Predicting the Accuracy of Ligand Overlay Methods with Random Forest Models

Ravi K. Nandigam,[†] David A. Evans,[‡] Jon A. Erickson,[§] Sangtae Kim,[†] and Jeffrey J. Sutherland*[,||]

School of Chemical Engineering, Purdue University, West Lafayette, Indiana, Lilly Research Centre, Erl Wood Manor, Surrey, United Kingdom, and Lilly Research Laboratories and Discovery Informatics, Eli Lilly and Company, Indianapolis, Indiana

Received June 27, 2008

The accuracy of binding mode prediction using standard molecular overlay methods (ROCS, FlexS, Phase, and FieldCompare) is studied. Previous work has shown that simple decision tree modeling can be used to improve accuracy by selection of the best overlay template. This concept is extended to the use of Random Forest (RF) modeling for template and algorithm selection. An extensive data set of 815 ligand-bound X-ray structures representing 5 gene families was used for generating ca. 70,000 overlays using four programs. RF models, trained using standard measures of ligand and protein similarity and Lipinski-related descriptors, are used for automatically selecting the reference ligand and overlay method maximizing the probability of reproducing the overlay deduced from X-ray structures (i.e., using rmsd ≤ 2 Å as the criteria for success). RF model scores are highly predictive of overlay accuracy, and their use in template and method selection produces correct overlays in 57% of cases for 349 overlay ligands not used for training RF models. The inclusion in the models of protein sequence similarity enables the use of templates bound to related protein structures, yielding useful results even for proteins having no available X-ray structures.

INTRODUCTION

Structure-based drug design (SBDD) remains a key component of hit identification and lead optimization activities in drug discovery. In order to effectively apply SBDD, prospective assignment of the bioactive conformation of newly designed compounds must be carried out. This is a challenging task given the various contributions to protein−ligand interaction energies, many of which are difficult to model accurately. The cumulative effect of these interactions results in conformational preferences of ligands that differ markedly from those in solution.[1−3] Molecular docking of ligands to the receptor X-ray structure is the typical method for binding mode prediction. Extensive comparative analysis of docking methods have shown that current methods can typically reproduce bioactive conformations in only 30−60% of cases, when starting from a protein conformation obtained from a different protein−ligand complex (i.e., cross-docking).[4−6]

The number of protein−ligand X-ray structures continues to grow at a tremendous rate, providing a large database of bioactive conformations. A number of molecular docking methods have taken advantage of this data in attempts to improve accuracy.[7−9] Overlay approaches employ the properties of one or more reference ligands in generating a binding mode hypothesis for a compound, allowing their application in the absence of an atomic level structure of the protein.[10,11]

New overlay methods continue to be actively developed: atomic property fields,[12] shape matching,[13] stochastic proximity embedding,[14] receptor-site models,[15] fuzzy pattern recognition,[16] grid-based interaction energies,[17] multiobjective optimization,[18,19] and Gibbs-sampling over multiconformer ligand sets[20] have been described for developing pharmacophore or overlay models. Recently, the ability of two overlay methods to reproduce bioactive conformations as defined by X-ray crystallography has been examined.[21] It was shown that molecular descriptors, in particular overlay vs reference ligand similarity, was a key determinant in overlay accuracy and that a decision tree model could be used to estimate the probability obtaining a "correct" overlay (i.e., an overlay with rmsd within 2 Å of that deduced from X-ray structures). Our recent work indicates that accuracy of binding mode prediction using ROCS can be comparable or superior to that obtained using docking programs, if molecular similarity is used as a basis for selecting the template.[6]

In the current study, we develop random forest models[22] that allow the automatic selection of an overlay method and reference ligand, while estimating the probability of obtaining a correct overlay. The models extend the applicability of ligand-based overlays (using protein-bound reference ligands) to proteins with no available X-ray structures. While we limit ourselves to four commercially available overlay programs (ROCS,[23,24] FlexS,[25] phase,[26] and FieldCompare[27]), the approach is generally applicable to any ligand-based overlay method.

METHODS

Overlay Algorithms. In this work, we examined the performance of ROCS, FlexS, phase, and FieldCompare in reproducing overlays deduced from X-ray structures. ROCS[23,24] involves a rigid body superposition of the overlay ligand onto the reference ligand, optimized using shape and

* Corresponding author phone: (317)655-0833; e-mail: sutherlandje@lilly.com.
[†] Purdue University.
[‡] Lilly Research Centre.
[§] Lilly Research Laboratories, Eli Lilly and Company.
[||] Discovery Informatics, Eli Lilly and Company.

10.1021/ci800216f CCC: $40.75
Published on Web 11/21/2008

Table 1. Composition of X-ray Complex Data Set Used for Ligand Overlay[a]

gene family	number of ligands	number of overlays
all	815 (408)	69915 (34473)
kinase	364 (185)[b]	44401 (21609)
NHR	186 (95)	9439 (4951)
protease	212 (102)	14617 (7187)
PDE	45 (21)	1434 (710)
other	8 (5)	24 (16)

[a] Values in parentheses indicate the number of ligands used for testing random forest models; only overlay pairs for which both ROCS and FlexS produce solutions are retained in this work; ROCS failed to produce an overlay for 879 pairs (202 pairs involving ligands of the same protein), and FlexS failed for 25526 pairs (7528 pairs involving ligands of the same protein); most FlexS failures arise due to its inability to generate a base fragment. [b] 299 kinase ligands (155 test set) excluding ATP and analogs.

Table 2. Percentage of Overlays within 2 Å of X-ray Overlay[a]

	ROCS	FlexS	phase	Field Compare
Reference and Overlay Ligand from Same Protein				
all	15 (22368)	16 (22368)	6 (16213)	9 (22285)
kinase	15 (9978)	15 (9978)	8 (6923)	11 (9931)
kinase non-ATP	16 (8949)	16 (8949)	9 (6499)	12 (8931)
NHR	37 (1701)	45 (1701)	31 (1007)	36 (1697)
protease	12 (10249)	11 (10249)	1 (7966)	3 (10220)
PDE	21 (424)	21 (424)	20 (305)	25 (421)
other	50 (16)	44 (16)	8 (12)	19 (16)
Reference and Overlay Ligand from Different Protein within 25% Seq. Identity				
all	5 (47547)	6 (47547)	3 (31003)	4 (47267)
kinase	3 (34423)	3 (34423)	2 (23734)	2 (34194)
kinase non-ATP	3 (27955)	3 (27955)	2 (20431)	3 (27874)
NHR	14 (7738)	19 (7738)	15 (3154)	14 (7707)
protease	1 (4368)	3 (4368)	1 (3288)	2 (4354)
PDE	12 (1010)	11 (1010)	10 (819)	12 (1004)
other	0 (8)	0 (8)	0 (8)	0 (8)

[a] Values in parentheses indicate the total number of overlays.

chemical (or "color") complementarity of the ligands. To allow for conformational flexibility, an ensemble of conformers is generated with OMEGA[2,23] and scored with ROCS, retaining the conformer with the highest score.

FlexS[25] takes a representative 3D conformer of the overlay ligand and disassembles it into several fragments. A base fragment is chosen using heuristic rules and optimally placed over the reference ligand. FlexS then performs a recursive construction of the overlay ligand by adding new fragments to the subsection of the ligand already placed over the template.

Phase builds a pharmacophore model using one or more template ligands.[26] In this work, the pharmacophore consists of features in the reference ligand only (the program is often used to develop pharmacophore models from multiple active molecules;[28] however, we did not explore the use of multiple references in this work). If the overlay ligand (i.e., scoring a new molecule against the pharmacophore model) satisfies the number and type of each of the pharmacophore site of the hypothesis, its conformational space is searched for a 3D structure where the pharmacophore sites satisfy the spatial constraints imposed by the hypothesis. A user-defined parameter is the number of pharmacophore features to consider, and 3 overlays were performed for each reference-overlay ligand pair: matching all features present in the reference ligand, 5 features or 3 features (otherwise, default parameters were used). The overlay with the highest number of matched features was retained; approximately 6% of overlays used all features, 16% used 5 features and the remainder used 3 features.

Table 3. Significance of Relationships between Molecular Properties and Overlay Accuracy[a]

	NROT	HDON	HACC	ClogP	MW	DaySim
ROCS						
kinase	*<0.0001*	*<0.0001*	*<0.0001*	<0.0001	*<0.0001*	<0.0001
NHR	0.012	<0.0001	*0.026*	0.0003	0.0002	<0.0001
protease	*<0.0001*	*<0.0001*	0.0046	<0.0001	<0.0001	<0.0001
PDE	0.0004	*0.02*	*0.0002*	0.001	0.33	<0.0001
FlexS						
kinase	*<0.0001*	*<0.0001*	*<0.0001*	<0.0001	*0.0008*	<0.0001
NHR	*0.14*	<0.0001	*0.39*	*0.42*	<0.0001	<0.0001
protease	*<0.0001*	*<0.0001*	*0.0009*	0.0026	*0.14*	<0.0001
PDE	0.42	*0.0002*	*<0.0001*	0.0037	*0.0017*	<0.0001
Phase						
kinase	*<0.0001*	*<0.0001*	*<0.0001*	<0.0001	*<0.0001*	<0.0001
NHR	*<0.0001*	*0.27*	*0.05*	*<0.0001*	*<0.0001*	<0.0001
protease	*<0.0001*	*0.0017*	*<0.0001*	0.0063	*<0.0001*	<0.0001
PDE	0.60	*0.63*	*0.0005*	0.16	*0.69*	0.03
FieldCompare						
kinase	*<0.0001*	*<0.0001*	*<0.0001*	<0.0001	*<0.0001*	<0.0001
NHR	*<0.0001*	0.0206	*0.0021*	*<0.0001*	*0.27*	<0.0001
protease	*<0.0001*	*<0.0001*	*0.42*	0.001	0.20	<0.0001
PDE	0.0052	*0.0005*	<0.0001	0.0001	*0.39*	<0.0001

[a] Values in table are P-values indicating the significance of each molecular property in a single parameter (plus intercept) logistic regression model, where the value fit is 1 for successful overlay, 0 otherwise; P-values less than or equal to 0.05 indicate with 95% certainty that the effect of the parameter does not arise by chance. Italicized values indicate a decrease in overlay accuracy with increasing value of the parameter. Overlays using templates bound to a different protein are excluded.

The FieldCompare program aligns two compounds based on molecular field similarity. Electrostatic, van der Waals, and hydrophobic fields are projected into the space surrounding each molecule, allowing two compounds to be aligned based on their potential interactions with a receptor, rather than by direct comparison of their atomic coordinates.[27] FieldCompare was used in a similar way to ROCS, in that an ensemble of conformers of the overlay compound was generated using the xedex program and the XED force field,[29,30] before being aligned onto the template compound in the reference conformation.

Data Set. A data set of 815 protein−ligand complexes from gene families of therapeutic interest was assembled from the PDB.[31] Only complexes having resolution better than 3.5 Å were selected, and complexes having protein−ligand contacts less than 1.75 Å were rejected (i.e., covalently bound ligands, or poorly refined complexes) (Table 1).

For each of 815 protein−ligand structures, other complexes containing proteins with sequence identity greater than 25% were identified (determined with the BLAST program[32] bl2seq, using a BLAST database of protein sequences converted from residue labels in PDB files). Approximately 70,000 overlays were obtained with ROCS, FlexS, phase, and FieldCompare, using as references each of the 815 ligands in turn and overlaying ligands extracted from complexes identified via BLAST searches.

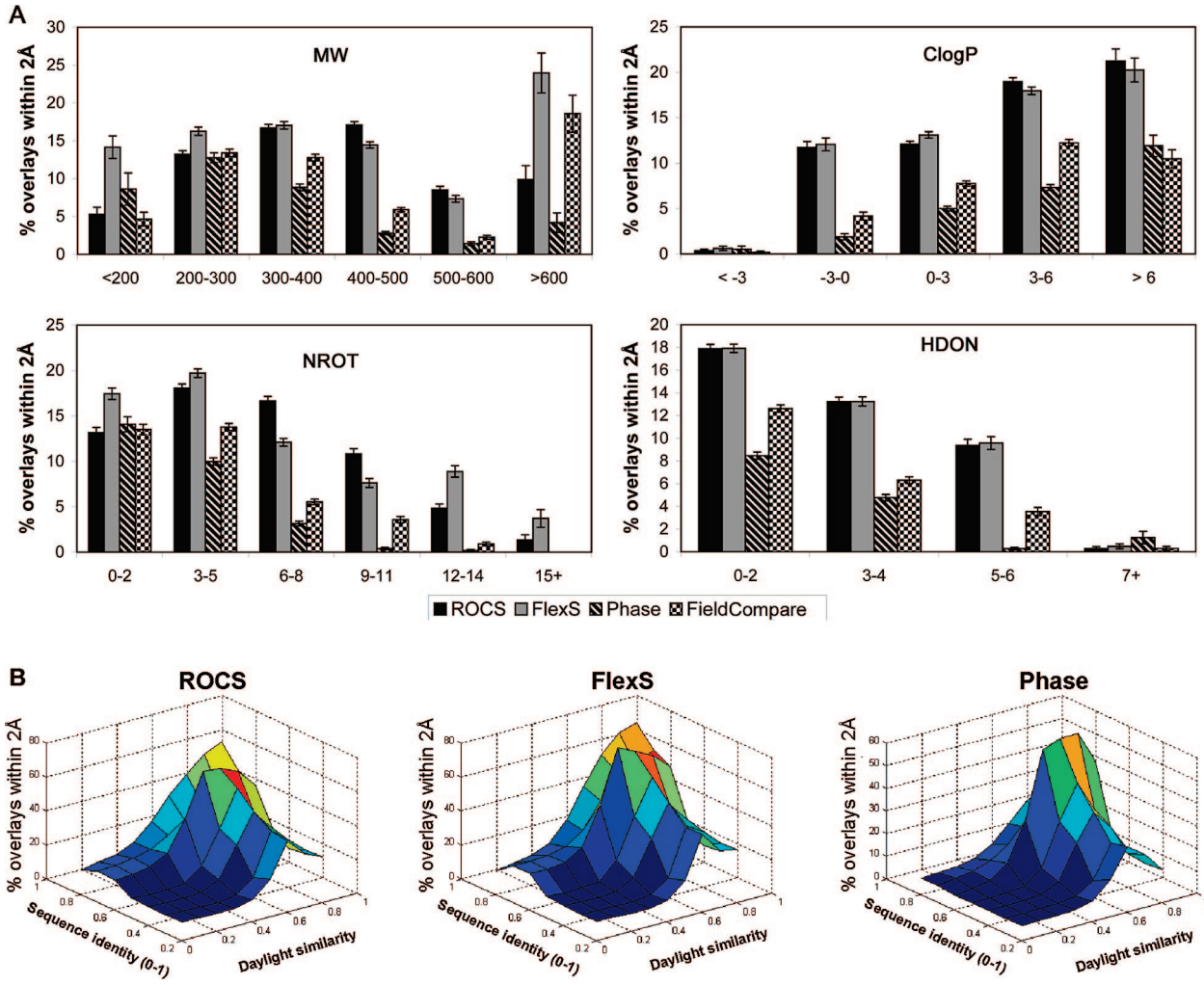

Figure 1. Relationships between molecular overlay accuracy for ROCS, FlexS, phase, and FieldCompare vs (A) molecular weight (MW), calculated logP (ClogP), rotatable bonds (NROT), and hydrogen bond donors (HDON) or (B) Daylight similarity of reference and overlay ligand and sequence identity (on a scale of 0−1) for proteins binding reference and overlay ligands. Profiles in (A) do not include overlays using templates from a different protein; profiles in (B) are obtained by binning overlay accuracy vs Daylight similarity and sequence identity into ranges of 0.1 and smoothed by averaging each point on the 3D grid over adjacent bins.

Ligand Preparation. Bond orders were added automatically to ligands in PDB format (without CONNECT records) using Maestro[33] and saved as SDF files. Bond orders were examined and manually fixed where appropriate. In order to remove any bias in overlay calculations, SDF files for overlay ligands were converted to Smiles strings (detecting chirality from the 3D arrangement of atoms in the SDF file) and converted back to 3D structures in SDF format using Corina.[34] Reference ligands were retained in the X-ray conformation. A standard scheme was applied for assigning formal charges (deprotonated acids, protonated amines, etc.). Where applicable, the tautomer observed in the X-ray structure (rationalized via hydrogen bonds to the protein) is used for producing overlays. For virtual-screening applications where the appropriate tautomer cannot be inferred *a priori*, it would be necessary to generate all energetically accessible tautomers and retain the highest-scoring vs the reference ligand.

Defining Successful Overlays from X-ray Complexes. For the evaluation of overlay accuracy, it is necessary to put each pair of structures into a common reference frame (i.e., the complex containing the ligand that will be overlaid, and the complex which provides the reference ligand). This is accomplished by the superposition of complete protein domains or active sites from each structure. Our internal organization of complexes uses a selected template for each protein, superposing other structures onto the template using the program LSQMAN[35] (a structure-based superposition algorithm which uses 3D positions of C-α atoms in the complete protein domain).

Throughout this work, "correct" overlays indicate those having heavy-atom root-mean-square deviations (rmsd) ≤ 2 Å compared to the X-ray overlay, in keeping with the usual convention used for docking studies.

Random Forest Models. Random Forest models consist of an ensemble of decision trees, each obtained by splitting object collections until terminal nodes contain only objects of the same class.[22] Models were trained using a number of computed properties describing an overlay pair (i.e., reference and overlay ligand), with the objective of modeling whether a given ligand is correctly fit to the reference (class 1) or not (class 0). Overlay pairs are described using reference ligand vs overlay ligand similarities (using Daylight fingerprints[36] hashed to 1024 bits using default path lengths), the sequence identity from a local sequence alignment, and standard molecular properties: heavy atoms counts (NA), molecular weight (MW), calculated logP (ClogP), H-donors (HDON), H-acceptors (HACC), polar surface area (PSA),

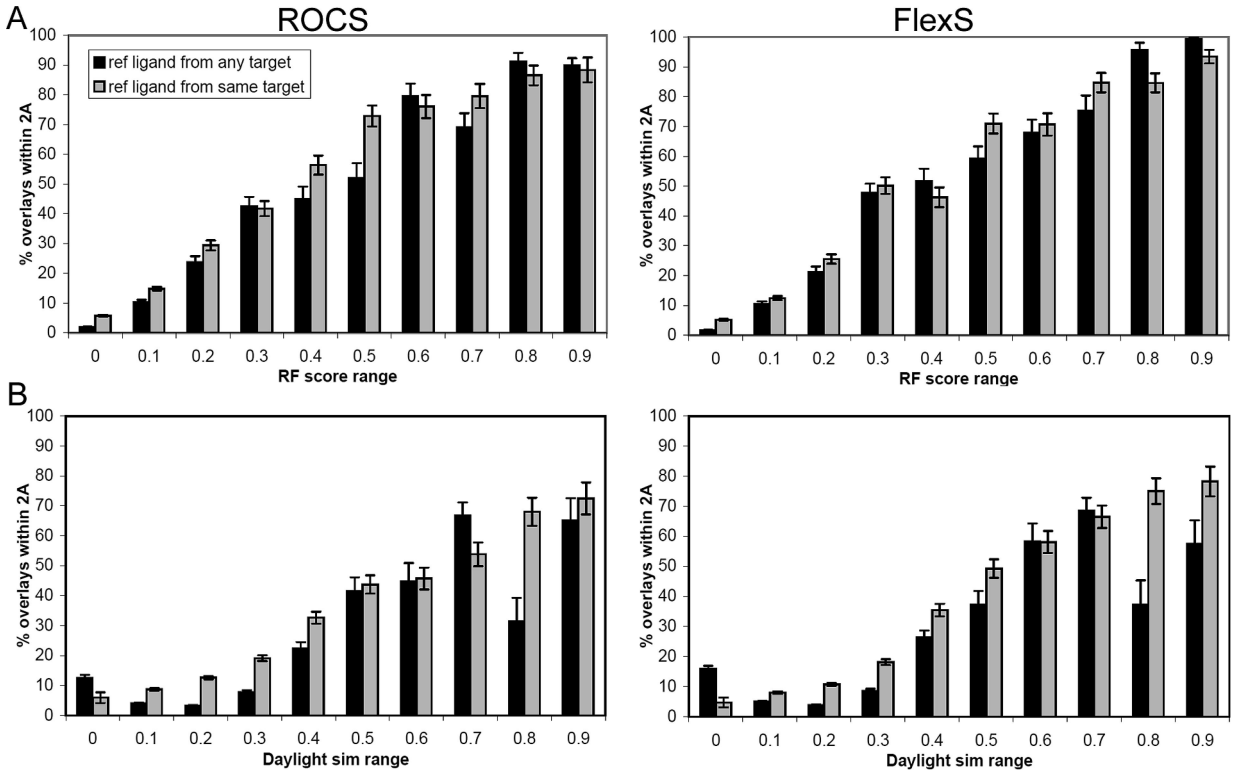

Figure 2. Test set overlay accuracy averaged over (A) RF score binned into ranges of 0.1 or (B) Daylight similarity binned into ranges of 0.1 for ROCS (left) and FlexS (right), e.g. the range with label 0.5 indicates average overlay accuracy for overlay-reference ligand pairs having RF scores between 0.5 and 0.6 in panel A and those with Daylight similarity between 0.5 and 0.6 in panel B. Black bars denote the selection of reference ligands from any protein receptor having 25% sequence identity or greater compared to the protein receptor for the overlay ligand, while gray bars denote selection of reference ligands bound to the same protein receptor only. Overlays of ATP and analogs bound to kinases are excluded (see Figure S.2 in the Supporting Information for their inclusion).

Table 4. Overlay Accuracy for Kinase Ligands When Using Daylight Similarity for Choosing Templates[a]

overlay pairs[b]	no. of overlays	% within 2 Å (std. error)			
		ROCS	FlexS	phase	Field Compare
ATP, diff kinases	800	9 (1)	15 (1)	3 (1)	2 (1)
ATP, same kinase	25	16 (7)	16 (7)	9 (6)	8 (5)
non-ATP, diff kinases	89	39 (5)	49 (5)	38 (5)	51 (5)
non-ATP, same kinase	142	72 (4)	77 (4)	50 (4)	65 (4)

[a] Only reference-overlay ligand pairs with Daylight similarity between 0.8 and 0.95 are included. [b] ATP or non-ATP indicates the nature of the ligand serving as reference; same vs different kinase indicates whether or not the template and overlay ligands are complexed to the same kinase.

and rotatable bond count (NROT). Overlay vs reference ligand differences, absolute values of differences, and ratios were calculated for the molecular properties (with the exception of HDON, HACC, and NROT, since some ratios will be indeterminate due to zero counts for these features in certain ligands). For each protein, 815 complexes were divided into 407 training and 408 test overlay ligands; all pairs involving training ligands become training pairs, and those involving test ligands become test pairs (e.g., 10 complexes for BACE1 are divided into 5 training and 5 test overlay ligands; 40 pairs in which the training ligand was overlaid are training pairs, the remaining 37 pairs are test pairs). This gave 35442 training and 34473 test pairs for ROCS and FlexS, 24913 training/22303 test pairs for phase, and 35261 training/34291 test pairs for FieldCompare. No overlay ligand in a training pair is used as an overlay ligand in a test pair. For each reference vs overlay ligand pair, a RF score was obtained by calculating the fraction of 1000 trees in which the overlay deduced from X-ray structures is reproduced (i.e., fraction of trees with a class 1 prediction).

Calculation of Standard Errors. For values with continuous distributions, standard errors are calculated from the relationship

$$\text{standard error} = \text{standard deviation of property}/\sqrt{N}$$

for success/failure properties (i.e., overlay within 2 Å rmsd or not), standard errors may be calculated using the standard deviation from Bernoulli trials

$$\text{standard error} = \sqrt{((P(1-P)/N)}$$

where P is the probability of success, and N is the number of samples.

RESULTS AND DISCUSSION

The programs ROCS, FlexS, phase, and FieldCompare were used to produce overlays for 815 ligands from the PDB, using as a reference each complex having >25% sequence identity calculated over the ligand-binding domain. This resulted in 69915 pairs for FlexS and ROCS, 47216 pairs for phase, and 69552 pairs for FieldCompare. As expected, overlays using a reference from a different protein have a lower probability of being within 2 Å of the overlay deduced from X-ray structures (Table 2). While ligands bound to nuclear hormone receptors have the highest incidence of "correct" overlays, overall accuracy is generally low.

Relationships between binding mode prediction and Lipinksi properties have been noted for docking[6,37] and ligand

Table 5. Comparison of Overlay Accuracy for Reference Selection by RF Score vs Daylight Similarity, Excluding ATP and Analogs

ref ligand vs overlay ligand relationship	ref. selection by RF score		ref. selection by Daylight sim[a]	
	no. of overlays	% within 2 Å (std. error)	no. of overlays	% within 2 Å (std. error)
Training Overlays (Out-of-Bag Sample[b])				
same protein[c]	275	55 (3)	246	61 (3)
different protein, ref from same protein available	68	59 (6)	97	35 (5)
restrict to refs from same protein	68	54 (6)	97	45 (5)
no refs from same protein available	28	25 (8)	28	25 (8)
Test Overlays				
same protein	274	58 (3)	233	61 (3)
different protein, ref from same protein available	75	52 (6)	116	34 (4)
restrict to refs from same protein	75	51 (6)	116	38 (5)
no refs from same protein available	29	28 (8)	29	21 (8)

[a] The reference ligand used for the overlay is selected by Daylight similarity, but the method (ROCS, FlexS, phase, or FieldCompare) is selected using RF scores since all 4 overlays have the same Daylight similarity. [b] Overlay accuracy assessed using RF out-of-bag samples, i.e. overlays not selected for training models in sampling with replacement process. [c] A reference ligand is selected by RF score or Daylight similarity among all complexes within 25% sequence identity, and the proteins happen to be the same for both ligands. For numerically equal scores or similarity (rounded to 2 decimals), a reference from the same protein as the overlay ligand is preferred over a reference from a different protein.

overlay methods.[21] In this work, we examined overlay accuracy vs molecular properties of overlay ligands, as a function of gene family and overlay algorithm (Table 3, Figure 1, Figure S.1 in the Supporting Information for HACC). Overlay accuracy generally decreases as the number of rotatable bonds increases. This relationship is statistically significant for all gene families except PDE ligands, which have at most 9 rotatable bond, and ROCS or FlexS vs NHRs. ROCS performs somewhat better for ligands having 6−11 rotatable bonds and somewhat worse for ligands outside this range. Overlay accuracy generally decreases vs increasing donors or acceptors and increases with increasing ClogP; ROCS and FlexS perform similarly for these three properties. Overlay accuracy vs molecular weight is highest for ligands in the 300−500 Dalton range (which contains 58% of overlays), with FlexS performing substantially better than ROCS for very high and low MW ligands. The unexpected increase in accuracy for FlexS for >600 MW bin arises due to THRB (thyroid hormone receptor beta), for which FlexS correctly predicts 24 X-ray overlays vs 8/24 for ROCS. While many trends are observed across different gene families and overlay methods, there are exceptions for each molecular property.

The relationship between the similarity of ligands and overlay accuracy is intuitive and has been noted in previous studies. This led us to examine the utility of employing reference ligands from a different protein and the impact of sequence similarity between the two proteins on overlay accuracy (i.e., the protein of interest, and related proteins having X-ray ligands that can serve as references). Overlay accuracy was calculated for each range of ligand and sequence similarity (henceforth, sequence similarity refers to identity/100) (Figure 1). There is a general increase in overlay accuracy with increasing sequence and ligand similarity, although the relationships are not strictly monotonic. Thus, choosing reference ligands from different but sequence-related proteins may be useful in some instances, even for relatively low similarity protein pairs.

There are 1249 overlay pairs involving reference and overlay ligands with greater than 0.95 Daylight similarity (663 training pairs). Most of these involve overlays of ATP or ADP bound to the same (116 overlays) or different (567 overlays) kinases; 235 pairs are for NHR ligands, many of which are steroid analogs. Given the artificial overlay problem they represent (and the fact that most of these are for ATP, a nondrug-like ligand of low affinity), they are excluded from accuracy calculations below.

While property-based analysis can help understand factors influencing overlay accuracy, establishing quantitative relationships between molecular properties of ligands and overlay accuracy would allow the identification of the reference ligand most likely to yield a correct overlay. Random forest models have been widely used for QSAR applications, relating molecular attributes to various biological end points. To capture complex relationships between molecular properties and overlay accuracy, we have sought to develop a model capable of predicting the probability of obtaining a correct overlay for a given pair of ligands, one serving as reference in the overlay (i.e., an overlay pair). Each overlay pair is described using the Daylight similarity of the ligands, sequence similarity of the proteins, molecular properties of reference and fit ligand (heavy atoms, MW, ClogP, H-donors, H-acceptors, polar surface area, and rotatable bond count), and various transformations of molecular properties (Methods). Approximately half of overlay pairs were used for training RF models for each overlay method (i.e., one model for each of ROCS, FlexS, phase, and FieldCompare). The other half were used for assessing relationships between overlay accuracy and RF model scores. The RF model score for a given overlay pair indicates the average purity (or fraction of pairs with class label '1') of terminal nodes containing the pair, calculated over all trees in the ensemble. The most highly used variables in the RF models are Daylight and sequence similarity and various ClogP-related measures (Supporting Information Table S.2).

The relationship between RF scores and overlay accuracy indicates that the former is a useful metric for estimating the probability that a given overlay is within 2 Å of the X-ray overlay, with few significant differences between methods for a given RF score range (Figure 2). When restricting overlay pairs to those using reference and overlay ligands from the same protein, the probability of obtaining a correct

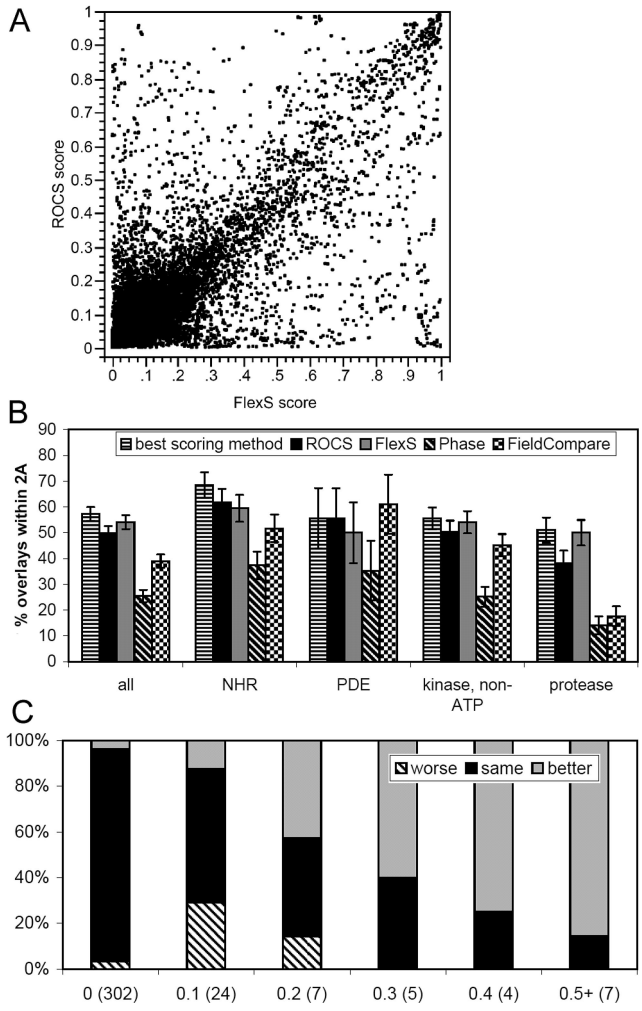

Figure 3. Using RF scores to select a reference ligand and overlay method most likely to generate a correct overlay. (A) Comparison of RF scores on test overlay pairs; ROCS vs FlexS $\rho = 0.88$, $N = 33887$; panels for phase and FieldCompare are shown in Figure S.4 in the Supporting Information); the Spearman's rho rank-order coefficient is used due to non-normal distribution of scores. (B) Overlay accuracy obtained by selecting the reference ligand (from any protein) and method which maximizes the RF score, averaged over 349 test set ligands, excluding ATP and analogs; FlexS/ROCS/Phase/FieldCompare provides the highest-scoring overlay for 195/111/13/30 ligands. (C) Differences between the method with the best RF score and FlexS binned into ranges vs improvements in overlay accuracy (i.e., "same" indicates that both are correct or incorrect, "better" indicates that the higher-scoring method is correct and FlexS is incorrect). The number of test overlay ligands falling into each range is indicated in parentheses.

overlay is slightly increased (<9%) for a given RF score range, although most differences are not statistically significant. Only a small fraction of overlays are in moderate or high accuracy RF score ranges: 4.5% (ROCS), 5.5% (FlexS), 3.2% (Phase), 3.4% (FieldCompare) of pairs with RF score ≥0.4 for overlays using references from any protein and 7.1% (ROCS), 8.5% (FlexS), 3.5% (Phase), 4.6% (FieldCompare) using references from the same protein only (a RF score ≥0.4 corresponds to ∼50% or higher probability of obtaining an accurate overlay).

In addition, we examined the relationship between the Daylight similarity of reference and overlay ligands as a simpler approach for quantifying the probability of obtaining correct overlays. For overlay pairs using ligands from the same protein, there is a monotonic increasing relationship

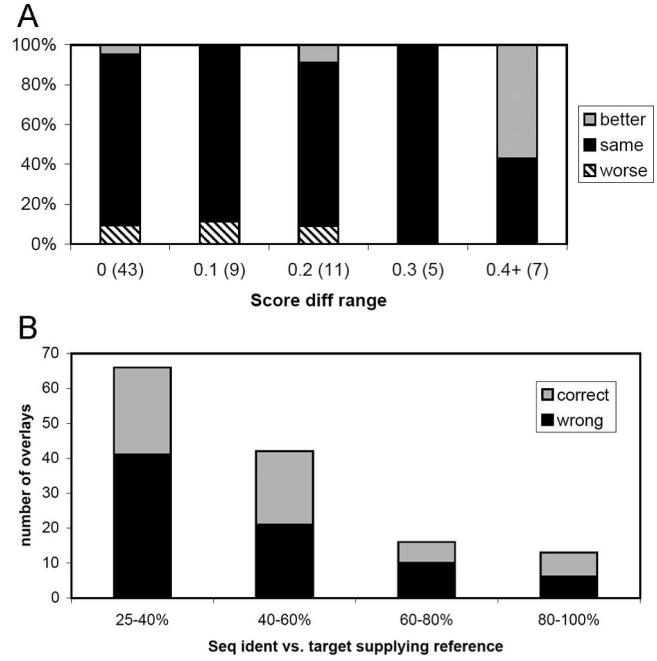

Figure 4. Reference ligand selection by RF score, where the reference ligand is bound to a different receptor than the overlay ligand: (A) RF score difference binned into ranges of 0.1, calculated by subtracting the highest scoring overlay for which both ligands are bound to the same receptor from the highest scoring overlay using a different protein (values in parentheses indicate the number of overlays in each range). (B) Sequence identity of protein from reference complex vs protein bound to overlay ligand, indicating fraction overlaid within 2 Å (gray).

between Daylight similarity and overlay accuracy (Figure 2). However, this relationship is not demonstrated for high-Daylight similarity overlay pairs when the reference ligand is bound to a different protein. Closer inspection reveals that the unexpected decrease in accuracy for high Daylight similarity (≥0.8) arises entirely from kinases, in particular pairs involving ATP and analogous ligands (Table 4). Most of these pairs have low RF scores (results not shown), indicating that overlays involving these ligands allowed their features to be associated with low overlay accuracy. ATP is a low affinity, nondruglike ligand, and rmsd computations are somewhat artificial because of the solvent-exposed triphosphate group which assumes many orientations in X-ray complexes (the second most frequently occurring kinase ligand is staurosporine, which is present in 7 complexes vs 65 for ATP and analogs).

Despite the exclusion of these pairs, only 53% of the remaining kinase overlays are accurate, when using a template with high Daylight similarity bound to a different protein. While it is an important variable in the RF models, other factors besides Daylight similarity account for their utility in quantifying the probability of obtaining correct overlays, especially when the reference is not in complex with the same protein. In addition, RF scores identify a higher fraction of overlays with ca. 50% or higher probability of being correct. For FlexS overlays using references from the same protein, 8.5% have RF score ≥0.4 vs 4.6% having Daylight similarity ≥0.5; both thresholds correspond to approximately 50% overlay accuracy in Figure 2. Another approach for quantifying differences in predictive accuracy employs logistic regression for fitting a response (correct or incorrect overlay) vs RF score or Daylight similarity. The

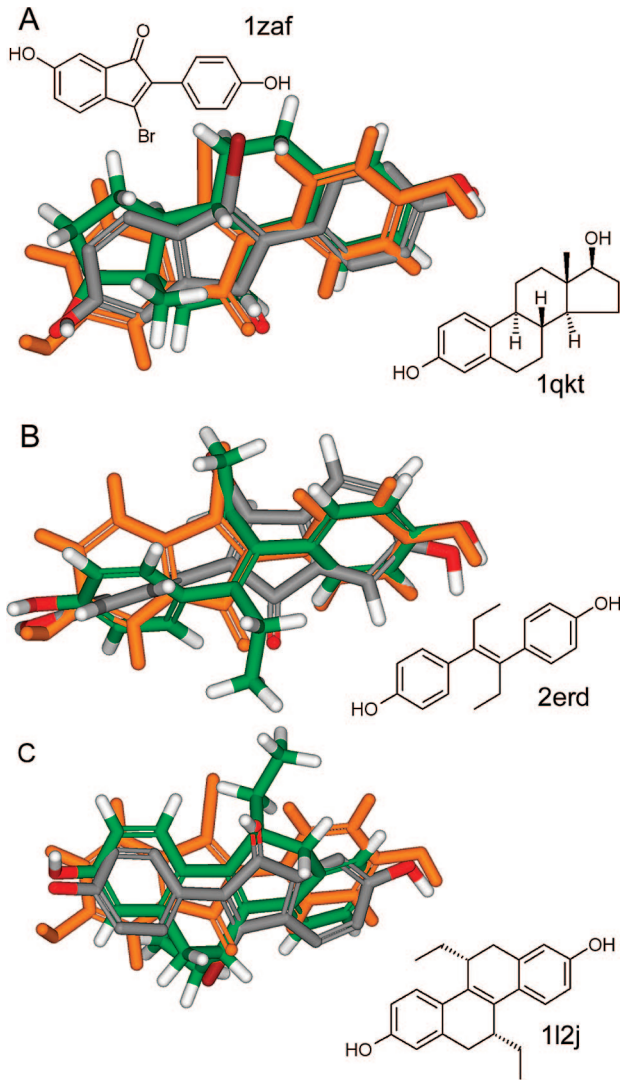

Figure 5. Overlay of 3-bromo-6-hydroxy-2-(4-hydroxyphenyl)inden-1-one (PDB 1zaf) bound to ESR2, which is overlaid by FlexS with rmsd of 0.9 Å when selecting by RF score a template from PDB 1qkt, an ESR1 complex (panel A). An overlay obtained using ROCS and the most similar ligand by Daylight similarity (PDB 2erd, an ESR1 complex) has rmsd of 6.5 Å (panel B), and that obtained with FlexS selecting a template by RF score from complexes for the same protein has rmsd of 6.9 Å (panel C). The method with maximum RF score is used to perform the overlay in all 3 cases. Carbon atoms in the templates are shown in green, and the X-ray pose from 1zaf is shown in orange.

area under the curve (AUC) for a receiver-operator characteristic (ROC) plot, which ranges from 0 (not predictive) to 1 (fully predictive) is 0.85 for RF scores and 0.73 for Daylight similarity, for FlexS test set overlays using templates bound to the same protein.

Given a ligand to overlay, scores from RF models can be used to select a reference ligand (in its X-ray conformation) from all available protein−ligand complexes for which the protein shares homology with the protein of interest. In addition, the overlay method most likely to correctly overlay a ligand can be chosen based on maximizing RF scores, since all four methods have a similar probability of generating correct overlays within a given RF score range (Figure S.3 in the Supporting Information). RF scores for different methods are moderately correlated (Figure 3). Using the highest scoring method results in overlay accuracy equal to or higher than that obtained using the best method for each gene family, although the differences are not statistically significant (Figure 3). Nevertheless, the RF models capture the variable performance of overlay methods across gene families (ROCS superior for NHRs and PDEs; FlexS for the other families; differences are not statistically significant). Larger RF score differences between the best method and any given individual method correspond to a higher fraction of improved overlay results.

Among the four methods studied, phase gives consistently lower overlay accuracy across all gene families. Using a pharmacophore method for this purpose requires ignoring all template atoms which are not contained in a pharmacophore feature. Further, we also allow partial matches of overlay ligand features vs the reference ligand, meaning not all template pharmacophore sites must be matched. A metric other than rmsd, such as reproducing protein−ligand contacts observed for both ligands, may be more suitable for evaluating success in this case.

This work has described the application of RF models to identify reference ligands and overlay methods that maximize the probability of obtaining a correct overlay (and quantify the probability). The inclusion of reference ligands from different proteins and the use of RF scores rather than Daylight similarity to quantify the above probability render the approach more complex than simpler alternatives. To understand any benefits arising from the use of RF models, we examined the percentage of test ligand overlays using a reference from a different protein and what the corresponding accuracy would be if selecting a reference ligand by RF score among those in complex to the *same* protein. Of 349 test ligand overlays having at least one reference from the same protein (excluding ATP and analogs), 21% employed a reference from a different protein, and 52% of these resulted in correct overlays. When restricting the selection of reference ligands to those cocrystallized with the same protein, 51% of overlays are correct, although the decrease in accuracy is not statistically significant (Table 5). As observed with method selection, larger differences in RF score between the best reference ligand from a different protein vs the same protein results in larger overlay accuracy improvements (Figure 4). Reference ligands from proteins having lower sequence similarity are used more frequently than those from higher similarity proteins (Figure 4), although this may simply reflect the larger number of references from lower similarity proteins.

In lieu of RF scores, one can select the reference ligand having the highest Daylight similarity compared to the overlay ligand. Using this approach, 67% of 349 test ligand overlays use a reference from the same protein, of which 61% are correctly overlaid; those using a reference from a different protein achieve 34% accuracy, which increases slightly to 38% when restricting template selection to ligands complexed with the same protein. Selecting a reference by Daylight similarity is comparable to selection by RF score when the reference is bound to the same protein and significantly worse than using RF scores when bound to a different protein. Figure 5 exemplifies a correct overlay for ESR2 (estrogen receptor beta) obtained by choosing a template by RF score from ESR1 complexes; it is not correctly overlaid by using either the most similar ligand across homologous proteins or by restraining the set of

templates to those from ESR2. Table S.1 indicates other successful overlays obtained using RF scores.

The use of either RF scores or Daylight similarity allowing the selection of a reference ligand from different proteins allows the application of molecular overlays to proteins having no available protein−ligand complexes (i.e., a situation which would require homology modeling or use of an apo structure for docking applications). Our test set contains 29 non-ATP ligands for which there were no other complexes for the same protein. Overlays that used RF scores for identifying a reference achieved an accuracy of 28% vs 21% when using Daylight similarity. Cross-docking studies frequently indicate lower accuracy when using apo structures,[4−6] suggesting that molecular overlay remains useful even in the absence of X-ray structures for a protein of interest, if other sequence-related complexes are available.

CONCLUSION

The field of ligand-based overlay or pharmacophore detection methods remains an area of active research. In contrast to docking approaches, relatively few larger scale studies have examined the ability of ligand-based overlay methods to reproduce overlays deduced from X-ray structures. Since overlays are frequently used for elucidating or comparing further modifications to actives in hit/lead optimization, understanding the uncertainty associated with a given overlay is important. The present work has described an approach using random forest models for quantifying this uncertainty and allows the automatic selection of an overlay algorithm and reference template most likely to reproduce results from X-ray crystallography. By incorporating the use of reference ligands from other proteins, the wealth of X-ray data available in public and proprietary repositories can be applied to a larger range of protein targets and provides binding mode predictions that compares favorably with those obtained from docking programs.

The selection of the random forest method for modeling overlay accuracy was governed by availability of software and its favorable comparison to other methods from other internal work and the literature. A key advantage of RF models is the class probability score obtained, which we have used for understanding the probability of obtaining accurate overlays. Our general experience indicates that different methodologies yield comparable results, especially when applied to large data sets. Other modern machine learning approaches would be expected to give similar results.

As Boström and co-workers have observed, binding modes are better conserved across structurally related ligands than are side-chain orientations and water network architecture.[38] Through the interpretation of binding modes from overlay programs in the protein environment, the approach might be extended to identify locations of bound water molecules, in particular those mediating protein−ligand contact.

ACKNOWLEDGMENT

R.N. was supported by a Research Assistantship from the Donald W. Feddersen Professorship funds for S.K. We thank members of the Lilly Global Computational Chemistry Group for suggestions and review of this manuscript.

Supporting Information Available: Parameters/methodology used for ROCS, FlexS, phase, and FieldCompare, a list of PDB codes for the complexes used in this work, and supplementary figures and tables indicated in the text. This material is available free of charge via the Internet at http://pubs.acs.org

REFERENCES AND NOTES

(1) Bostrom, J.; Norrby, P. O.; Liljefors, T. Conformational energy penalties of protein-bound ligands. *J. Comput.-Aided. Mol. Des.* **1998**, *12*, 383–96.
(2) Bostrom, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graphics Modell.* **2003**, *21*, 449–62.
(3) Perola, E.; Charifson, P. S. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J. Med. Chem.* **2004**, *47*, 2499–510.
(4) Murray, C. W.; Baxter, C. A.; Frenkel, A. D. The sensitivity of the results of molecular docking to induced fit effects: application to thrombin, thermolysin and neuraminidase. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 547–62.
(5) Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J. Med. Chem.* **2004**, *47*, 45–55.
(6) Sutherland, J. J.; Nandigam, R. K.; Erickson, J. A.; Vieth, M. Lessons in molecular recognition. 2. Assessing and improving cross-docking accuracy. *J. Chem. Inf. Model.* **2007**, *47*, 2293–302.
(7) Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499–511.
(8) Marialke, J.; Tietze, S.; Apostolakis, J. Similarity based docking. *J. Chem. Inf. Model.* **2008**, *48*, 186–96.
(9) Wu, G.; Vieth, M. SDOCKER: a method utilizing existing X-ray structures to improve docking accuracy. *J. Med. Chem.* **2004**, *47*, 3142–8.
(10) Lemmen, C.; Lengauer, T. Computational methods for the structural alignment of molecules. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 215–32.
(11) Wolber, G.; Seidel, T.; Bendix, F.; Langer, T. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug. Discovery Today* **2008**, *13*, 23–9.
(12) Totrov, M. Atomic property fields: generalized 3D pharmacophoric potential for automated ligand superposition, pharmacophore elucidation and 3D QSAR. *Chem. Biol. Drug Des.* **2008**, *71*, 15–27.
(13) Ebalunode, J. O.; Ouyang, Z.; Liang, J.; Zheng, W. Novel Approach to Structure-Based Pharmacophore Search Using Computational Geometry and Shape Matching Techniques. *J. Chem. Inf. Model.* **2008**, *48*, 889–901.
(14) Bandyopadhyay, D.; Agrafiotis, D. K. A self-organizing algorithm for molecular alignment and pharmacophore development. *J. Comput. Chem.* **2008**, *29*, 965–82.
(15) Todorov, N. P.; Alberts, I. L.; de Esch, I. J.; 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*, 1007–20.
(16) Nettles, J. H.; Jenkins, J. L.; Williams, C.; Clark, A. M.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Flexible 3D pharmacophores as descriptors of dynamic biological space. *J. Mol. Graphics Modell.* **2007**, *26*, 622–33.
(17) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for Ligands and Proteins (FLAP): theory and application. *J. Chem. Inf. Model.* **2007**, *47*, 279–94.
(18) Richmond, N. J.; Abrams, C. A.; Wolohan, P. R.; 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*, 567–87.
(19) Cottrell, S. J.; Gillet, V. J.; Taylor, R. Incorporating partial matches within multi-objective pharmacophore identification. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 735–49.
(20) Feng, J.; Sanil, A.; Young, S. S. PharmID: pharmacophore identification using Gibbs sampling. *J. Chem. Inf. Model.* **2006**, *46*, 1352–9.
(21) Chen, Q.; Higgs, R. E.; Vieth, M. Geometric accuracy of three-dimensional molecular overlays. *J. Chem. Inf. Model.* **2006**, *46*, 1996–2002.
(22) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–58.

(23) *ROCS*, *version 2.2*; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2006.
(24) Rush, T. S., III.; 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*, 1489–95.
(25) Lemmen, C.; Lengauer, T.; Klebe, G. FLEXS: A method for fast flexible ligand superposition. *J. Med. Chem.* **1998**, *41*, 4502–4520.
(26) Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A. PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 647–71.
(27) 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*, 665–76.
(28) Evans, D. A.; Doman, T. N.; Thorner, D. A.; Bodkin, M. J. 3D QSAR methods: Phase and Catalyst compared. *J. Chem. Inf. Model.* **2007**, *47*, 1248–57.
(29) Vinter, J. G. Extended electron distributions applied to the molecular mechanics of some intermolecular interactions. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 653–68.
(30) Vinter, J. G. Extended electron distributions applied to the molecular mechanics of some intermolecular interactions. II. Organic complexes. *J. Comput.-Aided Mol. Des* **1996**, *10*, 417–26.
(31) 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*, 235–42.
(32) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–402.
(33) *Maestro*, *version 7.5*; Schrodinger Inc.: New York, NY, 2006.
(34) *Corina*, *version 3.2*; Molecular Networks GmbH: Erlangen, Germany, 2005.
(35) *LSQMAN*, *version 9.6.2*; Uppsala Software Factory: Uppsala, Sweden, 2005.
(36) *Daylight*, *version 4.8.1*; Daylight Chemical Information Systems, Inc.: Aliso Viejo, CA, 2002.
(37) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **2004**, *56*, 235–49.
(38) Bostrom, J.; Hogner, A.; Schmitt, S. Do structurally similar ligands bind in a similar fashion. *J. Med. Chem.* **2006**, *49*, 6716–25.

CI800216F