Scaffold Hopping in Drug Discovery Using Inductive Logic Programming

Kazuhisa Tsunoyama,[†,‡] Ata Amini,[§,||] Michael J. E. Sternberg,[§] and Stephen H. Muggleton*[,†]

Computational Bioinformatics Laboratory, Department of Computing, Imperial College London, 180 Queen's Gate, London SW7 2AZ, United Kingdom, Molecular Medicine Research Laboratories, Drug Discovery Research, Astellas Pharma Inc., 21 Miyukigaoka, Tsukuba, Ibaraki 305-8585, Japan, and Structural Bioinformatics Group, Center for Bioinformatics, Division of Molecular Biosciences, Imperial College London, London SW7 2AZ, United Kingdom

Received November 16, 2007

In chemoinformatics, searching for compounds which are structurally diverse and share a biological activity is called scaffold hopping. Scaffold hopping is important since it can be used to obtain alternative structures when the compound under development has unexpected side-effects. Pharmaceutical companies use scaffold hopping when they wish to circumvent prior patents for targets of interest. We propose a new method for scaffold hopping using inductive logic programming (ILP). ILP uses the observed spatial relationships between pharmacophore types in pretested active and inactive compounds and learns human-readable rules describing the diverse structures of active compounds. The ILP-based scaffold hopping method is compared to two previous algorithms (chemically advanced template search, CATS, and CATS3D) on 10 data sets with diverse scaffolds. The comparison shows that the ILP-based method is significantly better than random selection while the other two algorithms are not. In addition, the ILP-based method retrieves new active scaffolds which were not found by CATS and CATS3D. The results show that the ILP-based method is at least as good as the other methods in this study. ILP produces human-readable rules, which makes it possible to identify the three-dimensional features that lead to scaffold hopping. A minor variant of a rule learnt by ILP for scaffold hopping was subsequently found to cover an inhibitor identified by an independent study. This provides a successful result in a blind trial of the effectiveness of ILP to generate rules for scaffold hopping. We conclude that ILP provides a valuable new approach for scaffold hopping.

Downloaded by IMPERIAL COLLEGE LONDON on September 4, 2015 | http://pubs.acs.org
Publication Date (Web): May 6, 2008 | doi: 10.1021/ci700418f

INTRODUCTION

Similarity search of compounds is often applied in drug discovery with the aim of finding molecules which have similar properties to an initial compound of interest. An important application of similarity search is scaffold hopping which aims to find compounds that are structurally diverse, while sharing a biological activity.[1] Several methods of scaffold hopping have been proposed so far.[2] Here, we propose a new scaffold hopping method based on inductive logic programming (ILP)[3] which learns from examples encoded as logical relationships and identifies human-readable rules describing structural features.

Pharmaceutical companies hold several million compounds in order to find leads for new drugs. They normally start their research with high-throughput screening (HTS) against a biological target which is considered to be associated with a certain disease. Within the academic community, there has been a recent increase of applications of HTS due to the emergence of the National Institutes of Health's Molecular Libraries Roadmap for identifying chemical probes to study the functions of genes, cells, and biochemical pathways.[4] In general, HTS includes several stages:[5,6] a large library of compounds is normally tested in the primary screening assay and successful "hit" compounds are evaluated in the secondary screenings for checking their sensitivity, specificity, pharmacology, and other properties. In drug discovery, the selected compound is called a "lead" for optimization by medicinal chemistry.

Although testing all available compounds with HTS tends to be thought of as a comprehensive search in chemical space, it has been pointed out that HTS has several limitations including the diversity of synthetically feasible compounds, quality, cost, time, and waste of valuable resources of materials.[7,8] In order to address these problems, virtual screening has been used to complement experimental screening.[9] In particular, integrating these two techniques in one workflow has been proposed as a smarter screening approach, which is often referred to as sequential screening.[8,10] In this approach, a subset of compounds is either randomly or rationally selected for the target and is tested initially. Any active compounds, together with the inactive compounds, provide an informative guide for selecting compounds in the next round of testing. Repeats of systematic selection of compounds are considered to be more efficient, especially in terms of cost. The selection method often involves techniques of similarity search.

In the search, the similarity of compounds is determined for finding neighbors of compounds of interest in chemical space.[11,12] If structurally diverse compounds are identified, this would help in finding new classes of compounds against

* Corresponding author. Telephone: +44 (0)20 7594 8307. Fax: +44 (0)20 7581 8024. E-mail: shm@doc.ic.ac.uk.
† Department of Computing, Imperial College London.
‡ Astellas Pharma Inc.
§ Division of Molecular Biosciences, Imperial College London.
|| Current address: Equinox Pharma Ltd., Bessemer Building, Prince Consort Road, London SW7 2BP, United Kingdom.

10.1021/ci700418f CCC: $40.75 © 2008 American Chemical Society
Published on Web 05/06/2008

Table 1. HTS Datasets Used in This Study

AID	assay name	source[a]	number of tested compounds	number of active compounds	threshold used to define active compounds
348	glucocerebrosidase-p2	NCGC (ncgca-glucocerebrosidase-p2)	4979	54	≤10 μM AC50
362	formylpeptide receptor−ligand binding assay	NMMLSC (UNM-FPR-01)	4282	61	≥65% inhibition[b]
373	SIP3 agonist primary HTS and Confirmation Assays	The Scripps Research Institute Molecular Screening Center (S1P3 AG BLA 1536%ACT)	59805	62	≥4% activation at 4.5 μM
408	voltage-dependent potassium channel beta subunit (KvBeta) inhibitor screen	Vanderbilt University Molecular Libraries Screening Center (VUMLSC) (VMLSCN00000003)	12369	112	slope values of the response curve[c]
409	voltage-dependent potassium channel beta subunit (KvBeta) substrate screen	Vanderbilt University Molecular Libraries Screening Center (VUMLSC) (VMLSCN00000002)	12369	20	slope values of the response curve[c]
412	discovery of novel allosteric agonists of the M4 muscarinic receptor	Vanderbilt University Molecular Libraries Screening Center (VUMLSC) (VMLSCN00000004)	12369	72	slope values of the response curve[c]
422	HTS for 14-3-3 protein interaction modulators	Emory University Molecular Libraries Screening Center (Emory 14-3-3 Screening May 19, 2006)	15157	24	≥30% inhibition at 40 μM
425	MKP-3 in vitro HTS assay	San Diego Center for Chemical Genomics (SDCCG-A002-MKP3)	64393	29	≥50% inhibition at 20 μM
428	measurement of GPCR-mediated thallium flux through GIRK channels	Vanderbilt University Molecular Libraries Screening Center (VUMLSC) (VMLSCN00000005)	8536	49	slope values of the response curve[c]
429	HTS for tumor Hsp90 inhibitors	Emory University Molecular Libraries Screening Center	15157	44	≥50% inhibition at 30 μM

[a] The information of assays and the tested compounds was downloaded on June 2006. [b] Compound concentration is not available. [c] The slope values of the response curve differ from the mean sample distribution at a 99.7% confidence level. Compounds were tested at 10 μM.

Downloaded by IMPERIAL COLLEGE LONDON on September 4, 2015 | http://pubs.acs.org
Publication Date (Web): May 6, 2008 | doi: 10.1021/ci700418f

the target protein. This approach is called "scaffold hopping,"[1] which has recently been described by Renner and Schneider as "finding isofunctional but structurally dissimilar molecular entities".[13] Scaffold hopping is useful for pharmaceutical companies where it can be used to obtain alternative compound structures when the initial compound under development has unexpected side-effects or when the initial compound is patented by competitors. Importantly, it has been reported that the sequential process which combines scaffold hopping methods with experimental screening is an efficient procedure for finding novel classes of compounds.[14,15]

Here we propose a new scaffold hopping method based on inductive logic programming (ILP).[3] ILP algorithms are designed specifically to learn human-readable rules from observed relationships in positive and negative examples (e.g., relationships between atoms in active and inactive compounds). ILP has been applied to various areas including bioinformatics and chemoinformatics.[16–21] The inter-relationships of the component atoms in the molecules are used by the learning algorithm as part of the background knowledge. ILP algorithms search for logically encoded hypotheses in the form of rules which cover as many of the positive examples as possible and the fewest negative examples. During learning, an ILP algorithm produces many hypothesized rules. The predictions of these rules can be represented as a binary string which encodes the classification of the set of covered compounds. Such binary strings can be used as inputs to other algorithms. One strategy, which has recently been developed, is to feed the binary strings generated by ILP into a support vector machine.[19] This is known as support vector inductive logic programming (SVILP) and has been demonstrated to produce improved predictive accuracy both in toxicology and screening.[20,21] In this paper, we follow a different, but related, approach in which we use the binary strings as inputs to a similarity search algorithm to perform scaffold hopping.

Table 2. Numbers of Active and Inactive Scaffolds Selected by Cluster Analysis[a]

AID	total number of selected compounds as scaffolds	number of active scaffolds	modified Tanimoto coefficients		number of inactive scaffolds	modified Tanimoto coefficients	
			average	SD[b]		average	SD[b]
348	1074	28	0.40	0.06	1046	0.40	0.05
362	449	25	0.43	0.06	424	0.42	0.05
373	7294	38	0.43	0.05	7256	0.41	0.04
408	3396	74	0.44	0.05	3322	0.42	0.05
409	3353	20	0.42	0.05	3333	0.42	0.05
412	3364	64	0.42	0.04	3300	0.42	0.05
422	3144	23	0.42	0.05	3121	0.42	0.05
425	7611	23	0.41	0.05	7588	0.41	0.04
428	2739	45	0.42	0.04	2694	0.42	0.05
429	3172	39	0.41	0.04	3133	0.42	0.05

[a] The average values and standard deviations of modified Tanimoto coefficients within active and inactive scaffolds each are also shown. These scaffolds were used for the benchmarking of scaffold hopping. [b] SD = standard deviation.

Downloaded by IMPERIAL COLLEGE LONDON on September 4, 2015 | http://pubs.acs.org
Publication Date (Web): May 6, 2008 | doi: 10.1021/ci700418f

MATERIALS AND METHODS

Data Preparation. Ten data sets of results from HTS and the associated structure-data files of tested compounds were taken from the PubChem database[22] (accessed June 2006; see Table 1). The target proteins for these HTS results include enzymes, channels and G-protein-coupled receptors, which are frequently subjects of HTS in pharmaceutical companies.

Next, we prepared sets of diverse active and inactive compounds from each of the HTS results. The structure-data files of tested compounds were converted to binary fingerprints by Open Babel software based on the occurrence of linear fragment up to seven atoms in length.[23] Active and inactive compounds in each HTS result were respectively clustered by their chemical similarities based on modified Tanimoto coefficient.[24,25] This coefficient measures the similarity between two compounds based on the presence of common molecular fragments. The modified Tanimoto coefficient ranges from 0 to 1, with identical molecules having a score of 1.0. The criterion for excluding similar molecules was that they had a modified Tanimoto coefficient of greater than 0.7. The compound closest to the centroid of each cluster was selected as the representative of the scaffold structure of the cluster. For simplicity and in keeping with the terminology of Renner and Schneider,[13] in the context of our study, these centroids will be referred to as scaffolds. We note, however, that the term "scaffold" can also be used to refer to the molecular backbone and this use is not identical to our use in this paper. These active and inactive scaffolds provided a chemically diverse data set and the average values of modified Tanimoto coefficients within active and inactive scaffolds each were about 0.4 as shown in Table 2. The aim of the study is to identify different scaffolds without removing similar compounds from the results. The evaluation and the comparative assessment of the results including similar compounds identified by different methods would be more complex.

For each HTS assay, 10% of active scaffolds were randomly set aside as part of the test data set. The remaining 90% of active scaffolds were used as positive examples in the training data set. These ratios were chosen in order to achieve a reliable estimate of performance in unseen data given the number of positive examples in the data. Concept learning algorithms, such as those used in ILP, gain from a process of selecting roughly equal numbers of positive and negative examples. This can be achieved in various ways. One obvious approach includes random selection of an equal number of positive and negative examples. A contrasting approach involves selection of "near misses", i.e. pairings of similar positive and negative examples. This focuses the learning on boundary discrimination. For each active scaffold, we construct a near miss by finding the closest inactive scaffold. The distance between scaffolds is measured according to the modified Tanimoto coefficient. All the remaining inactive scaffolds were incorporated into the previously selected 10% of active scaffolds and were used as a test data set. The numbers of scaffolds in the training data set and the test data set are summarized in Table 3.

ILP-Based Method. Our method is illustrated in Figure 1a. We used the ILP system CProgol[26,27] version 5.

CProgol requires examples and background knowledge for the learning process. An example might be as follows. In

Scaffold 1000 is active.

the above, "1000" is a scaffold name. Background knowledge described the chemistry of the scaffolds. We assigned pharmacophore types to each atom of every scaffold in the training data set to develop the background knowledge. Five pharmacophore types, defined in the literature,[14] were used: positively charged, negatively charged, hydrogen-bond donor, hydrogen-bond acceptor, and lipophilic. The coordinate information of atoms was included along with pharmacophore types. The coordinate information was prepared from a conformer which was selected as that having lowest energy according to the CONCORD[28] program. The definition for calculating spatial distances between two atoms, and a tolerance value of 0.5 Å, were also added to the background knowledge. A typical piece of background knowledge could be as follows: where "a10" is a unique label for an atom in scaffold 1000.

Atom a10 in scaffold 1000 is a lipophilic atom.

The x, y and z coordinate of atom a10 are respectively

5.40, 1.21, -1.82.

When learning, CProgol searches for rules describing combinations of spatial relationships between pharmacophore types. An example of such a rule might be as follows. The part of the rule which follows "if" is called the body of the rule. We limited the search space to rules which consist of

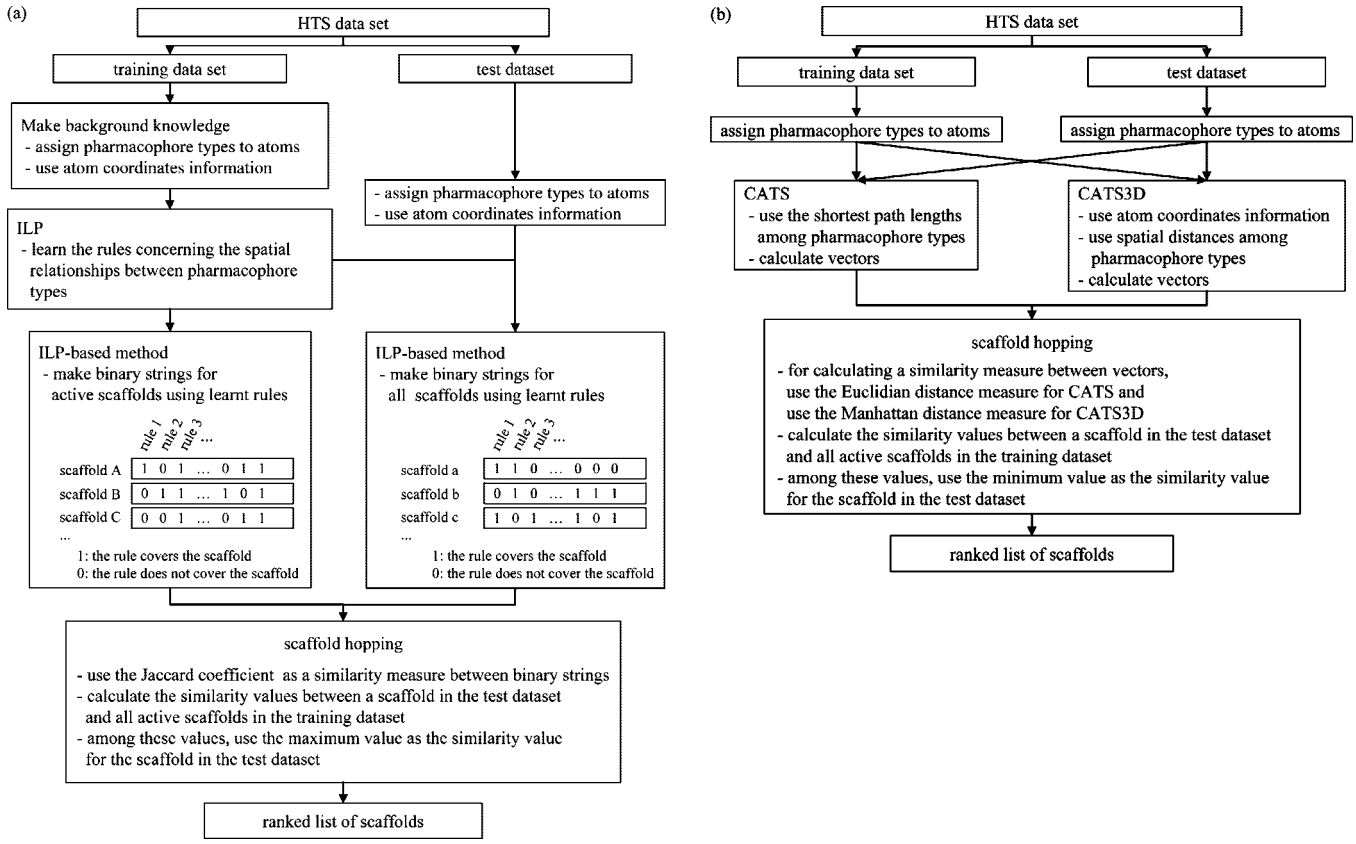

Figure 1. Process of scaffold hopping by three algorithms. (a) ILP-based method. (b) CATS and CATS3D.

Scaffold X is active if

X contains atoms A, B and C, and

the distance between atom A and B is 8.8±0.5Å, and

the distance between atom B and C is 4.8±0.5Å, and

the distance between atom A and C is 12.9±0.5Å, and

A and C are lipophilic atoms, and

B is an acceptor atom.

relationships between two atoms (say, atom A and atom B), concatenated pairs of relationships (atom A and B, then atom B and C), and triangle relationships (atom A and B, atom B and C, atom A and C). Owing to time considerations, we also excluded relationships among only lipophilic atoms from the calculations. Each rule is evaluated in CProgol with a measure known as compression (f):

$$f = P(p - (n + c))/P$$

Where, P is the total number of positive examples, p is the number of positive examples explained by the rule, n is the number of negative examples incorrectly explained by the rule, and c is the number of relationships in the body of the rule. Only rules with positive compression were extracted and used for in the next step.

The number of extracted rules differs depending on the HTS data set as shown in Table 3. On average, 1760 rules were obtained. These rules were used for producing the binary string for each scaffold (Figure 1a). Each bit in a binary string corresponds to an individual rule and has the value 1 when the rule covers the scaffold and has the value 0 when the rule does not. The active scaffolds in the training data set were converted to binary strings. The scaffolds in the test data set were also converted to binary strings based on the same rules. The Jaccard coefficient was used for calculating the similarity between binary strings of an active scaffold in the training data set and a scaffold in the test data set. The Jaccard coefficient is computed as the ratio of the number of rules covering both scaffolds to the number of rules covering only one scaffold.

Table 3. Numbers of Active and Inactive Scaffolds in the Training Data and the Test Dataset

AID	number of active scaffolds in training data set	number of scaffolds in test data set	number of active scaffolds in test data set	number of rules used in ILP-based method
348	25	1024	3	2592
362	22	405	3	2572
373	34	7226	4	1018
408	66	3264	8	823
409	18	3317	2	568
412	57	3250	7	3395
422	20	3104	3	119
425	20	7571	3	253
428	40	2659	5	2725
429	35	3102	4	3531

Comparison of Methods. We compared our ILP-based method with two algorithms: (a) CATS (chemically advanced template search),[1] the first published scaffold hopping method, and (b) CATS3D,[28–32] an extended method considering the three-dimensional shape of the molecules. These algorithms have the advantage of being able to perform

Downloaded by IMPERIAL COLLEGE LONDON on September 4, 2015 | http://pubs.acs.org
Publication Date (Web): May 6, 2008 | doi: 10.1021/ci700418f

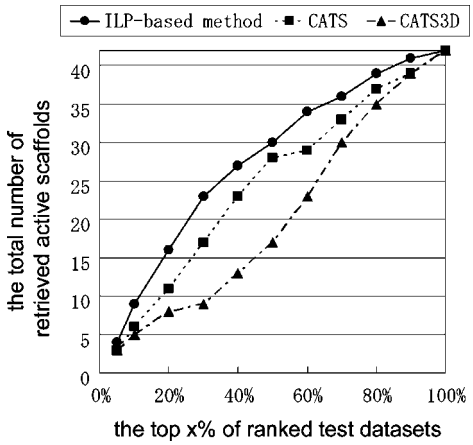

Figure 2. Graph of the total number of retrieved active scaffolds in the top *x*% of the ranked test data set.

Downloaded by IMPERIAL COLLEGE LONDON on September 4, 2015 | http://pubs.acs.org
Publication Date (Web): May 6, 2008 | doi: 10.1021/ci700418f

scaffold hopping from one positive example and do not require negative examples.

Figure 1b shows the implementation of CATS and CATS3D used in this study. CATS uses the two-dimensional topology of a molecule. CATS assigns five predefined pharmacophore types (hydrogen-bond donor, hydrogen-bond acceptor, positively charged, negatively charged, and lipophilic) to relevant atoms. The algorithm is based on calculating vectors that describe frequencies of shortest paths along bonds between pharmacophore types. The shortest paths from 1 to 10 bonds are considered. Each vector is scaled by the total number of non-hydrogen atoms in a molecule. Obtained vectors are compared using the Euclid distance. CATS3D uses spatial distances instead of paths along bonds to generate the vectors. The vectors are produced by using twenty bins (from 0 to 20 Å) of distances and are also scaled in the same manner as CATS. The Manhattan distance (a sum of absolute values of differences between values in bins) is used as the similarity metric between vectors. CATS3D adds a new feature, polar, to the definitions of pharmacophore types if the atom is a hydrogen-bond donor and also a hydrogen-bond acceptor atom.

The definitions of the pharmacophore types differ between CATS and CATS3D. CATS uses its own definitions of the pharmacophore types, but CATS3D uses the ph4_aType function of the commercially available software, MOE.[29,33] The choice of these definitions will influence the accuracies of methods and should be validated, but in this study, we used the same definitions of the pharmacophore types as in the ILP-based method. These definitions were originally used in the report of the analogous CATS methodology.[14] We consider that using the same definitions of the pharmacophore types provides an unbiased comparison between these algorithms.

Benchmark of Scaffold Hopping. In order to compare the results of the ILP-based method, CATS, and CATS3D, we conducted virtual screening and calculated their retrieval rates. Following the division into a training and test data set, the active scaffolds in the training data set were used to search active scaffolds in the test data set. In the ILP-based method, each active scaffold in the training data set and all scaffolds in the test data set were converted to binary strings using the ILP rules. Each scaffold in the test data set was then considered in turn. The Jaccard coefficients between it and all the active scaffolds in the training data set were evaluated and the maximum coefficient was used as the score for the scaffold in the test data set. The scaffolds in the test data set were ranked based on these scores. CATS and CATS3D vectors were also calculated. For ranking scaffolds in the test data set, an analogous procedure to the ILP-based method was used with the exception that the minimum Euclid or Manhattan distance was used.

For the ten HTS data sets, we compared the success rate of scaffold hopping between the three algorithms. We counted the number of active scaffolds which were included in the top *x*% of the ranked test data set and compared the numbers between the three algorithms.

RESULTS

Numbers of Retrieved Active Scaffolds. Figure 2 shows the total numbers of retrieved scaffolds for the ten HTS data sets. For all the values of *x* considered in the top *x*% of the ranked test data set, more active scaffolds are retrieved by the ILP-based method compared to CATS and CATS3D. Table 4 reports the numbers of retrieved scaffolds for the top 5, 10, 20, and 30% of the ranked test data set. In the top 5% of the ranked test data set, the ILP-based method retrieved four scaffolds while CATS and CATS3D each retrieved three (the difference is not statistically significant). However, comparisons in other cases of the top *x*% of ranked test data set showed larger differences in the performance. For example, in the top 10% of the ranked test data set, the ILP-based method retrieved nine active scaffolds, whereas CATS and CATS3D retrieved six and five active scaffolds, respectively. The retrieved scaffold structures are shown in Figure 3. In three HTS data sets, active scaffolds were retrieved only by the ILP-based method but not by CATS or CATS3D. In only one HTS data set (AID 422), both CATS and CATS3D found an active scaffold while the ILP-based method did not. Only in AID 428, CATS found more active scaffolds than the ILP-based method.

Comparison to Random Selection. To assess the significance of these results, we first examined whether each of the three methods performed better than chance selection. The ILP-based method found nine scaffolds which is significantly better than random selection (cumulative binomial probability $P = 0.0211$, where the probability of random success was 1/10). Neither CATS (which found six scaffolds) nor CATS3D (which found five) were significantly better than random selection ($P = 0.2396$ and 0.4121 for CATS and CATS3D, respectively). The ILP-based method also showed significantly better performance than random selection for the top 20 and 30% ($P < 0.01$). The other methods did not ($P > 0.096$).

Comparison of Three Methods. Next, we examined whether the ILP-based method can be shown to be statistically better than CATS and CATS3D. In the top 10% of the ranked test data set, we used the CATS result of finding 6 scaffolds out of 42 to estimate the probability of success and then evaluated the cumulative binomial probability of obtaining 9 scaffolds (i.e., the result from the ILP-based method) as 0.136. Similarly, the cumulative binomial probability of obtaining 9 scaffolds with the CATS3D result (5 hits) was 0.056. Thus the ILP-based method is not significantly better than these two approaches for the top 10%. In the case of the top 20%, the number of scaffolds obtained

Table 4. Numbers of Retrieved Active Scaffolds in the Top 5, 10, 20, and 30% of the Test Dataset Ranked by Three Algorithms

AID	number of scaffolds in test data set	number of active scaffolds in test data set	number of retrieved active scaffolds											
			ILP-based method				CATS				CATS3D			
			top 5%	10%	20%	30%	top 5%	10%	20%	30%	top 5%	10%	20%	30%
348	1024	3	0	1	1	1	0	0	0	1	1	1	1	1
362	405	3	0	0	2	2	0	0	1	1	0	0	0	0
373	7226	4	1	1	2	3	0	0	0	0	0	0	0	1
408	3264	8	0	1	2	2	0	1	1	2	0	0	0	0
409	3317	2	0	0	1	1	0	0	0	1	0	0	0	0
412	3250	7	1	1	3	3	1	1	2	3	0	0	0	0
422	3104	3	0	0	0	2	0	1	2	2	0	1	1	1
425	7571	3	1	2	2	3	0	0	2	2	1	1	1	1
428	2659	5	1	1	1	3	2	2	2	3	1	1	2	2
429	3102	4	0	2	2	3	0	1	1	2	0	1	3	3
total	*34922*	*42*	*4*	*9*	*16*	*23*	*3*	*6*	*11*	*17*	*3*	*5*	*8*	*9*

by the ILP-based method also was not significantly better compared with CATS ($P = 0.061$), but was significantly better than CATS3D ($P = 0.003$). For the top 30%, the ILP-based method showed significantly better performance than both methods ($P < 0.05$). We conclude that the ILP-based method is significantly better than random selection while

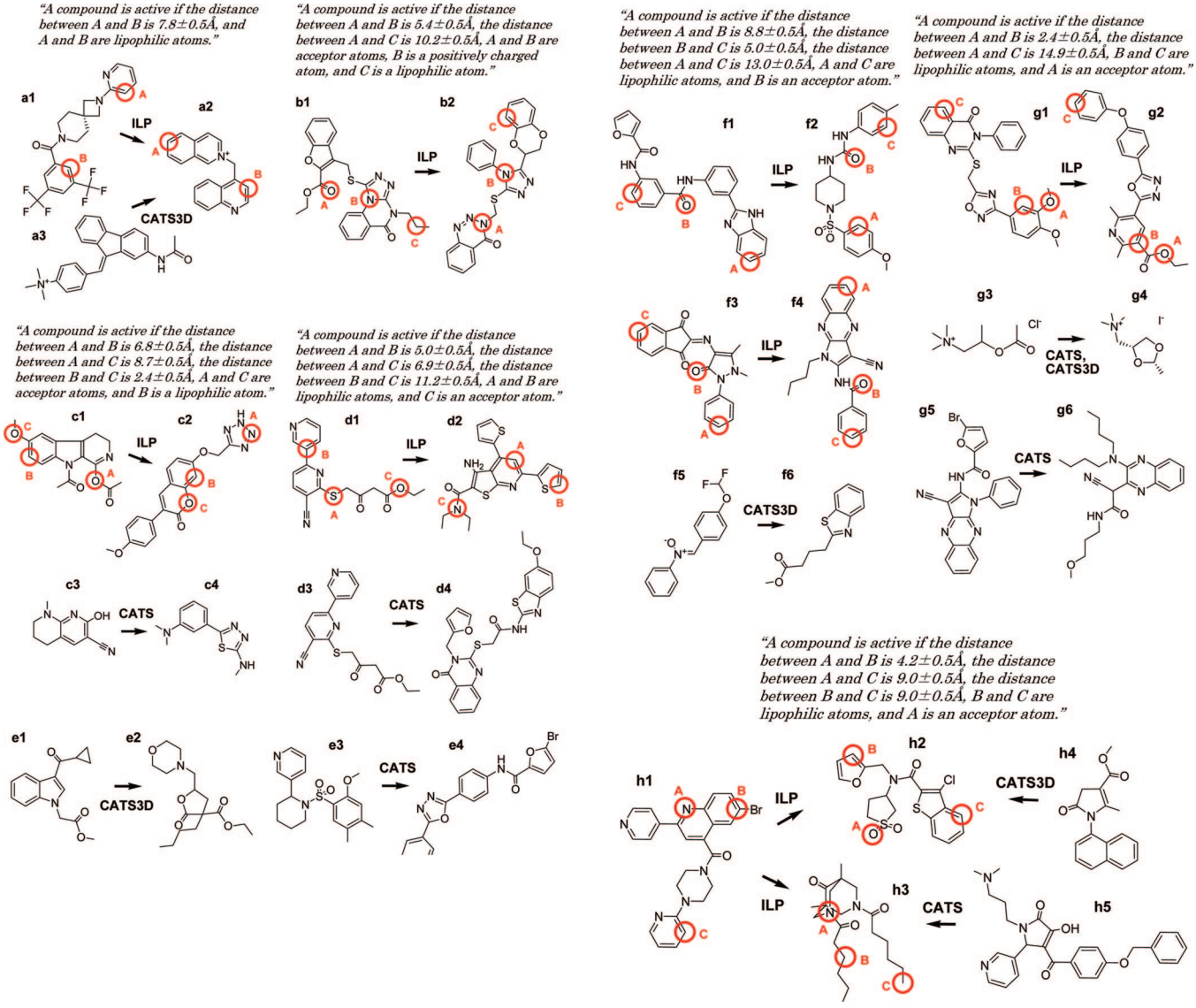

Figure 3. Active scaffolds retrieved by three algorithms in the top 10% of ranked test data set. The retrieved active scaffolds have an pointing to them arrow from the most similar active scaffolds in the training data set: a1−a3 in AID 348; b1 and b2 in AID 373; c1−c4 in AID 408; d1−d4 in AID 412; e1−e4 in AID 422; f1−f4 in AID 425; g1−g6 in AID 428; h1−h5 in AID429. The dominant rules are also shown. Corresponding atoms in the rule are marked with red circles. In ref 14, pharmacophore types as defined by the Sybyl line notation (SLN) language are the following: Acceptor N[not = NH]|O[not = OH], Donor Het[is = HetH], Negative O[is = O(H)Hev = Het], Positive N[(is = N(Any)(Any)Any and not = N-Hev = O)], and Lipophilic C[not = CN, CO, CS=O,CP=O]|S[not = SH, SO].

Downloaded by IMPERIAL COLLEGE LONDON on September 4, 2015 | http://pubs.acs.org
Publication Date (Web): May 6, 2008 | doi: 10.1021/ci700418f

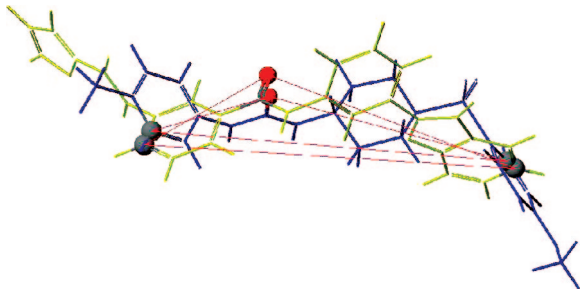

Figure 4. Three-dimensional depiction of a dominant rule for scaffold hopping. The alignment contains the retrieved active scaffold (blue) and the most similar compound (yellow) in the training data set (the corresponding 2D structures are f2 and f1 in Figure 3). The red lines represent the dominant rule. Gray and red balls indicate lipophilic and acceptor atoms, respectively.

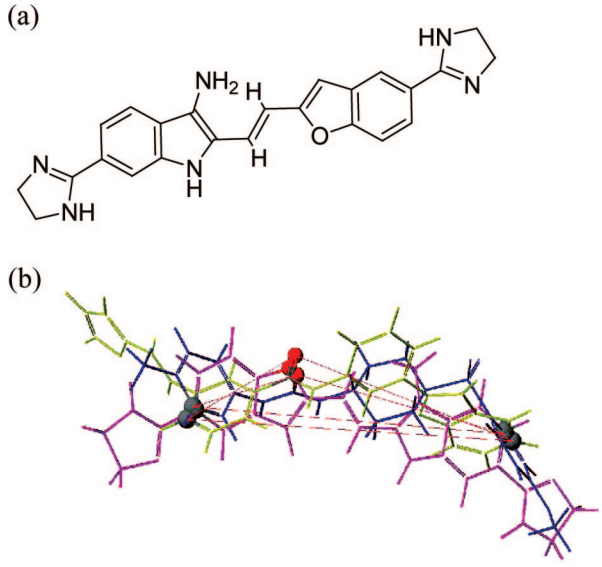

Figure 5. Alignments of NSC357756 with two molecules in Figure 4 based on the dominant rule. (a) Structure of NSC357756. (b) Alignments of the molecules. The figure depicts NSC357756 (violet), molecule f1 (yellow), and molecule f2 (blue). The red lines represent the dominant rule shown in Figure 4.

Downloaded by IMPERIAL COLLEGE LONDON on September 4, 2015 | http://pubs.acs.org
Publication Date (Web): May 6, 2008 | doi: 10.1021/ci700418f

the other two methods are not. In addition, the ILP method is at least as good as the other algorithms and therefore has a valuable role in computer-based methods for scaffold hopping.

Insights into the Reason for Scaffold Hopping. In addition to just finding active scaffolds, it is helpful to identify the three-dimensional features that lead to different scaffolds having the required biological activity. In general most machine learning algorithms, including support vector machine and artificial neural networks, are exceptionally limited in the insight they can provide. However, a general feature of ILP is its ability to generate human-comprehensible rules. In this application, the ILP-based method used bit strings and each bit corresponds to a human-readable rule. We are therefore able to examine the reason for scaffold hopping. We note that CATS and CATS3D can provide which relationship between pharmacophore types is efficient for obtained similarity score because the values in CATS and CATS3D vectors reflect the frequency of the relationship between pharmacophore types. The rules obtained by ILP, however, more directly link to the substructure of molecules.

Identification of Dominant Rules. For each active scaffold in the test data set, we identified the most similar scaffold in the training data set as determined by the ILP-based method. We then identified the rules which predict the pairing of the active scaffold in the training data set with the retrieved active scaffold in the test data set. From these rules, we identified the rule which covered the most active scaffolds but the fewest inactive scaffolds in the training data set. We refer to this as the dominant rule for the pair of scaffolds. Figure 4 shows an example of the dominant rule associated with an alignment for a pair of scaffolds (the corresponding 2D structures are f1 and f2 in Figure 3). These scaffolds were retrieved in the top 10% of the ranked test data set of AID 425. The dominant rule for scaffold hopping is: the distance between atom A and B is 8.8 ± 0.5 Å, the distance between atom B and C is 4.8 ± 0.5 Å, the distance between atom A and C is 12.9 ± 0.5 Å, A and C are lipophilic atoms, and B is an acceptor atom. This rule described the triangle relationships between a double-bonded oxygen atom in the keto group and two carbon atoms in the two ring systems. Figure 4 indicates which spatial relationships are common between the different scaffolds. We note that despite these spatial relationships, there remains the possibility that the two active scaffolds actually bind to different sites of the same protein.

After we completed this study, we identified another study which considered the target protein, mitogen-activated protein kinase phosphatase-3 (MKP-3) in AID 425.[34] They reported three inhibitors for MKP-3 in vitro. Of these, the most active compound, NSC 357756, had in vivo antitumor activity in mouse models. With an increase of the distance tolerance from 0.5 to 0.6 Å, the dominant rule shown in Figure 4 now covers NSC357756. Figure 5 shows the alignments of NSC357756 with the two compounds shown in Figure 4 together with the associated dominant rule. This independent confirmation suggests that the dominant rule has a general applicability for scaffold hopping.

DISCUSSION

Comparison of Predicted Scaffolds between Algorithms. To gain further insight into the differences and similarities between the three algorithms, we examined the overlapping sets of identified active scaffolds found by three approaches in the top 10, 20, and 30% of the ranked test database (Figure 6). It has been pointed out that CATS and CATS3D performed differently.[13,29] We also observed only a limited overlap between the active scaffolds retrieved by CATS and CATS3D. Interestingly, no active scaffold was found in common by the three algorithms in the top 10%. For the top 20 and 30%, one and four scaffolds were retrieved in common, respectively. Thus, the overlap still remains small. In general, there was little overlap of retrieved scaffolds between any pair of the three methods suggesting that each operates quite differently. What are the features in these algorithms that lead them to finding different scaffolds?

CATS uses topological information whereas CAT3D and ILP use the same three-dimensional information of compounds and this probably explains some of the differences in the results. However since CATS3D and ILP both employ three-dimensional information, we need to identify differences between these two approaches. First, the predicted three-dimensional structures of molecules generally contains errors in atom coordinates and the tolerance of such errors

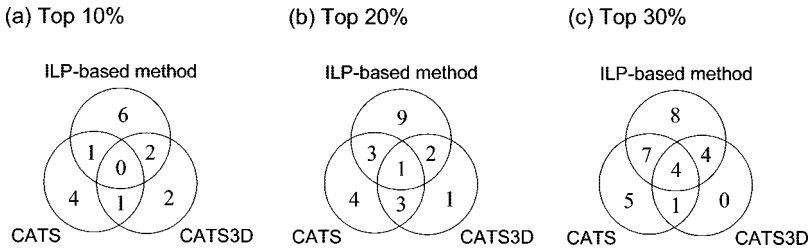

Figure 6. Venn diagrams representing retrieved active scaffolds by three algorithms in the top (a) 10%, (b) 20%, and (c) 30% of the ranked test data sets.

in these two algorithms might differ and thus explain the small number of scaffolds found in common. However, there are other major differences between the three algorithms. ILP finds the features that describe structures of active molecules and hence can discover new features of scaffold hopping which were not able to be found by the other two methods. In addition, ILP employs information about the inactive examples and this can help in the derivation of rules which would not be identified by CATS and CATS3D.

This study was specifically aimed at assessing scaffold hopping and so we needed to establish an appropriate test set that met our criterion of a new scaffold. To do this we employed the Open Babel fingerprints which are based on atom pair distances. This inevitably introduces a bias in the data set. CATS employs topological atom pairs to find scaffolds and this approach is related to the Open Babel fingerprints and thus CATS may therefore perform less well on our test data set than ILP. However CATS3D does not use topological atom pairs and so this bias unlikely to affect the CATS3D approach compared to ILP.

Quality of Data Sets. Our data sets contain the results for the primary and the confirmation HTS assays. Of the ten HTS data sets used, seven data sets come from a single experiment, two include AC50 values and one data set comes from a confirmed assay. The comments associated with these data sets note that these HTS results contain artifacts due to measurement errors. In general, HTS results are noisy due to many factors including degradation of compounds on screen plates, measurement errors, or using one concentration in a single assay.[8] It is envisaged that our approach often would be applied to more focused and higher quality data sets but these were not available to us. Accordingly we used these publicly available data sets to benchmark our method. Indeed, the fact that the ILP-based method performed significantly better than random selection suggests that the method has a sufficient degree of robustness to noise. In addition, we note that this study has considered HTS data. Further work could investigate the applicability of our approach to iterative virtual screening compared to a standard HTS approach. In such a study, one should consider the tradeoff of cost, time and yield in terms of hits.

The results of the evaluation in this study are affected by the data sets studied and different data sets might have yielded different results, particularly as the differences between the three methods involve small numbers. We expect that more data sets will be publicly available in future to assess our method with various target protein and assay types. We emphasis at this stage at the top 5% of the ranked data set (which is the useful level for real-world application), we can only conclude that ILP, CATS and CATS3D offer complementary approaches for scaffold hopping and cannot claim that any approach is superior.

Conformational Flexibility of Compounds. We used one conformer for a molecule in CATS3D and in the ILP-based method. A recent study with 150 crystal structures of protein−ligand complexes showed that over 60% of the ligands do not bind in a local minimum conformation.[35] Therefore, it is inappropriate to select one conformation for each molecule and several potential conformations should be used in the calculation. Interestingly, it was shown that the benefit of using multiple conformations is not as high as one might expect for CATS3D and the authors recommend the use of single conformation for large databases.[30] ILP can incorporate multiple conformations in the calculation, which produces rules describing structural features for each conformer. However, if ILP were to be used to learn rules from data containing large numbers of multiple conformations, calculation time and memory consumption would become a serious problem. Further work is required to extend our method efficiently to manage multiple conformations.

Further Developments in the ILP-Based Scaffold Hopping. Other improvements for our method would include using new atom types or different similarity measures, learning more diverse rules, and, combining learnt rules with other algorithms. The major advantage of ILP is its flexibility to describe objects (molecules) and to declare rules for learning. ILP can incorporate other atom types such as base or acid, fragments of molecules, and even values of LogP (octanol−water partition coefficient) or LUMO (lowest unoccupied molecular orbital) for molecules. It is also able to learn various rules involving four, or more, point pharmacophore types. The obtained rules produce the binary vector for a molecule as described and other algorithms can use in accordance with the purpose. SVILP is one of such efforts. These improvements should be involved in the future study.

CONCLUSIONS

We propose a new scaffold hopping method using ILP. ILP uses the observed spatial relationships between pharmacophore types in pretested active and inactive compounds and learns rules identifying the structures of active compounds. Our ILP-based method was compared to two previously published algorithms (CATS and CATS3D) with sets of diverse scaffolds in ten HTS data sets. The ILP-based method is significantly better than random selection while the other two methods are not. In addition, the ILP-based method found new active scaffolds which were not retrieved by the other methods. The results indicate that the ILP-based method performs at least as well as previous algorithms.

Downloaded by IMPERIAL COLLEGE LONDON on September 4, 2015 | http://pubs.acs.org
Publication Date (Web): May 6, 2008 | doi: 10.1021/ci700418f

Importantly, ILP produces human-readable rules which provide insight into scaffold hopping. We consider that elucidation of scaffold hopping knowledge by ILP will provide guidance to identify new active compounds in medicinal chemistry.

ACKNOWLEDGMENT

We thank Dr. Alireza Tamaddoni-Nezhad for helpful discussion and Dr. Michael P. H. Stumpf for statistical advice.

REFERENCES AND NOTES

(1) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: a Contribution to Virtual Screening. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 2894–2896.
(2) Brown, N.; Jacoby, E. On Scaffolds and Hopping in Medicinal Chemistry. *Mini. Rev. Med. Chem.* **2006**, *6*, 1217–1229.
(3) Muggleton, S. H. Inductive Logic Programming. *New Generat. Comput.* **1991**, *8*, 295–318.
(4) Austin, C. P.; Brady, L. S.; Insel, T. R.; Collins, F. S. NIH Molecular Libraries Initiative. *Science* **2004**, *306*, 1138–1139.
(5) Dove, A. Screening for Content - The Evolution of High Throughput. *Nat. Biotechnol.* **2003**, *21*, 859–864.
(6) Malo, N.; Hanley, J. A.; Cerquozzi, S.; Pelletier, J.; Nadon, R. Statistical Practice in High-Throughput Screening Data Analysis. *Nat. Biotechnol.* **2006**, *24*, 167–175.
(7) Lipinski, C.; Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **2004**, *432*, 855–861.
(8) Stahura, F. L.; Bajorath, J. Virtual screening methods that complement HTS. *Comb. Chem. High Throughput Screen.* **2004**, *7*, 259–269.
(9) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* **2002**, *1*, 882–894.
(10) Engels, M. F.; Venkatarangan, P. Smart screening: approaches to efficient HTS. *Curr. Opin. Drug Discov. Dev.* **2001**, *4*, 275–283.
(11) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
(12) Sheridan, R. P.; Kearsley, S. K. Why Do We Need So Many Chemical Similarity Search Methods. *Drug Discov. Today* **2002**, *7*, 903–911.
(13) Renner, S.; Schneider, G. Scaffold-Hopping Potential of Ligand-Based Similarity Concepts. *ChemMedChem* **2006**, *1*, 181–185.
(14) Naerum, L.; Norskov-Lauritsen, L.; Olesen, P. H. Scaffold Hopping and Optimization towards Libraries of Glycogen Synthase Kinase-3 Inhibitors. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1525–1528.
(15) Franke, L.; Schwarz, O.; Muller-Kuhrt, L.; Hoernig, C.; Fischer, L.; George, S.; Tanrikulu, Y.; Schneider, P.; Werz, O.; Steinhilber, D.; Schneider, G. Identification of Natural-Product-Derived Inhibitors of 5-Lipoxygenase Activity by Ligand-Based Virtual Screening. *J. Med. Chem.* **2007**, *50*, 2640–2646.
(16) King, R. D.; Whelan, K. E.; Jones, F. M.; Reiser, P. G.; Bryant, C. H.; Muggleton, S. H.; Kell, D. B.; Oliver, S. G. Functional Genomic Hypothesis Generation and Experimentation by a Robot Scientist. *Nature* **2004**, *427*, 247–252.
(17) King, R. D.; Muggleton, S. H.; Srinivasan, A.; Sternberg, M. J. Structure-Activity Relationships Derived by Machine Learning: the Use of Atoms and Their Bond Connectivities to Predict Mutagenicity by Inductive Logic Programming. *Proc. Natl. Acad. Sci. U S A* **1996**, *93*, 438–442.
(18) Sternberg, M. J.; Muggleton, S. H. Structure Activity Relationships (SAR) and Pharmacophore Discovery Using Inductive Logic Programming (ilp). *QSAR Comb. Sci.* **2003**, *22*, 527–532.
(19) Vapnik, V. Methods of Pattern Recognition. In *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, 1995; pp 123−169.
(20) Amini, A.; Muggleton, S. H.; Lodhi, H.; Sternberg, M. J. A Novel Logic-Based Approach for Quantitative Toxicology Prediction. *J. Chem. Inf. Model.* **2007**, *47*, 998–1006.
(21) Cannon, E. D.; Amini, A.; Bender, A; Sternberg, M. J.; Muggleton, S. H.; Glen, R.; Mitchell, J. Support Vector Inductive Logic Programming Outperforms the Naive Bayes Classifier and Inductive Logic Programming for the Classification of Bioactive Chemical Compounds. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 269–280.
(22) PubChem. http://pubchem.ncbi.nlm.nih.gov (accessed June 2006).
(23) The Open Babel Package, version 2.0.2. http://openbabel.sourceforge.net/ (accessed Jul 3, 2006).
(24) Flinger, M. A.; Verducci, J. S.; Blower, P. E. A Modification of the Jaccard-Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings. *Technometrics* **2002**, *44*, 110–119.
(25) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747–750.
(26) Muggleton, S. H. Inverse Entailment and Progol. *New Generat. Comput.* **1995**, *13*, 245–286.
(27) Muggleton, S. H.; Bryant, C. H. Theory Completion Using Inverse Entailment. In *Proceedings of the 10th International Conference on Inductive Logic Programming (ILP-00)*; Cussens, J., Frisch, A. M., Eds.; Springer-Verlag: Berlin, 2000; pp 130−146.
(28) Pearlman, R. S. *Concord User's Manual*; Tripos, Inc.: St Louis, MO, 2000.
(29) Fechner, U.; Franke, L.; Renner, S.; Schneider, P.; Schneider, G. Comparison of Correlation Vector Methods for Ligand-Based Similarity Searching. *J. Comput.-Aided. Mol. Des.* **2003**, *17*, 687–698.
(30) Renner, S.; Schwab, C. H.; Gasteiger, J.; Schneider, G. Impact of Conformational Flexibility on Three-Dimensional Similarity Searching Using Correlation Vectors. *J. Chem. Inf. Model.* **2006**, *46*, 2324–2332.
(31) Renner, S.; Ludwig, V.; Boden, O.; Scheffer, U.; Gobel, M.; Schneider, G. New Inhibitors of the Tat-TAR RNA Interaction Found with a "Fuzzy" Pharmacophore Model. *ChemBioChem* **2005**, *6*, 1119–1125.
(32) Renner, S.; Noeske, T.; Parsons, C. G.; Schneider, P.; Weil, T.; Schneider, G. New Allosteric Modulators of Metabotropic Glutamate Receptor 5 (mGluR5) Found by Ligand-Based Virtual Screening. *ChemBioChem* **2005**, *6*, 620–625.
(33) *MOE (Molecular Operating Environment)*; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2006.
(34) Vogt, A.; Cooley, K. A.; Brisson, M.; Tarpley, M. G.; Wipf, P.; Lazo, J. S. Cell-active dual specificity phosphatase inhibitors identified by high-content screening. *Chem. Biol.* **2003**, *10*, 733–742.
(35) 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*, 2499–2510.

CI700418F

Downloaded by IMPERIAL COLLEGE LONDON on September 4, 2015 | http://pubs.acs.org
Publication Date (Web): May 6, 2008 | doi: 10.1021/ci700418f