bond acceptor atoms that are not donors. Refer to Table 1 for the details of these terms. This table gives the average number of good poses, n25, within the top 25 scoring poses for various values of w_{R2}.

Table 2h. Parameter Optimization for w_{PF} and α_p[a]

w_{PF}	0	0.25	0.5	1	2	4	8
			For $\alpha_p = 0.5$ Å^{-2}				
n25	6.45	6.48	6.50	6.52	6.48	6.35	6.01
			For $\alpha_p = 0.25$ Å^{-2}				
n25	6.45	6.51	6.54	6.60	6.61	6.52	6.22
			For $\alpha_p =$ **0.125** Å^{-2}				
n25	6.45	6.54	6.61	6.68	**6.71**	6.60	6.28
			For $\alpha_p = 0.0625$ Å^{-2}				
n25	6.45	6.57	6.64	6.69	6.66	6.44	5.98

[a] The scoring function at this stage is ($T_0 + T_1 + T_2 + T_3 + T_4 + T_5 + T_8 + T_9$). The new terms T_8 and T_9 correspond to the attraction between hydrogen-bond donor/acceptor projected features. Refer to Table 1 for the details of these terms. This table gives the average number of good poses, n25, within the top 25 scoring poses for various values of w_{PF} and α_p.

Table 3. Optimal Parameters for the Scoring Function[a]

α_a	α_p	w_{DA}	w_{HH}	w_{R1}	w_{PF}
0.5 Å^{-2}	0.125 Å^{-2}	4	1	−1	2

[a] The optimal scoring function consists of the terms T_0−T_5, T_8, and T_9, as given in Table 1, with the various parameters as given by this table.

Table 4. Success Rates of the 28 Rigid-Body Cross-Alignments of 8 Thermolysin Ligands Using Various Methods[a]

SURFCOMP,[24] electrostatic potential	75%
SURFCOMP,[24] lipophilic potential	71%
Cosgrove et al.[25]	50%
Shapelets[6]	71%
current work	93%

[a] The eight thermolysin ligands have PDB ID's 1THL, 1TLP, 1TMN, 3TMN, 4TMN, 5TLN, 5TMN, and 6TMN. For each pair of nonidentical ligands, one ligand was chosen as the template, and the other ligand was aligned onto the template. If the top scoring solution yielded an rmsd of under 2 Å with the crystallographic pose for the heavy atoms, then it was considered successful. This table gives the success rates for the 28 cases, as reported on Table 4 of Proschak et al.[6]

ignored. The only parameter to be optimized was α. Obviously a smaller value for α would result in a more diffuse Gaussian function. Consider the simple case to align one atom (the source) onto a diatomic molecule (the target). This is similar to trying to align methane, with one heavy atom, onto ethane, with two heavy atoms. Suppose the internuclear distance of the diatomic molecule is 1.5 Å (for comparison, a C−C bond in a benzene ring is about 1.40 Å, and a C−C bond in an ethane is about 1.54 Å). The shapes of the scoring function for three different values of α are plotted in Figure 2. For large values of α, the source atom will see a score with two distinct peaks located very close to the two target atoms. It will have a strong tendency to align with either one of the target atoms. For example, if α is 2.0 Å^{-2} (the blue curve of Figure 2), then the distance between the peaks will be 1.46 Å. The value of the score at the trough midway between the peaks will be 64% of the height of the peaks. When α decreases to 1.2 Å^{-2} (the black curve of Figure 2), the peaks will move toward each other, away from the atomic positions. The distance between the peaks will drop to 1.18 Å, and the trough to peak height ratio of the score will increase to 94%. In other words, the score has already very much smoothed out the atomic positions. The source atom can settle for any position between the two atoms with little change on the score.

Given a scoring scheme and an initial aligned pose between two rigid ligands, one can refine the alignment

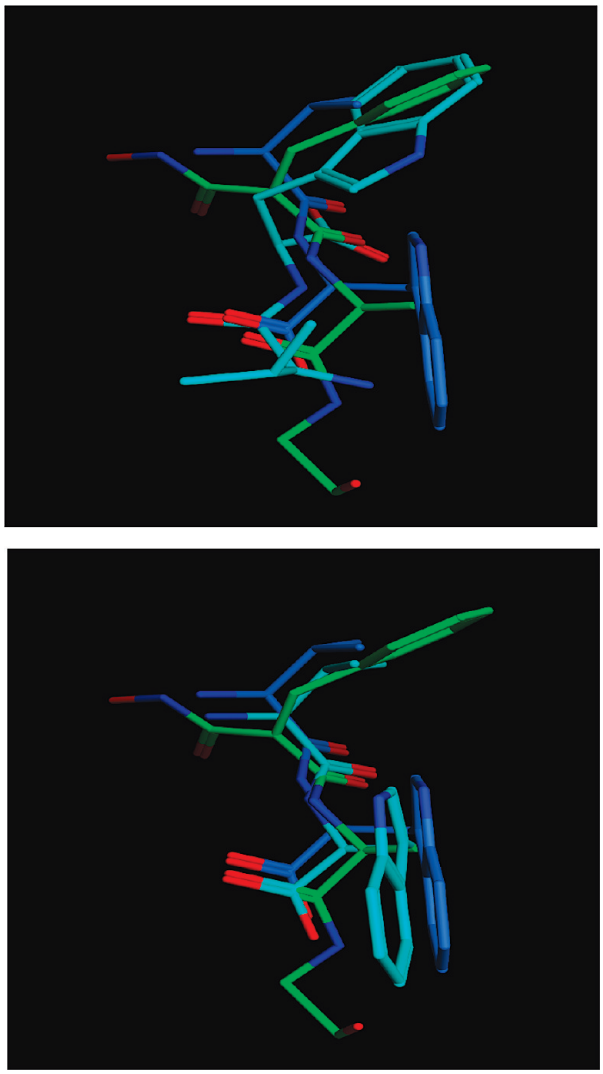

Figure 3. Rigid-body alignment of 3TMN onto 5TLN (carbons in green). Carbons of the correct answer for 3TMN are in dark blue. Carbons of our solutions (top and bottom are best and second best scoring, respectively) are in cyan.

by optimizing the score. When we did this with many starting poses, we found that we ended up with fewer resulting aligned poses when we used smaller values of α (unpublished results). This is in line with our expectation that smaller values of α smooth out the score landscape.

According to previous works of Gaussian function-based alignment scores, values of α of between 0.1 and 0.5 Å^{-2} were found to be reasonable.[4,16,19,20] In light of this, we

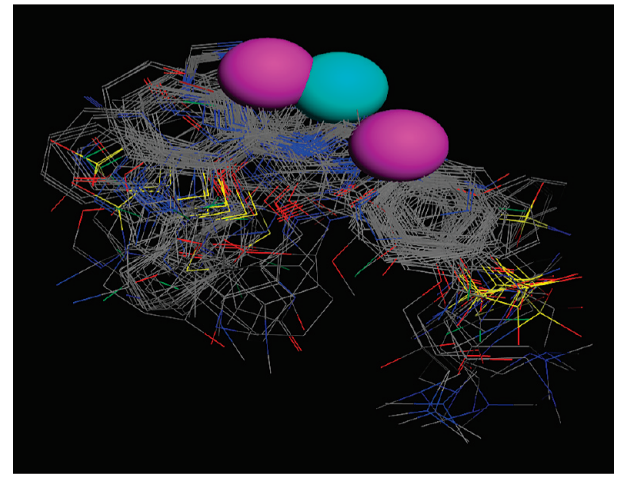

Figure 4. Important pharmacophore features for the CDK2 ligands include two hydrogen-bond donors (magenta sphere) and one acceptor (cyan sphere).

tried out a total of 7 values for α_a, from 0.06 to 0.7 Å^{-2}. The corresponding numbers of good poses among the top 25 are tabulated in Table 2a. It can be seen that for the values of α_a tested, a value of 0.5 Å^{-2} was the best, with an average of 4.83 good poses among the top 25.

After establishing a value for α_a, the relative weights of the atom-based terms were optimized one by one. We first looked at the weight w_{DA} for the hydrogen-bonding atoms (donor and acceptor) similarity terms (T_1 and T_2 in Table 1). For this we tried out 7 values for w_{DA}, from 0.5 to 32, in logarithmic intervals. The corresponding numbers of good poses among the top 25 are tabulated in Table 2b. A value of 4 turned out to be the most favorable, with an average of 6.10 good poses among the top 25.

After this, we turned to the similarity term between hydrophobic atoms (T_3 in Table 1). Besides trying to establish the optimal weight w_{HH} for this term, we also tried out two schemes of defining hydrophobic atoms. One used the conventional hydrophobic atom assignment in MOE.[22] The other scheme was based on how far topologically an atom was from any hydrogen-bonding atom. Atoms that were two or more bonds away from any hydrogen-bond acceptors or donors were considered hydrophobic. Tables 2c and 2d give the numbers of good poses among the top 25 for various values of w_{HH} using the two schemes. It can be seen that the two schemes gave very similar results, with the optimal weighting factor w_{HH} to be around 1.0 for both schemes.

Table 5. Percentage of Correct Results Given by the Current Scoring Function Compared to Those from Two Published Methods[a]

alignment mode	rigid			flexible		
method	ROCS	FLEXS	current	ROCS	FLEXS	current
CDK2	30%	25%	40%	20%	21%	22%
HIV	39%	24%	85%	6%	8%	16%
P38	27%	27%	43%	22%	24%	30%
ESR1	44%	47%	59%	25%	28%	41%
trypsin	57%	73%	80%	55%	29%	61%
rhinovirus	50%	52%	50%	50%	50%	50%

[a] For each of the six systems, pairwise cross-alignments were carried out for all $n \times n$ ligand pairs, where n is the number of ligands in the system. If the top scoring result yielded an rmsd of under 2 Å for the heavy atoms, then it was considered correct. This table gives the percentage of correct results for the $n \times n$ ligand pairs in the system, as reported on Table 2a of Chen et al.[23]

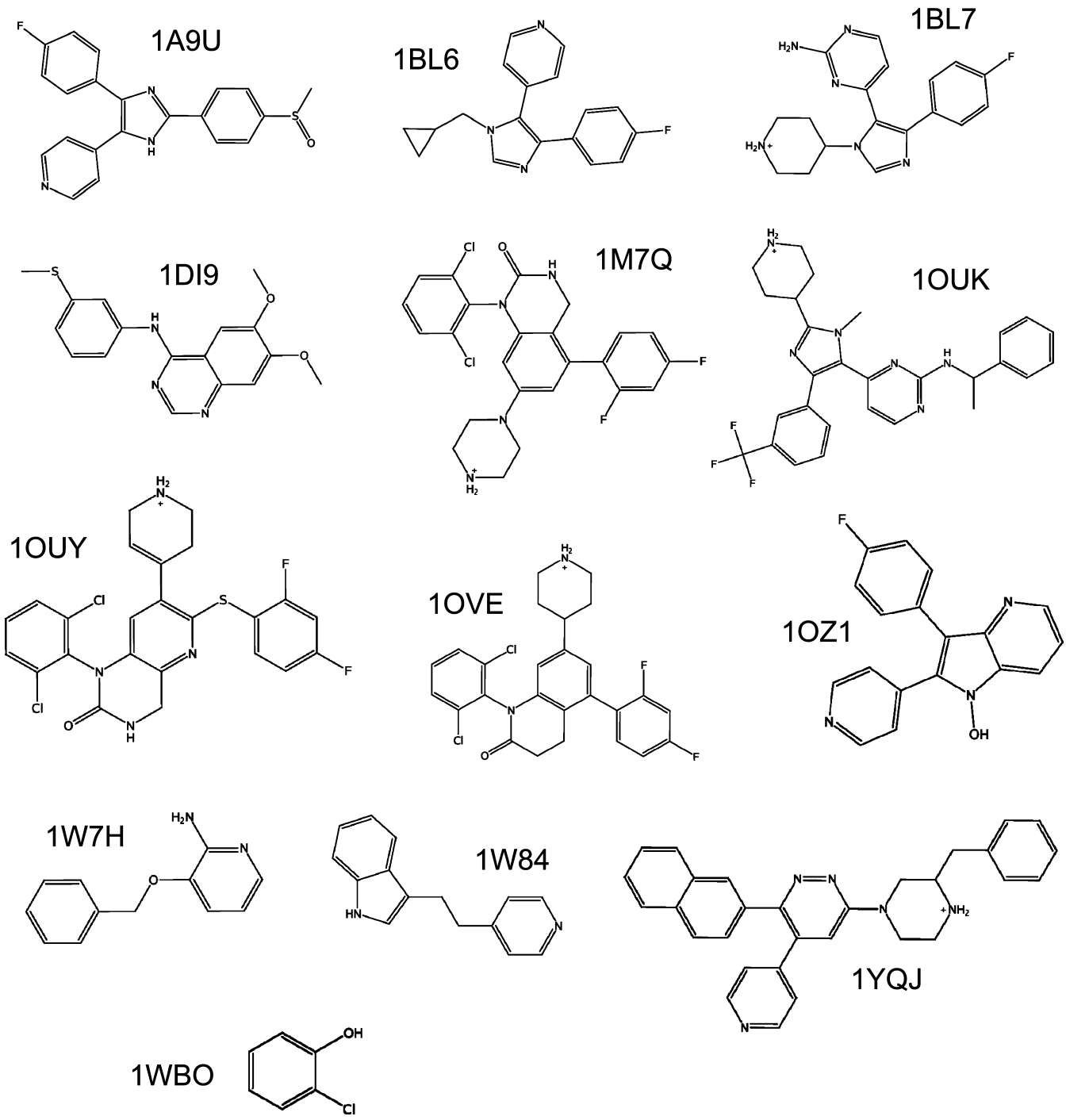

Figure 5. The 13 ligands of the p38 system.

Next we considered the repulsion term between hydrogen-bonding atoms and hydrophobic atoms (T_4 and T_5 in Table 1). We continued to test both schemes for defining hydrophobicity. Results are given in Tables 2e and 2f. Here we see that the first scheme of defining hydrophobicity was slightly better and that the optimal weight for the repulsion term, w_{R1}, was around -1.0. So we chose to define hydrophobic atoms simply using the MOE definition.

After this we investigated a repulsion term between hydrogen-bond donor and acceptor atoms (T_6 and T_7 in Table 1). Naturally, atoms that can be both donor and acceptor were excluded from consideration. Perhaps surprisingly, it was found that this term did not seem to improve the scoring function (see Table 2g). So these terms were dropped from the scoring function.

Finally the projected acceptor and donor feature points were considered (T_8 and T_9 in Table 1). These points correspond to the expected positions of donors and acceptors on the binding pocket. In this sense they are different from the previously considered terms that are atom based. So we tried out different values of α as well as different weights, w_{PF}, for this term. The results are given in Table 2h. It can be seen that an α_p value of 0.125 $Å^{-2}$ coupled with a weight of 2.0 was a good combination. Otherwise an α_p value of 0.0625 $Å^{-2}$ coupled with a weight of 1.0 was also respectable.

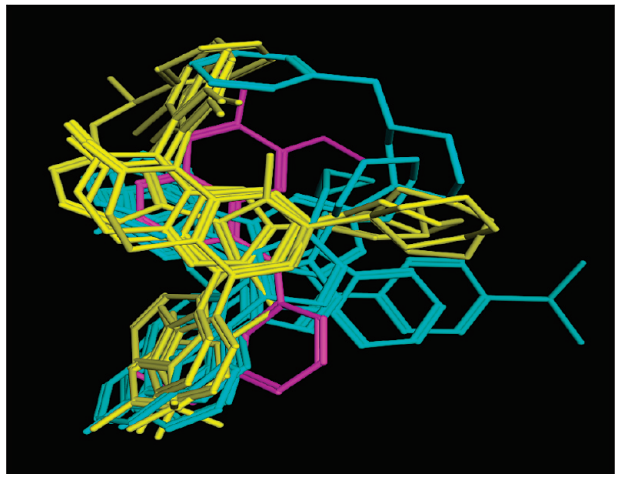

Figure 6. Different binding modes of the p38 ligands. 1DI9 is in magenta. 1M7Q, 1OUK, 1OUY, and 1OVE are in yellow.

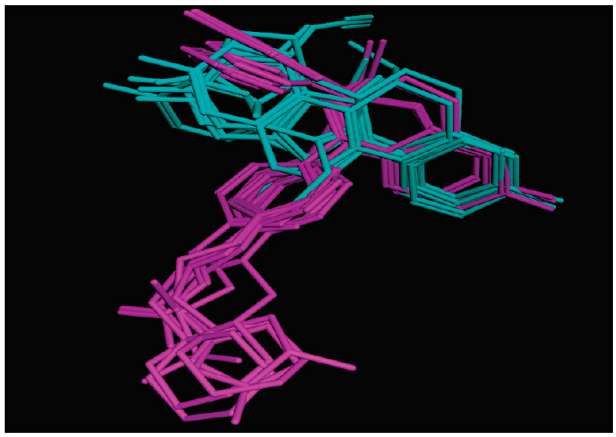

Figure 7. Two binding modes of the ESR1 ligands. The 6 smaller ligands (1A52, 1GWQ, 1L2I, 1X7E, 1X7R, and 3ERD) are in cyan. The 7 larger ligands (1R5K, 1SJ0, 1UOM, 1XP1, 1XP9, 1XQC, and 2BJ4) are in magenta.

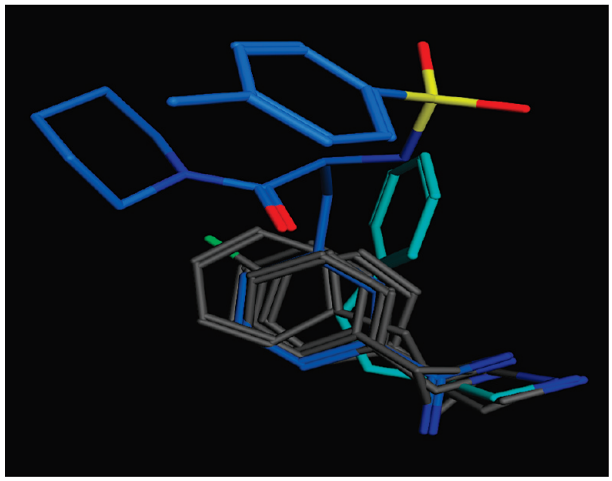

Figure 8. Crystallographic alignment of the 7 trypsin ligands. Carbon atoms of 1PPH are in dark blue. Those of 1TNI are in cyan. The remaining ligands are 1TNH, 1TNJ, 1TNK, 1TNL, 3PTB.

To summarize, it was found that a good scoring function is of the form given by the terms of Table 1, except terms T_6 and T_7, with the values of the various coefficients given in Table 3.

Uncertainty Considerations. For each of the parameters considered, only one run was performed. However, there was a considerable number of ligand−pocket systems, and in each system, there was a considerable number (at least 25, but often much more) of pairs of source−target ligands. Therefore random noise can be expected to have averaged out. More importantly, the fact that there are smooth trends in the numbers in Tables 2a−2h, as the parameters were varied, serves to confirm that the numbers are accurate enough for our purpose of locating the optimal values for the parameters.

Validation Runs: The Procedure. MOE's flexible alignment (FlexAlign) functionality[22] was used for the validation runs. For rigid-body alignments, conformers from the PDB were used as the input. FlexAlign generated random poses of the source molecule superposed onto the target molecule and optimized the alignment score S, while keeping the molecules rigid.

For the flexible alignments, for each source molecule, one randomized conformation was used as input. FlexAlign generated a random conformer from this input, randomly superposed it onto the target molecule, and then optimized this pose under the influence of the alignment score between the source and target molecules. The target molecule remained fixed throughout the process. In order to avoid unrealistic conformations, the internal energy U of the source molecule must also be considered alongside the alignment score S. For the internal energy, we used the MMFF94 force field[22,30] with a distance-dependent dielectric. Pose optimization was carried out by minimizing the effective energy, E, which was defined by

$$E = U - kS$$

Alignment solutions were also ranked according to their E values. A value of 4 kcal/mol was used for k. This value seemed to give a good balance between the internal energy U and the alignment score S. The difference in the force field energy U between the conformation of the source molecule in the top scoring (i.e., the one with the lowest E value) solution and its corresponding local energy minimum was generally small. The average of this difference (over the $n \times n$ cross-alignments for each system) for the CDK2, HIV, p38, ESR1, trypsin, and rhinovirus systems were, respectively, 0.32, 0.76, 0.25, 0.33, 0.25, 0.25 kcal/mol per heavy atom. In contrast, the average of the difference in the internal energy U between the ligand conformer in the crystal structure and its corresponding local energy minimum is over 1 kcal/mol per heavy atom for all the six systems.

We also investigated whether the value of the alignment score S or the effective energy E gives an indication of the goodness of the alignment. Unfortunately there did not seem to be any correlation (unpublished results).

Validation Runs: The Results. Our results were first compared with those in Table 4 of Proschak et al.[6] This involved pairwise rigid-body alignments on a system of eight thermolysin ligands. For each ligand pair, one ligand was used as the target, while the other was used as the source. Self-alignments were not performed. Therefore there was a total of 28 cross-alignments. An alignment was considered successful if the top scoring pose had a heavy atom rmsd of under 2 Å from the correct answer. Our results, together with those obtained by Proshak et al.,[6] are tabulated in Table 4.

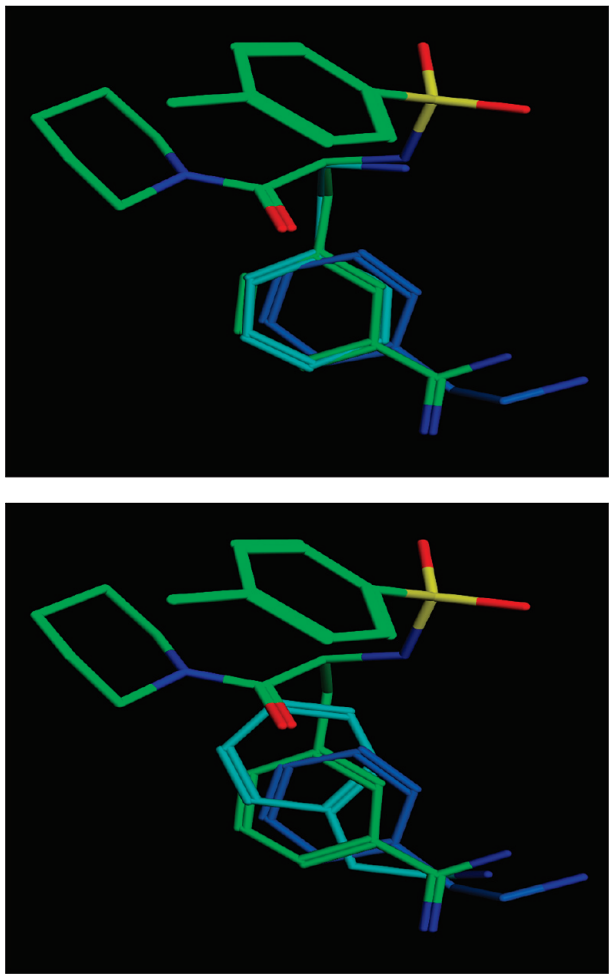

Figure 9. Flexibly aligning 1TNJ onto 1PPH (carbons in green). Carbons of the correct pose of 1TNJ are in dark blue. Carbons of our solutions (top and bottom: best and second best scoring) are in cyan.

According to Proshak et al., when SURFCOMP[24] was used with the electrostatic potential, the success rate was 75%. If the lipophilic potential was used, then the success rate was 71%. The method of Cosgrove et al.[25] gave a success rate of 50%. The success rate for the authors' own Shapelets method was 71%. In contrast, our scoring function yielded a success rate of 93%; there were only 2 failed cases out of 28 alignments. One of our failed cases was for aligning 3TMN onto 5TLN. Our result is shown in Figure 3. Our scoring function returned the correct answer as the second best scoring solution (Figure 3, bottom). In any case, our best scoring solution looks highly plausible with heavy overall volumetric overlap as well as the overlap of the aromatic ring and two carbonyl groups (Figure 3, top).

Tests were also run on six systems mentioned in Chen et al.,[23] where ROCS[26] and FLEXS[27] were used. An alignment was considered correct if the top scoring solution had a heavy atom rmsd of under 2 Å with the crystallographic overlay. Table 5 gives the overall results for the six systems. Except for the rhinovirus system, which will be examined in more details below, and for the flexible alignment of the CDK2 system, our results are significantly better than those obtained by Chen et al. using FLEXS or ROCS. For rigid-body alignments, our fraction of correct results was at least 40% and reached 80% or more in two systems. Naturally, flexible alignment is more difficult than rigid-body alignment, and so the percentages of correct results are lower for all systems except rhinovirus (see analysis below).

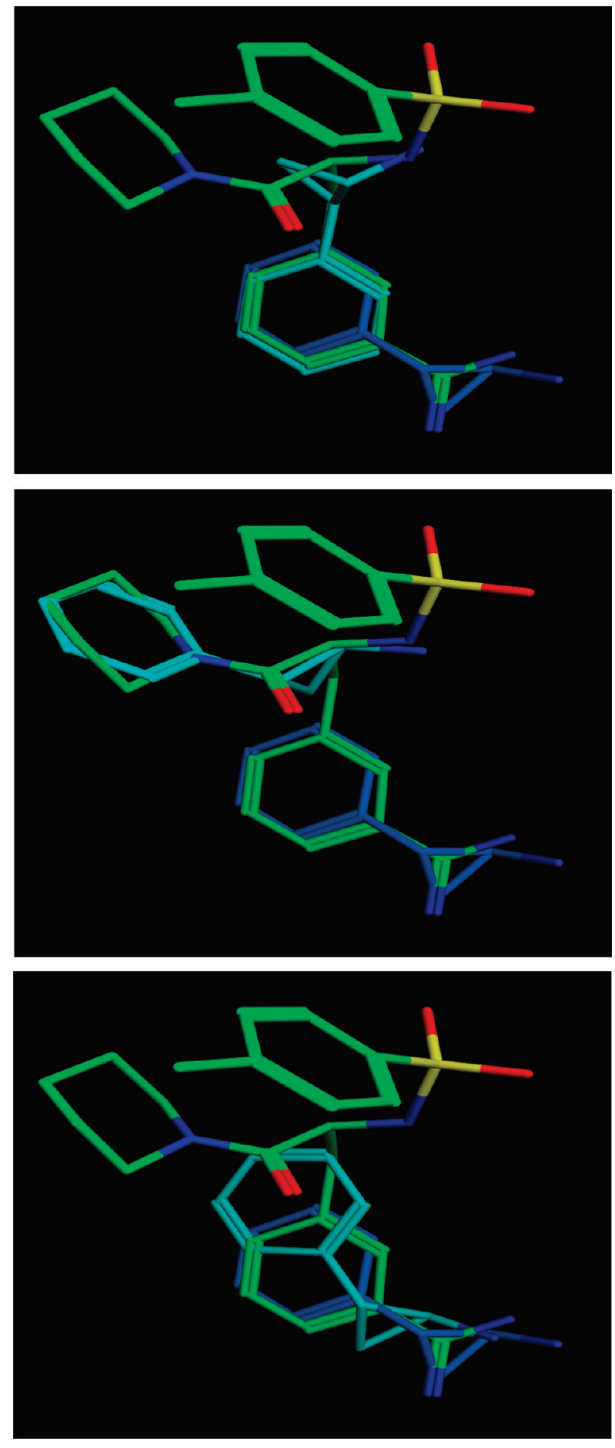

Figure 10. Flexibly aligning 1TNL onto 1PPH (carbons in green). Carbons of the correct pose of 1TNL are in dark blue. Carbons of our solutions (from top to bottom: first, second, and third best scoring) are in cyan.

For the CDK2 system, our results for the rigid-body alignments but not for the flexible alignments are significantly better than those of Chen et al. using ROCS or FLEXS. CDK2 is a large system (57 ligands) with a complex pharmacophore model.[31] If the six ligands 1DI8, 1P5E, 1PKD, 1PXI, 1PXJ, and 1WCC were excluded, then our results would improve from 39.6% to 47.6% correct for the rigid-body alignments and from 22.4% to 26.6% correct for

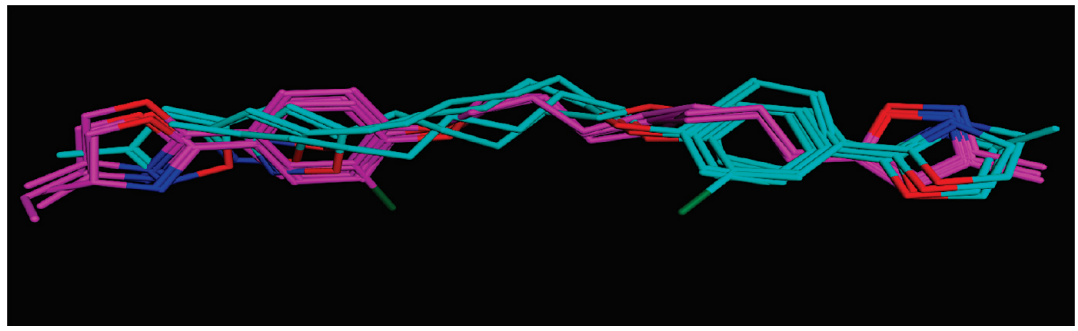

Figure 11. The eight rhinovirus ligands bind in two modes. The carbons are in magenta for one mode (2RM2, 2RR1, 2RS1, and 2RS3) and in cyan for the other mode (2R04, 2R06, 2R07, and 2RS5).

the flexible alignments. Of the 57 ligands in the system, only 7 ligands have more than 33 or less than 17 heavy atoms, and 1P5E, 1PKD, 1PXI, 1PXJ, and 1WCC are among them, with 13, 36, 14, 14, and 8 heavy atoms, respectively. Since the pharmacophore model is complex,[31] it seems that volumetric overlap is an important factor for alignment. Hence ligands that are much smaller or much larger than average would be difficult to align. Moreover, an important pharmacophore element is the set of two hydrogen-bond donors and one acceptor interacting with the hinge region of the protein. These correspond to pharmacophore features D1, D2, and A1 mentioned in Zou et al.[31] and are shown in Figure 4. Out of the 57 ligands, only 4 do not have at least 2 out of these 3 features, and ligands 1DI8, 1P5E, and 1WCC are among them.

The HIV ligands are big. All but one have 35 or more heavy atoms. The fact that they are relatively comparable in size (all but three have between 39 and 61 atoms), do not have a homogeneous (spherical) shape, and have rich feature sets might have helped the rigid-body alignment result, because there are not that many possibilities for the molecules to align well. One possibility for our superior results is that we have trained our scoring function using a significantly larger data set from the PDB than the version of FLEXS or ROCS that were used. Of all six test systems, the drop in performance in going from the rigid-body to the flexible mode is most pronounced for the HIV system. This is because the HIV ligands are particularly big and flexible. Of the 28 HIV ligands, 5 have 10 rotatable bonds. 7 have between 12 and 16 rotatable bonds. The other 16 have 18 or more rotatable bonds. In comparison, no molecules in the other 5 systems have more than 11 rotatable bonds.

For the 13 ligands of the p38 system, our scoring function produced an overall correct rate of 43% and 30% for the rigid-body and flexible alignments, respectively (Table 5). It is worth noting that one of the ligands, 1WBO, is significantly smaller than the others (Figure 5). The crystallographic alignment of the remaining 12 ligands is given in Figure 6. It can be seen that the ligand 1DI9 (magenta in Figure 6) binds in a mode different from the others. The four ligands 1M7Q, 1OUK, 1OUY, and 1OVE (yellow in Figure 6) also bind in a mode somewhat different from the others. For the remaining 7 ligands, our scoring function actually got 73% of the pairwise cross-alignments correct in the rigid-body mode, and 53% correct in the flexible mode. And for the 16 pairwise cross-alignments of the four ligands 1M7Q, 1OUK, 1OUY, and 1OVE, our scoring function got them all, 100%, correct in the rigid-body mode and 94% (all but one of the 16) correct in the flexible mode.

The 13 ligands of the estrogen receptor system can be divided into two groups. Ligands within each group have comparable sizes and can be considered to bind in a similar mode, as can be seen in Figure 7. For the pairwise cross-alignments of all 13 ligands, our scoring function got 59% and 41% correct in the rigid-body and flexible modes, respectively. However, for the group of 6 smaller ligands, we got 94% of the 36 rigid-body cross-alignments correct and 72% of the flexible alignments correct. For the group of 7 larger ligands, we got 92% of the 49 rigid-body cross-alignments correct and 65% of the flexible alignments correct.

The crystallographic alignment of the 7 trypsin ligands is given in Figure 8. Ligand 1PPH (dark blue in Figure 8) is significantly larger than the others. Ligand 1TNI (cyan in Figure 8) assumes a binding mode different from the others. For the remaining 5 ligands, our scoring function got all 25 cross-alignments correct in both the flexible and rigid-body modes. Figure 9 gives our result for flexibly aligning 1TNJ onto the much larger ligand 1PPH. Our second best scoring solution (Figure 9, bottom) yielded the correct alignment mode. As for our best scoring solution (Figure 9, top), it involves a perfect atom to atom matching for all atoms of 1TNJ. The aromatic ring is mapped to an aromatic ring of 1PPH, all aliphatic carbons are mapped to aliphatic carbons on 1PPH, and the nitrogen atom is mapped to a nitrogen on 1PPH. Figure 10 gives our result for flexibly aligning 1TNL onto 1PPH. Our third best scoring solution (Figure 10, bottom) roughly yielded the correct binding mode. Our best scoring solution (Figure 10, top) involves a matching of atoms similar to that of the best scoring solution of aligning 1TNJ onto 1PPH (Figure 9, top). Our second best scoring solution (Figure 10, middle) also involves a good matching of atoms, with the phenyl ring of 1TNL aligned to a six-membered ring of 1PPH.

For the 8 rhinovirus ligands, our scoring function got 50% of the pairwise cross-alignments correct in both the rigid-body and flexible modes. As shown in Figure 11, all eight ligands have a long shape and are almost symmetric. There is a heterocyclic ring on either end of each ligand, connected by a long, flexible linker chain. As noted in the original publication for the X-ray structures,[32] the ligands have two binding modes, one being the inverse of the other, as shown in Figure 11. It is easy for the alignment algorithm to

superpose the two rings on either end but difficult to distinguish between the two binding modes. Hence all the cross-alignments within a binding mode were obtained correctly, but the inverse binding mode was obtained for each pair across the different binding modes.

Scoring Function Training with All Data. When the scoring function was trained for the validation runs, systems involved in the test sets were excluded. After the validation runs, the training process was repeated to make use of all data, including systems involved in the test sets. It turned out that using all the data did not lead to any change in the scoring function. The updated Tables 2a−2h are given in Tables a−h of section B of the Supporting Information. Although the numbers of good poses among the top 25 (n25) have changed slightly, the positions for the optimal values for all the parameters have not. Hence the scoring function remained the same when all data was included for training.

CONCLUSION

Based on the PDB[17] and SCOP,[28] a comprehensive set of aligned ligands binding in the same pocket has been compiled. We used only high-resolution structures and considered only pockets that are highly similar and have diversified, drug-like ligands. For each pair of ligands binding in the same pocket, one ligand was used as the target, and dummy aligned poses of the other ligand were generated. By studying the correlation between the score of these dummy poses and their similarities with the correct answer, a small molecule alignment scoring function was built up term by term. Eventually we obtained the scoring function composed of terms given by Table 1, except terms T_6 and T_7, with parameters as given by Table 3.

We verified our scoring function by comparing results of rigid-body and flexible alignments performed using the flexible alignment functionality in MOE[22] with those reported in the literature. We found that our results are superior to those of five methods reported in two recent publications.[6,23]

We are happy to share our comprehensive aligned small molecules data set with other scientists so that they can use it to optimize their own scoring functions.

ACKNOWLEDGMENT

We thank our colleagues in Chemical Computing Group for valuable comments.

Supporting Information Available: In the main text above, parameters of the scoring function were trained by considering the "top 25 scoring poses" (Results and Discussion Section) and by using data excluding the validation test sets. If "top 100" instead of "top 25" were considered for training, but still using data excluding the validation test sets, the optimal parameters for the scoring function would change very slightly, and so would the validation results, as given in section A of the Supporting Information. If all available data was used for training, and still by considering the "top 25", the optimal scoring function would remain the same. However, Tables 2a−2h would look like Tables a−h in section B of the Supporting Information. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
(2) Lemmen, C.; Langauer, T. Computational Methods for the Structural Alignment of Molecules. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 215–232.
(3) Gardiner, E. J.; Cosgrove, D. A.; Taylor, R.; Gillet, V. J. Multiobjective Optimization of Pharmacophore Hypotheses: Bias Toward Low-Energy Conformations. *J. Chem. Inf. Model.* **2009**, *49*, 2761–2773.
(4) Shin, W.; Hyun, S. A.; Chae, C. H.; Chon, J. K. Flexible Alignment of Small Molecules Using the Penalty Method. *J. Chem. Inf. Model.* **2009**, *49*, 1879–1888.
(5) Taminau, J.; Thijs, G.; de Winter, H. Pharao: Pharmacophore Alignment and Optimization. *J. Mol. Graphics Modell.* **2008**, *27*, 161–169.
(6) Proschak, E.; Rupp, M.; Derksen, S.; Schneider, G. Shapelets: Possibilities and Limitations of Shape-Based Virtual Screening. *J. Comput. Chem.* **2008**, *29*, 108–114.
(7) Todorov, N. P.; Alberts, I. L.; de Esch, I. J. P.; Dean, P. M. QUASI: A Novel Method for Simultaneous Superposition of Multiple Flexible Ligands and Virtual Screening Using Partial Similarity. *J. Chem. Inf. Model.* **2007**, *47*, 1007–1020.
(8) Marialke, J.; Körner, R.; Tietze, S.; Apostolakis, J. Graph-Based Molecular Alignment (GMA). *J. Chem. Inf. Model.* **2007**, *47*, 591–601.
(9) Wolber, G.; Dornhofer, A. A.; Langer, T. Efficient Overlay of Small Organic Molecules Using 3D Pharmacophores. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 773–788.
(10) Richmond, N. J.; Abrams, C. A.; Wolohan, P. R. N.; Abrahamian, E.; Willett, P.; Clark, R. D. GALAHAD: 1. Pharmacophore Identification by Hypermolecular Alignment of Ligands in 3D. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 567–587.
(11) Cheeseright, T.; Mackey, M.; Rose, S.; Vinter, A. Molecular Field Extrema as Descriptors of Biological Activity: Definition and Validation. *J. Chem. Inf. Model.* **2006**, *46*, 665–676.
(12) Cho, S. J.; Sun, Y. FLAME: A Program to Flexibly Align Molecules. *J. Chem. Inf. Model.* **2006**, *46*, 298–306.
(13) Rönkkö, T.; Tervo, A. J.; Parkkinen, J.; Poso, A. BRUTUS: Optimization of a Grid-Based Similarity Function for Rigid-Body Molecular Superposition. II. Description and Characterization. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 227–236.
(14) Feng, J.; Sanil, A.; Young, S. S. PharmID: Pharmacophore Identification Using Gibbs Sampling. *J. Chem. Inf. Model* **2006**, *46*, 1352–1359.
(15) Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein-Protein Interaction. *J. Med. Chem.* **2005**, *48*, 1489–1495.
(16) Labute, P.; Williams, C.; Feher, M.; Sourial, E.; Schmidt, J. M. Flexible Alignment of Small Molecules. *J. Med. Chem.* **2001**, *44*, 1483–1490.
(17) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(18) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. A Molecular-Field-Based Similarity Study of Non-Nucleoside HIV-1 Reverse Transcriptase Inhibitors. 2. The Relationship Between Alignment Solutions Obtained from Conformationally Rigid and Flexible Matching. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 39–51.
(19) Miller, M. D.; Sheridan, R. P.; Kearsley, S. K. SQ: A Program for Rapidly Producing Pharmacophorically Relevent Molecular Superpositions. *J. Med. Chem.* **1999**, *42*, 1505–1514.
(20) 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.
(21) McMartin, C.; Bohacek, R. S. Flexible Matching of Test Ligands to a 3D Pharmacophore Using a Molecular Superposition Force Field: Comparison of Predicted and Experimental Conformations of Inhibitors of Three Enzymes. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 237–250.
(22) *MOE*; Chemical Computing Group: Montreal, Quebec, Canada, 2009.
(23) Chen, Q.; Higgs, R. E.; Vieth, M. Geometric Accuracy of Three-Dimensional Molecular Overlays. *J. Chem. Inf. Model.* **2006**, *46*, 1996–2002.
(24) Hofbauer, C.; Lohninger, H.; Aszódi, A. SURFCOMP: A Novel Graph-Based Approach to Molecular Surface Comparison. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 837–847.
(25) Cosgrove, D. A.; Bayada, D. M.; Johnson, A. P. A Novel Method of Aligning Molecules by Local Surface Shape Similarity. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 573–591.
(26) *ROCS*; OpenEye Scientific Software: Santa Fe, NM, 2005.
(27) Lemmen, C.; Lengauer, T.; Klebe, G. FLEXS: A Method for Fast Flexible Ligand Superposition. *J. Med. Chem.* **1998**, *41*, 4502–4520.

(28) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247*, 536−540.

(29) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.-P.; Bertrand, H.-O. Virtual Screening Workflow Development Guided by the "Receiver Operating Characteristic" Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *J. Med. Chem.* **2005**, *48*, 2534–2547.

(30) Halgren, T. A. MMFF VII. Characterization of MMFF94, MMFF94s, and Other Widely Available Force Fields for Conformational Energies and for Intermolecular-interaction Energies and Geometries. *J. Comput. Chem.* **1999**, *20*, 730–748.

(31) Zou, J.; Xie, H.-Z.; Yang, S.-Y.; Chen, J.-J.; Ren, J.-X.; Wei, Y.-Q. Towards More Accurate Pharmacophore Modeling: Multicomplex-Based Comprehensive Pharmacophore Map and Most-Frequent-Feature Pharmacophore Model of CDK2. *J. Mol. Graphics Modell.* **2008**, *27*, 430–438.

(32) Badger, J.; Minor, I.; Oliveira, M. A.; Smith, T. J.; Rossmann, M. G. Structural Analysis of Antiviral Agents that Interact with the Capsid of Human Rhinoviruses. *Proteins* **1989**, *6*, 1–19.

CI100227H