

Finding Discriminating Structural Features by Reassembling Common Building Blocks

Kevin P. Cross,[†] Glenn Myatt,[†] Chihae Yang,[†] Michael A. Fligner,[‡] Joseph S. Verducci,[‡] and Paul E. Blower, Jr.*[†]

Leadscope, Inc., 1245 Kinnear Road, Columbus, Ohio 43212 and Department of Statistics, The Ohio State University, Columbus, Ohio 43210

Received June 3, 2003

We present a new method for constructing discriminating substructures by reassembling common medicinal chemistry building blocks. The algorithm can be parametrized to meet differing objectives: (1) to build features that discriminate for biological activity in a local structural neighborhood, (2) to build scaffolds for R-group analysis, (3) to construct cluster *signatures* that discriminate for membership in the cluster and provide a graphical representation for its members, and (4) to identify substructures that characterize major classes in a heterogeneous compound set. We illustrated the results of the algorithm on a literature dataset is of 118 compounds with in vitro inhibition data against recombinant human protein tyrosine phosphatase 1B (PTP-1B).

Introduction

Identifying the structural feature responsible for the pharmacological activity of a compound and finding classes of compounds with enhanced activity are key objectives of drug discovery research. A number of two-dimensional (2D) and three-dimensional (3D) computational techniques have been developed to solve this problem. In the 3D arena, these include pharmacophore mapping,¹ 3- and 4-point pharmacophore-based similarity searching² and Comparative Molecular Field Analysis³ (CoMFA). In the 2D arena, the oldest and most widely used technique to identify chemical classes is structure-based clustering.^{4,5} Compounds are typically represented as binary fingerprints, and the Tanimoto coefficient is used to measure the pairwise distance between compounds.

Recursive partitioning⁶ (RP) has proved to be another valuable tool for identifying structurally homogeneous classes with high mean values for a biological activity. It begins with a potentially large heterogeneous set of compounds and attempts to produce smaller and more homogeneous sets by finding features that successively split a set of compounds into two subsets that are more similar to each other in terms of their biological activity than the original set. In the standard application of RP, each feature can be used to divide the set of compounds into two groups, those with the feature and those without. Then the procedure is reapplied to each of the daughter nodes. An extension to the standard RP algorithm, called RP/SA,⁷ uses simulated annealing (SA) to determine which molecular descriptors best partition each node of the tree. In particular, the SA algorithm searches for a combination of k features to maximize a particular splitting criterion such as a t -statistic comparing the difference in the means of the two groups.

Although these tree classification algorithms can quickly identify structurally homogeneous classes with high mean activity, there is a disadvantage that they require a numerical property for the splitting criterion.

Another problem with both clustering and tree classification algorithms is that they group compounds based only on presence of common descriptors and, in the case of tree algorithms, the common absence of descriptors. But there is no indication that the compounds in a group share a common substructure larger than any structural features used by the classification algorithm. Maximal common substructure (MCS) algorithms^{8,9} can be used to build larger substructures. However, this procedure is very sensitive to outliers, which can reduce the MCS to a much smaller common substructure, and it only produces one substructure.

Two recent papers address the problem of constructing characterizing substructures for structural classes. Nicolaou, et al.¹⁰ describe a *phylogenetic-like tree* (PGLT) algorithm that is a method for analyzing large heterogeneous datasets to identify chemical classes that correlate with increased biological activity. For each root node comprising a set of compounds, the PGLT algorithm first clusters compounds into structurally homogeneous groups, then for each group, derives a chemical fragment common to all molecules. The common fragment is either the MCS of the compounds in the cluster or the *significant common substructure*.¹⁰

Miller¹¹ describes a chemical class generation procedure implemented in the ChemTK software package. Classes are generated on the basis of common scaffold. Each scaffold is either a single ring system or several ring systems each connected by an unbranched chain of atoms. Thus, all atoms in a scaffold are connected to at least two other atoms. Scaffolds are limited to those comprising 2–4 rings, and scaffolds contained in fewer than a threshold number of compounds in the target set are eliminated.

We present a new method for constructing characterizing features for structural classes. Leadscope auto-

* To whom correspondence should be addressed. Phone: 614-675-3766, fax: 614-675-3732, pblower@leadscope.com.

[†] Leadscope, Inc.

[‡] The Ohio State University.

matically breaks down molecules into common structural fragments such as functional groups and heterocycles when compounds are loaded into the system. The new process we describe, Macrostructure Assembly (MSA), reassembles structural fragments in a directed way to produce larger substructures that are commonly occurring within a group of compounds or that discriminate for a biological response within the group. The structural fragments generated by the reassembly process are guided by one of several, potentially conflicting, objectives:

- Create substructures that discriminate for a biological response within a designated set of compounds. With this objective, the algorithm uses the response to optimize the MSAs generated.
- Create substructures that discriminate for membership in a set of structurally homogeneous compounds. These MSAs have application to structure-based clustering and provide *signatures* that summarize of the structural contents of the clusters.
- Create large, commonly occurring substructures. MSAs generated from this objective provide scaffolds for R-group analysis within a congeneric series.

Methods

Molecular Building Blocks. The initial set of molecular building blocks are those defined in the Leadscape Structural Feature Hierarchy.¹² This is based on structural features and combinations of features commonly used for experimental design in drug discovery programs. When Leadscape loads a set of compounds to create a project, the software performs a systematic substructure analysis using predefined structural features stored in a feature library. The structural features chosen for analysis are motivated by those typically found in small molecule drug candidates. At the present time, the feature library contains over 27 000 structural features. The major structural classes include: *amino acids; bases, nucleosides; benzenes; naphthalenes; carbocycles; carbohydrates; elements; functional groups; heterocycles; natural products; peptidomimetics; pharmacophores; protective groups and spacer groups.* The features represent a wide range of structural specificity from very specific substructures such as *benzene, 1-hydroxymethyl, 3-methoxy* to generic features such as the pharmacophores which define pairs of generalized physicochemical atom types joined by a path of atoms/bonds of indeterminate type.

The Algorithm. The algorithm is guided by one of the three objectives described above: it searches for (1) large, commonly occurring substructures; (2) substructures that discriminate for a biological response; or (3) substructures that discriminate for membership in a set of compounds. It is further controlled by a set of adjustable parameters:

Minimum Compounds per MSA. This parameter eliminates MSAs that are not contained in a minimum number of compounds and restricts MSAs to those that occur most commonly within the target set.

Minimum Number of Atoms per MSA. This parameter controls the size of MSAs that are generated.

Maximum Number of Rotatable Bonds per MSA. This parameter can be used to control the flexibility of the MSAs that are generated.

Minimum Absolute z-Score per MSA. This parameter can be used to filter the MSAs that discriminate for a biological response. Higher values restrict MSAs to those that occur in compounds with unusually high or low mean values of the response. The z-score compares the mean activity of a subset to the expected value according to eq 1:

$$z = (\bar{x}_1 - \bar{x}_0) \sqrt{\frac{n_1 n_0}{s_0^2 (n_0 - n_1)}} \quad (1)$$

where \bar{x}_1 , \bar{x}_0 are the mean activities of the subset and full set, respectively, n_1 and n_0 are the set sizes and s_0^2 is the sample variance of the full set.

The algorithm proceeds in three phases: preprocessing, the iterative feature reassembly phase, and a postprocessing phase.

Preprocessing. 1. Filter/process structural features. This step prepares the basic structural building blocks used for MSA generation. First, generic features such as pharmacophore binding pairs are eliminated as building blocks. Second, most pattern modifiers are removed from structural feature templates. These pattern modifiers are atom/bond restrictions that control the external environment of a template match. For example, an atom modifier may require that the matching atom be *closed*; that is, the atom matching the template atom may have no neighbors except those matching neighbors of the template atom.

2. Generate ring systems. This step generates additional structural building blocks by identifying all unique ring systems within the target compound set. If a ring has an exocyclic tautomeric bond to the ring, the exocyclic, tautomeric atoms are included in the ring system. In such cases, the ring both with and without exocyclic tautomers is also kept.

3. Prune building block list. This step removes duplicate and infrequent templates that may be generated in the previous steps. The Leadscape Feature Hierarchy contains numerous rings so many ring systems identified in step 2 will be duplicates. As another example, the building blocks derived from *pyridine, 4-alkyl*, and *pyridine, 4-methyl* will be identical because the atom restriction that differentiated them are removed in step 1. Those templates that occur less frequently than the specified parameter for minimum compounds per MSA are also removed from consideration, since any MSA built from them, by definition, would be too infrequent to keep.

Reassembly Phase. Repeat for user-specified number of cycles:

1. Prioritize templates. The list of building block templates is prioritized using a scoring function applied to each template T_i based on the general formula $w_1 F_i + w_2 S_i + w_3 Z_i$. F_i is the frequency of template T_i in the target set, S_i is the size of template T_i (number of atoms), and Z_i is the absolute value of the z-score for the subset of compounds containing T_i compared to the target set, according to eq 1. All values of F_i , S_i , Z_i are standardized to the range [0.0–1.0], and in the case where no response data is available for the compounds, we set $Z_i = 0$. The weighting parameters are adjustable and depend on the objective. For example, if the objective is to generate structural feature that discriminate

for high or low values of a biological response, the Z_i term is most considered most important and given a higher weight.

2. Prune list of template pairs. Each pair of building block templates (where at least one member of the pair was created in the previous iteration) on the prioritized list is examined for possible elimination. The size of the candidate template combination (that would be formed by merging the building block templates) is estimated assuming a one-atom overlap of the two building blocks in the target compounds. If the two building blocks share no common atom in the target compounds, the pair is discarded. The frequency and z -score of the candidate template combination is estimated from the logical AND of the pair of structure bitsets. If the minimum frequency criteria (a user-specified parameter) is not satisfied, the pair is discarded.

3. Repeat for each building block pair:

a. For each target compound containing both building blocks, generate one or more merged templates by mapping the building blocks onto the target structure. For each mapping of the building block pair, all atoms and bonds of the target not covered by the map are eliminated. In this way, the target structures guide the reassembly of building blocks to form merged templates.

b. Prune list of newly generated templates to eliminate duplicates and templates not satisfying requirements for minimum frequency or maximum number of rotatable bonds.

4. Remove duplicate and contained templates. Processing the list of building block pairs in step 3 generates many duplicates. In addition if one template is a substructure of another and they both match the same set of compounds, the smaller template is discarded.

5. Add the list of newly generated templates to the master list.

Postprocessing. 1. Prune final list of templates to eliminate the initial building blocks from the base set and templates not satisfying requirements for minimum atom count, or minimum z -score.

2. Remove redundant templates. Duplicate templates are identified and eliminated at several points in the reassembly process. However, the list of templates from the previous step typically contains many nonduplicate, but highly redundant, clusters of similar templates. The final step is to select a representative from each template cluster, where members of a cluster are structurally similar and correspond to nearly the same compound subset.

a. Any template matching every compound in the target set is retained.

b. For each pair of templates (T_i , T_j), calculate the Pearson correlation coefficient P between the structure bitsets. If $P < 0.85$, retain both templates.

c. Let Z_i (Z_j) be the z -score for the subset of compounds containing T_i (respectively, T_j) compared to the target set. If $|Z_i - Z_j| > 0.3$, retain both templates.

d. Let Z_{OR} be the z -score of compound set containing either T_i or T_j ; let Z_{AND} be the z -score of compound set containing both T_i and T_j . Compare the absolute values of the four z -scores Z_i , Z_j , Z_{OR} , Z_{AND} . If Z_i (respectively, Z_j) has the largest absolute value of all four scores, discard T_j (respectively, T_i). Otherwise, retain both T_i and T_j .

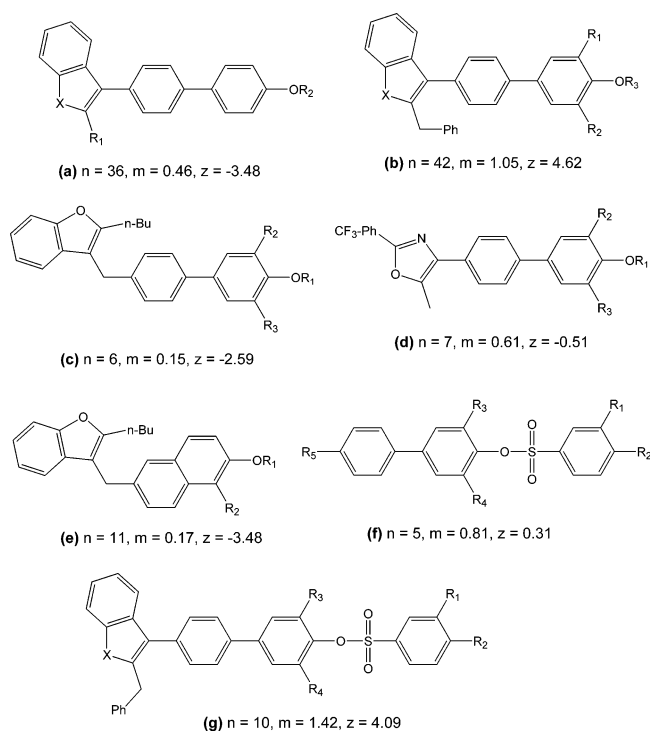


Figure 1. Summary of structure–activity relationships of PTP-1B inhibitors following the original authors' presentation.¹³ Classes are labeled **a–g**, and for each class we give the frequency (n), mean (m) pIC_{50} value, and z -score (z) calculated according to eq 1. The overall mean pIC_{50} and standard deviation for the 118 compounds is $\bar{x}_0 = 0.73$, and $s_0 = 0.56$.

This final pruning step typically eliminates 70–80% of the candidate MSAs.

Results and Discussion

Protein tyrosine phosphatase 1B (PTP-1B) has been proposed as an important negative regulator of the insulin signaling pathway, thus PTP-1B inhibitors are potential therapeutic agents for treatment of Type 2 diabetes and obesity.¹³ To illustrate the reassembly algorithm described above, we selected structure–activity data from a literature study of a series of benzofuran/benzothiophene biphenyls as PTP-1B inhibitors with anti-hyperglycemic activity. The dataset is composed of 118 compounds with in vitro inhibition data against recombinant human PTP-1B taken from Malamas et al.¹³ For another 19 compounds, the in vitro inhibition data was reported as percent inhibition at 2.5 μ M or 1 μ M. The latter compounds were excluded from our analysis.

Structure–activity data for the PTP-1B inhibitors, following the original author's presentation,¹³ are summarized in Figure 1. In this diagram, classes are labeled **a–g**, and for each class we give the frequency (n), mean (m) pIC_{50} value, and z -score (z) calculated according to eq 1. The overall mean pIC_{50} and standard deviation for the 118 compounds is $\bar{x}_0 = 0.73$, and $s_0 = 0.56$. The initial series is shown in Figure 1 (a), where the original lead compound was $X = O$, $R_1 = n$ -Bu, and $R_2 = H$. The most potent members of this series were substituted acetic acids ($R_2 = CH(CH_2Ph)CO_2H$) with benzothiophene substituents ($X = S$) and $R_1 = CH_2$ -aryl. Stereoisomers at the α -carbon of R_2 were approximately equipotent. Ten of the 19 compounds excluded from our study were

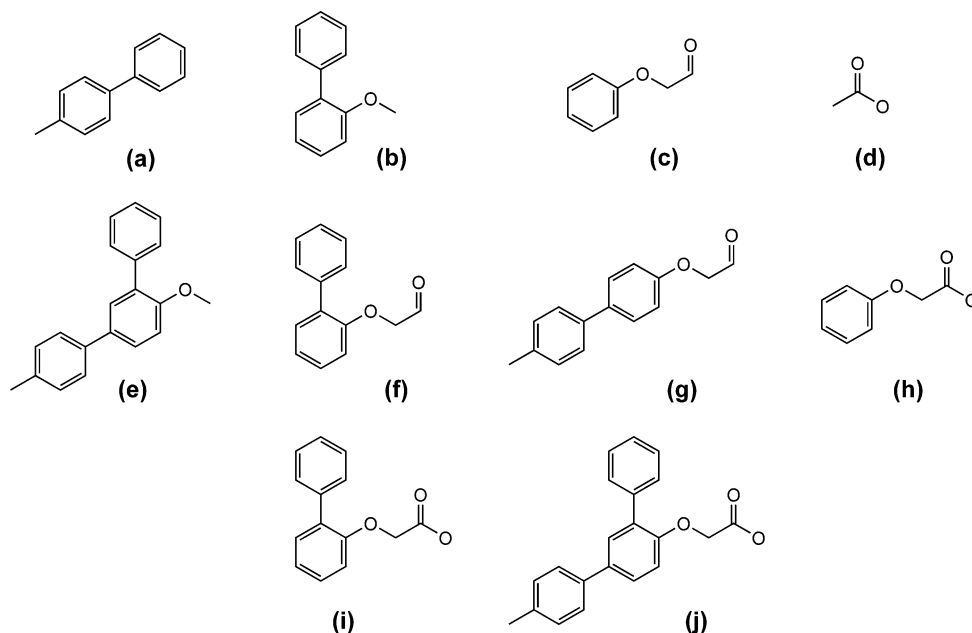


Figure 2. The macrostructure assembly process. Substructures **a–d** are several features from the Leadscape feature hierarchy that occur in many compounds in the set of 118 PTP-1B Inhibitors. MSA merges pairs of overlapping features to form larger features; for example, **a + b = e**, **b + c = f**, **a + c = g**, and **c + d = h**. These features are added to the building blocks, and the process is repeated producing the substructures **i** and **j** among others.

members of this series because the activity data was reported as percent inhibition.

The second series (Figure 1, **b**) investigated hydrophobic substituents at the ortho positions of the phenolic ring. Typical substituents at R_1 and R_2 are Br and 4-MeO-Ph. Overall, this series showed a 4-fold increase in potency over the initial series (Figure 1, **a**) with $n = 42$, $m = 1.05$, and $z = 4.62$. In this series, unsubstituted acetic acid substituents ($R_3 = \text{CH}_2\text{CO}_2\text{H}$) had higher potency than series 1 with IC_{50} values in the range 0.025–0.1 μM .

Other aryl groups in place of benzothiophene or benzofuran such as oxazole (Figure 1 (**d**), $n = 7$, $m = 0.61$, and $z = -0.51$) or adding a one-carbon spacer (Figure 1 (**c**), $n = 6$, $m = 0.15$, and $z = -2.59$) decreased potency. Replacing the 4,4'-biphenyl scaffold with a 2,6-naphthalenyl scaffold also resulted in a loss of potency (Figure 1 (**e**), $n = 11$, $m = 0.17$, and $z = -3.48$). However, replacing the acetic acid substituent by a sulfonylbenzoic acid resulted in a nearly 10-fold increase in potency over the initial series (Figure 1 (**g**), $n = 10$, $m = 1.42$, and $z = 4.09$). Within this series, unlike the acetic acid series, hydrophobic substituents at the ortho positions of the phenolic ring had little affect on potency.

Starting with this 118 compound subset, macrostructure assembly constructs larger substructures by reassembling the commonly occurring building blocks as illustrated in Figure 2. Substructures **a–d** in Figure 2 are several features from the Leadscape feature hierarchy that occur in many compounds in the target set. MSA merges pairs of overlapping features (i.e., two features that share a common atom in the target set) to form larger features. Referring to Figure 2, **a + b = e**, **b + c = f**, **a + c = g**, and **c + d = h**. The newly constructed features are added to the list of building blocks, and the process is repeated producing the substructures **i** and **j** in Figure 2. Of the substructures

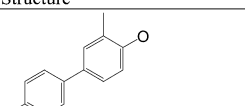
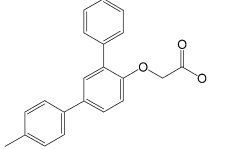
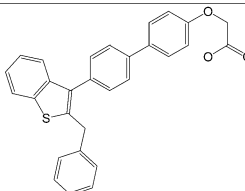

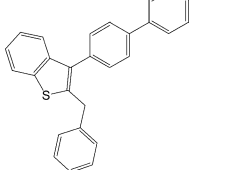
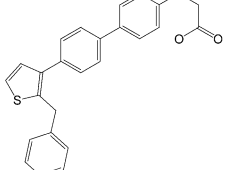
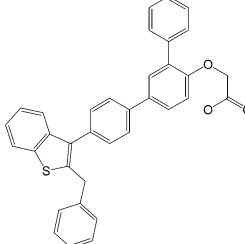
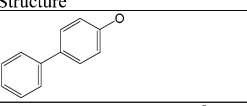
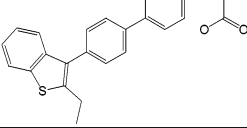
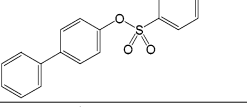
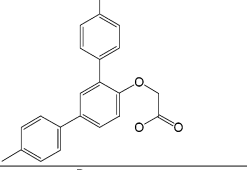
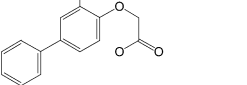
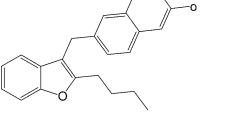
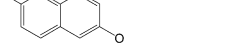
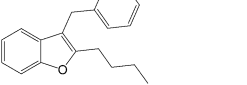

shown in Figure 2, only **j** was selected for inclusion in the final list (MSA **2** in Table 1). This macrostructure occurs in a 15-compound subset with mean PTP-1B activity (pIC_{50}) = 1.33, z -score = 4.4.

Results of MSA generation for the set of 118 PTP-1B inhibitors are shown in Table 1; the algorithm objective was to construct structural features that discriminate for activity. Each entry gives the ID, structure diagram, the number of compounds matching the MSA, the percent of the full dataset (118 compounds), and the mean pIC_{50} value and z -score for the matching subset. The table is sorted in decreasing order by z -score.

MSAs **8** and **14** are complementary and cover the full set. Thus, the subset of 106 compounds containing the 4-phenylphenoxy substructure has a mean pIC_{50} value of 0.79 compared to 0.73 for the full set, giving a z -score of 3.8 for this subset. In contrast, the subset of 12 compounds containing the 6-carbo-naphthalen-2-oxy substructure has a mean pIC_{50} value of 0.15 and a z -score of -3.8 .

The majority of MSAs in Table 1 are specializations of MSA **8**. Of the 106 compounds in this series, 20 compounds are phenols, 69 compound have an α -acetic acid substituent on the oxygen (class: 4'-phenylphenoxyacetic acid), and 15 compounds contain the phenylsulfonyl substituent (MSA **10**). Of the series of 69 compounds containing the 4'-phenylphenoxyacetic acid substructure, 31 have a 4''-benzothiophenyl substituent with mean $\text{pIC}_{50} = 1.05$, 30 have a 4''-benzofuranyl with mean $\text{pIC}_{50} = 0.74$, and the remaining eight compounds have other aryl groups. Thus, the 4''-benzothiophenyl substituent is an important discriminant for enhanced activity while the 4''-benzofuranyl analogues are neutral. This is reflected in MSAs **3**, **6**, **7**, and **9**. The benzofuranyl group appearing as a negative discrimi-

Table 1. MSAs Generated from PTP-1B Inhibitors^a

ID	Structure	Mean	Z-Score	Count	%
1		1.31	5.2	21	17.8
2		1.33	4.4	15	12.7
3		1.15	4.4	27	22.9
4		1.21	4.1	19	16.1
5		1.03	3.9	36	30.5
6		1.08	3.9	29	24.6
7		1.31	3.8	12	10.2
8		0.79	3.8	106	89.8
9		1.05	3.7	31	26.3
10		1.22	3.6	15	12.7
11		1.33	3.6	10	8.5
12		1.14	3.5	19	16.1
13		0.17	-3.5	11	9.4
14		0.15	-3.8	12	10.2
15		0.16	-4.5	17	14.4
16		0.15	-5.1	20	16.9

^a Each entry gives the I.D., structure diagram, the number of compounds matching the MSA, the percent of the full dataset (118 compounds), and the mean pIC₅₀ value and z-score for the matching subset. The table is sorted in decreasing order by z-score.

nant in MSAs **13**, **15**, and **16** is primarily due to its association with the naphthalen-2-oxo group as seen in MSA **13**.

Murthy and Kulkarni¹⁴ performed a comparative molecular field analysis (CoMFA) with the same set of PTP-1B inhibitors, using a training set of 92 of the compounds. This provided electrostatic and steric 3D contour maps, which are valuable aids in understanding the quantitative structure–activity relationships and designing new compounds for testing. However, the CoMFA analysis is quite complex. It requires selection of a template structure for molecular alignment, conformational analysis of the data set, molecular alignment with the template structure, determination of the number of factors used in the statistical model building, and statistical validation of the predictivity of the model. In this study the authors also evaluated multiple alignment procedures and inclusion of additional elec-

tronic, spatial, and thermodynamic descriptors. Many of these steps require expert assistance to obtain reliable results.

Table 1 illustrates results of the algorithm where the objective is to build features that discriminate for activity in a local neighborhood. These MSAs are useful descriptors for building prediction models based on statistical techniques such as multiple regression, *k*-nearest neighbors, or recursive partitioning. Because MSAs are assemblies of familiar medicinal chemistry building blocks, the models are easy to interpret and present 2D structure–activity relationships that can be used to design follow-up experiments. MSAs also provide the basis for R-group analysis to further refine the SAR within the corresponding compound set. For example, MSA **3** from Table 1 can be used as an R-group scaffold to study the effects of positional and substituent

variations on potency of PTP-1B inhibition within the set of 27 compounds.

Another complementary objective for building MSAs is to provide a structural or, for visualization purposes, a graphical representation to characterize a local structural neighborhood. This has a natural application to structure-based clustering. For each cluster we generate a substructural *signature*; that is, a MSA that discriminates for membership in the cluster. Signatures are related to the maximal common substructure, but do not necessarily include the MCS. The MCS tends to be very sensitive to outliers. Since it must be contained in every compound, an outlier that does not contain a large substructure common to most compounds in the set limits the MCS to a potentially much smaller common substructure. A large signature that is contained in most—but not necessarily every—compound often provides a better representation of the cluster. Cluster signatures also have application in comparing two compound sets. The cluster signatures generated from clustering one compound set can be used as *seed compounds* to analyze or select compounds from an external set.

Conclusion

We have presented a new method for constructing discriminating substructures by reassembling common medicinal chemistry building blocks. The algorithm can be parametrized to meet differing objectives: (1) to build features that discriminate for activity in a local neighborhood, (2) to build scaffolds for R-group analysis, (3) to construct cluster *signatures* that discriminate for membership in the cluster and provide a graphical representation for its members, and (4) to identify substructures that characterize major classes in a heterogeneous compound set. We illustrated the results of MSA generation on a literature dataset is of 118 compounds with in vitro inhibition data against recombinant human PTP-1B.

Three-dimensional techniques such as pharmacophore mapping and CoMFA-like analysis provide geometric models that are valuable aids in understanding and using the quantitative structure–activity relationships. However, the processes involved in building the models are complex and require expert assistance to obtain reliable results. In contrast the MSA analysis described here is simple to perform, provides results that are easy to interpret, and reveals the essential 2D structure–activity relationships in the experimental data. Thus,

MSA can provide useful information that complements that from the well-established 3D geometrical techniques.

Acknowledgment. Leadscope, Inc. would like to acknowledge partial financial support of Technology Action Fund Grant #TECH 02-066 from the State of Ohio, Department of Development: Chemical Genomics discovery platform with novel informatics methods to link genes to drugs.

References

- (1) Martin, Y. C. Pharmacophore Mapping. In *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*; Martin, Y. C., Willett, P., Eds.; American Chemical Society: Washington D. C., 1998; pp 121–148.
- (2) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C. et al. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251–3264.
- (3) Cramer, R. D. I.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. The Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
- (4) Barnard, J. M.; Downs, G. M. Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644–649.
- (5) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 93–996.
- (6) Chen, X.; Rusinko, A., 3rd; Tropsha, A.; Young, S. S. Automated pharmacophore identification for large chemical data sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 887–896.
- (7) Blower, P.; Fligner, M.; Verducci, J.; Bjoraker, J. On combining recursive partitioning and simulated annealing to detect groups of biologically active compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 393–404.
- (8) Raymond, J. W.; Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput. Aided Mol. Des.* **2002**, *16*, 521–533.
- (9) Wang, T.; Zhou, J. EMCSS: A New Method for Maximal Common Substructure Search. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 828–834.
- (10) Nicolaou, C. A.; Tamura, S. Y.; Kelley, B. P.; Bassett, S. I.; Nutt, R. F. Analysis of large screening data sets via adaptively grown phylogenetic-like trees. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1069–1079.
- (11) Miller, D. W. A chemical class-based approach to predictive model generation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 568–578.
- (12) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E., Jr. Leadscope: software for exploring large sets of screening data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302–1314.
- (13) Malamas, M. S.; Sredy, J.; Moxham, C.; Katz, A.; Xu, W. et al. Novel benzofuran and benzothiophene biphenyls as inhibitors of protein tyrosine phosphatase 1B with antihyperglycemic properties. *J. Med. Chem.* **2000**, *43*, 1293–1310.
- (14) Murthy, V. S.; Kulkarni, V. M. 3D-QSAR CoMFA and CoMSIA on protein tyrosine phosphatase 1B inhibitors. *Bioorg. Med. Chem.* **2002**, *10*, 2267–2282.

JM0302703