

Structural Unit Analysis Identifies Lead Series and Facilitates Scaffold Hopping in Combinatorial Chemistry

Philippa R. N. Wolohan,[†] Lakshmi B. Akella,^{†,§} Roman J. Dorfman,[†] Peter G. Nell,[‡]
Stefan M. Mundt,[‡] and Robert D. Clark^{*,†}

Tripos, Inc., 1699 S. Hanley Road, St. Louis, Missouri 63144, and Bayer HealthCare AG, Pharma R&D,
Discovery Research Europe, D-42096 Wuppertal, Germany

Received September 30, 2005

Combinatorial chemistry and high-throughput screening technologies produce huge amounts of data on a regular basis. Sieving through these libraries of compounds and their associated assay data to identify appropriate series for follow-up is a daunting task, which has created a need for computational techniques that can find coherent islands of structure–activity relationships in this sea. Structural unit analysis (SUA) examines an entire data set so as to identify the molecular substructures or fragments that distinguish compounds with high activity from those with average activity. The algorithm is iterative and follows set heuristics in order to generate the structural units. It produces graphs that represent a set of units, which become SUA rules. Finding all of the input structures that match these graphs generates clusters. The Apriori algorithm for association rule mining is adapted to explore all of the combinations of structural units that define useful series. User-defined constraints are applied toward series selection and the refinement of rules. The significance of a series is determined by applying statistical methods appropriate to each data set. Application to the NCI-H23 (DTP Human Tumor Cell Line Screen) database serves to illustrate the process by which structural series are identified. An application of the method to scaffold hopping is then discussed in connection with proprietary screening data from a lead optimization project directed toward the treatment of respiratory tract infections at Bayer Healthcare. SUA was able to successfully identify promising alternative core structures in addition to identifying compounds with above-average activity and selectivity.

INTRODUCTION

The combination of combinatorial chemistry and high-throughput screening (HTS) techniques produces large compound collections and reams of assay data. One challenge encountered is to identify hits from about 500 000 to 4 million compound libraries where the hit rate is typically 0.1–1% in the industry.² The nature of the hits obtained may vary from one screen to another. The reliability of HTS data is compromised by false positives (inactive compounds which show activity in the screen) as well as by false negatives (active compounds which fail to show activity). The techniques developed so far to analyze HTS data have proved to be effective on high-quality HTS data sets, but the derived models need to be validated on large data sets and noisy data.^{3,4} The computational approaches most frequently employed are based on pharmacophores, docking, or substructural similarity methods such as feature trees.⁵ In a feature tree, the connectivity and physicochemical properties of functional groups are captured in the nodes of a reduced graph. Comparison of feature trees is based on matching subtrees of two feature trees onto each other. A similar approach to similarity searching based on classical reduced graphs describes structures as topological pharmacophores.

Converting them to SMILES-type representations makes comparisons of reduced graphs possible.⁶

In practice, it is rare to find more than a few distinct chemotypes that have activity against a given target. Structural unit analysis (SUA) identifies distinct chemotypes and then finds the series of analogues defined by these chemotypes. Incorporating important information from nonhit compounds is virtually impossible using traditional methods because of their large numbers. This method utilizes both active and inactive compounds to delineate the structure–activity relationship within high-throughput data and thereby aids in scaffold hopping.

COMPUTATIONAL METHODS

Structural unit analysis is carried out in three steps: fragmentation, buildup of association rules, and significance analysis.

Fragmentation. The first stage involves fragmenting each molecule into substructures by removing rotatable bonds to nonrotatable parts of the molecule, using a procedure similar to that used in RECAP.⁷ Bonds in rings are untouched, whereas any bond to an atom that is in a ring is deleted so long as the bond is not itself in a ring. Bonds to terminal halogen atoms are also deleted, as are rotatable bonds between sp³ carbons and heteroatoms. Each fragment that remains constitutes a “structural unit.”

Association. A modified form of the Apriori algorithm developed for association rule mining⁸ is then applied that

* Corresponding author tel.: (314) 951-3365; e-mail: bclark@tripos.com.

[†] Tripos, Inc.

[‡] Bayer HealthCare AG.

[§] Current address: John F. Welch Technology Center, GE Global Research Center EPIP, Phase-2, Hoodi Village, Whitefield Road, Bangalore 560066, India.

examines all connected graphs formed by combining structural units present in the training set. Such graphs are built up by considering all unique combinations of structural units and identifying those that are found within some minimum number of compound structures, that is, that represent “rules” having at least some minimum support. SUA produces unit graphs wherein each rule describes a set of structural units as well as their connectivity. A unit graph is defined as a set of units and their connectivities, with each node in the unit graph corresponding to a specific structural unit. In order for a compound to match the unit graph, it must contain each element of the unit graph. By utilizing unit graphs, the connectivity and relative topological distances between various substructures within molecules are largely preserved.

The exact connectivity of the individual units is not necessarily preserved, however. Hence, a compound could match a unit graph, but the vertices connecting the individual units can differ. This is a clear limitation of the current algorithm and will be addressed in the future. To accommodate a variation in scaffolds while still maintaining a notion of connectivity and distance, the concept of a wildcard is also supported. The wildcard is a placeholder unit which, when inserted into the unit graph, will match any other single structural unit. The wildcard must match exactly one structural unit in the compound being considered, and all other connectivity requirements must still be met.

The algorithm is progressive; single units are examined initially, then combinations of two units, then combinations of three, and so on. At every subsequent step, unit graphs (“rules” v_i) that do not have the minimum support (specified as the number of actives) are eliminated, as are those for which the average activity is less than the average activity of the training set as a whole. The remaining rules make up the frontier set of rules from which the next batch of candidate rules is generated. The process continues until the specified maximum rule size has been reached.

The procedure can be formulated more formally as:

1. Let n_{\min} be the minimum number of actives (“support”) required for a rule to survive at any level, and let k_{\max} be the maximum unit graph size (complexity) that is of interest.

2. Let $\mathbf{U}_0 = \{u_i\}$ be the set of all unique structural units u_i generated from the molecules that make up the training set $\mathbf{S}_0 = \{S_j\}$, where S_j is the connected graph corresponding to molecule j , the nodes of which are individual structural units.

3. Let $a(S_j)$ be the activity of the compound corresponding to S_j ; let a_{\min} be the minimum activity required for a compound to qualify as “active”.

4. Let $\bar{a}(v_i)$ be the average activity across the set $\mathbf{S}(v_i) = \{S_j: v_i \subseteq S_j\}$. Note that v_i and S_j are graphs, so “ \subseteq ” in this context means “is contained in”.

5. Set $k = 1$, and create the first (and simplest) set of rules:

$$\mathbf{V}_0 = \{u_i: |\mathbf{S}(u_i)| \geq n_{\min} \text{ and } \bar{a}(u_i) > \bar{a}_0\}$$

6. Create the next set of candidate rules by augmenting each rule (graph) from the previous step by one structural unit:

$$\mathbf{U}_k = \{u_i: u_i \in \mathbf{U}_0\} \times \{v_j: v_j \in \mathbf{V}_{k-1}\}$$

7. Filter out rules that do not have enough support or are not discriminating enough:

$$\mathbf{V}_k = \{v_i: v_i \in \mathbf{U}_k \text{ and } |\mathbf{S}(v_i)| \geq n_{\min} \text{ and } \bar{a}(v_i) > \bar{a}_0\}$$

8. If $k = k_{\max}$, stop; otherwise, increment k and go to step 6.

Significance Analysis. SUA generates a large number of potential combinations of units and thereby explores many potential chemical series. The third and final stage of the method entails determining the significance of each SUA rule in some \mathbf{V}_k with respect to activity. For every combination v_i of units that is generated by the Apriori algorithm, the data set is split into those molecules that contain this unit graph $[\mathbf{S}(v_i)]$ and those that do not $[\mathbf{S}_0 - \mathbf{S}(v_i)]$. The probability that the observed distribution among the two sets arises by chance is then estimated using one of three statistical tests: nonparametric Kruskal–Wallis (K–W), χ^2 , or parametric ANOVA. The K–W test, which is based on ranks rather than individual data values, is the most generally applicable and is used by default. It works well for both normal and non-normal distributions. In cases where the distribution is known to be normal, however, ANOVA is a more powerful test. In practice, the difference between the results obtained using ANOVA and K–W was usually minimal. The χ^2 test is the significance test of choice for data with large numbers of tied activity values.

The series are ranked on the basis of the rank enrichment relative to the data set as a whole, and the best series are selected for detailed examination.

RESULTS AND DISCUSSION

Application of the Method to the NCI Data Set. To test the SUA methodology, we used the publicly available NCI-H23 database, which is comprised of 35 985 compounds, tested at five different concentrations against the 60 human cancer cell lines used by the NCI anticancer drug discovery program (http://dtp.nci.nih.gov/docs/cancer/cancer_data.html). The anticancer activity is expressed as GI_{50} , the concentration required to inhibit 50% of cell growth. Benchware HTS DataMiner⁹ was used to identify several scaffold classes within the NCI-H23 data set, and the corresponding derivatives served as a proof of concept application. The actives were selected using an activity cutoff of GI_{50} greater than or equal to 6. This resulted in a subset of 1806 active compounds. Several chemotypes were selected from among these by considering their calculated properties, that is, $\log P$, solubility, and general drug likeness.¹⁰ Substructure searches and maximum common substructure searches were conducted on these scaffolds, and the resulting sets of structures were used to retrieve structurally related inactives from the NCI-H23 database. This resulted in a working data set of 2013 compounds comprising eight distinct scaffold classes with GI_{50} 's ranging from -1.86 to 10.81 .

SUA was then used to identify structural series. The program found 349 compounds having a $\text{GI}_{50} \geq 6$ in the data set and was able to use all of the 2013 compounds for data analysis. The minimum number of members to form a series was set to 3, and default values of a minimum of two actives per cluster and a range of 1–3 structural units per series were set. No wildcards were allowed.

Table 1

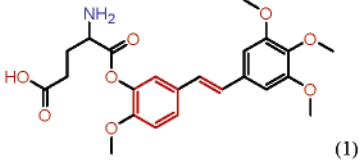


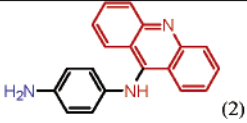

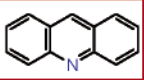
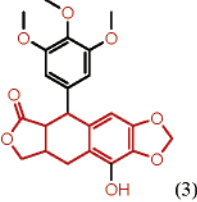
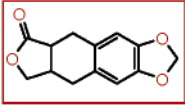
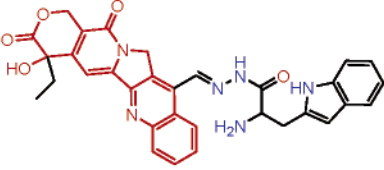
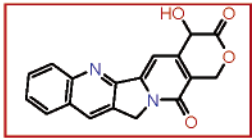
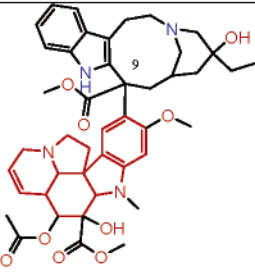
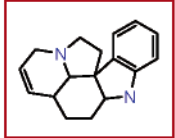
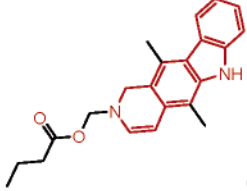
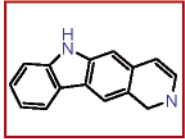
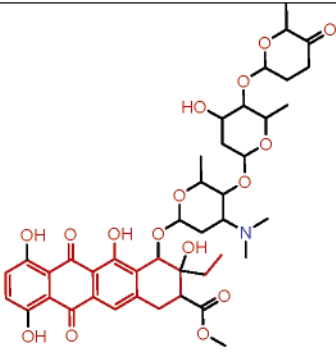
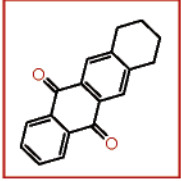

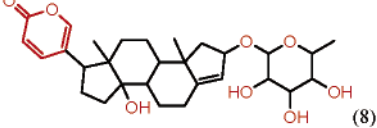
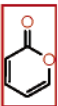
Rule	Example	Structural Units	Connectivity
R91	 (1)	 	A-B
R194	 (2)	 	A-B
R72	 (3)		A
R35	 (4)		A
R36	 (5)		A
R39	 (6)		A
R156	 (7)	 	A-B
R69	 (8)		A

Table 2

rule	series	total hits	mean activity (GI ₅₀)
R91	combretastatin derivatives	40	6.62
R194	acridine derivatives	58	6.47
R72	podophyllotoxin analogues	96	6.42
R35	camptothecin derivatives	5	7.55
R36	vinblastine analogues	9	7.36
R39	ellipticine analogues	25	6.05
R156	anthracycline derivatives	21	7.60
R69	scillaridin analogues	5	7.69

This data set is skewed both in terms of the size of each compound class and in terms of the percentage of actives in each compound class. Therefore, the nonparametric Kruskal–Wallis test with a confidence threshold of 0.95 was selected as the test for the significance of the series. In addition, rules with substantial membership overlap (redundant rules) and rules with fewer than a five heavy-atom count were filtered out.

The analysis was complete in less than 5 min on 2013 compounds and generated 25 rules based on the analysis parameters specified. Upon manual examination of the 25 rules, eight were found to represent distinct structural classes within the database. The resulting SUA rules, the number of hits associated with each rule, and example structures from the corresponding series identified are shown in Tables 1 and 2.

Scaffold Hopping. To elucidate the capabilities of SUA for the analysis of structure–activity relationships across sets of compounds with structurally diverse scaffolds, screening data from a lead optimization project for the treatment of respiratory tract infections (RTIs) at Bayer Healthcare were analyzed using the SUA implementation of Bayer's proprietary Pharmacophore Informatics (PIx) platform.^{11–13} This

data set contains approximately 1600 compounds that were tested in a panel of eight pathogens that included gram-negative as well as gram-positive bacteria. The measured minimal inhibitory concentration values of the different pathogens were normalized to range from 0 to 100, with 0 representing compounds with the highest activity. Values of 30 or higher are considered to be weakly active or inactive. Approximately 90% of all single pharmacological data points for the compounds were available; missing values were treated as average values. The goal of the project was to identify compounds with a broad antibacterial profile covering all major respiratory tract pathogens.

A simple approach would be to consider the average activity value of all eight pathogens as the activity value for the SUA algorithm. However, this approach is optimal only when the individual pathogen assays are statistically independent. Also, using a simple average may not distinguish a compound with good activity against all species from a compound with mediocre activity against most species but high activity against one particular species; that is, it will not reflect differences in variation across assays.

For the example described here, we created a more discriminating parameter for the selectivity profile by applying principal components analysis. The first principal component reflects the average across treatments, whereas the second principal component reflects the differences among them. By calculating a weighted average from the summed coefficients for the first two principal components, we can derive a selectivity factor that takes both mean activity and variation in activity across pathogens into account. For this data set, a low value of the selectivity factor parameter represents consistently high activity against all pathogens. The histogram shown in Figure 1 represents the distribution of the calculated selectivity factor across the data

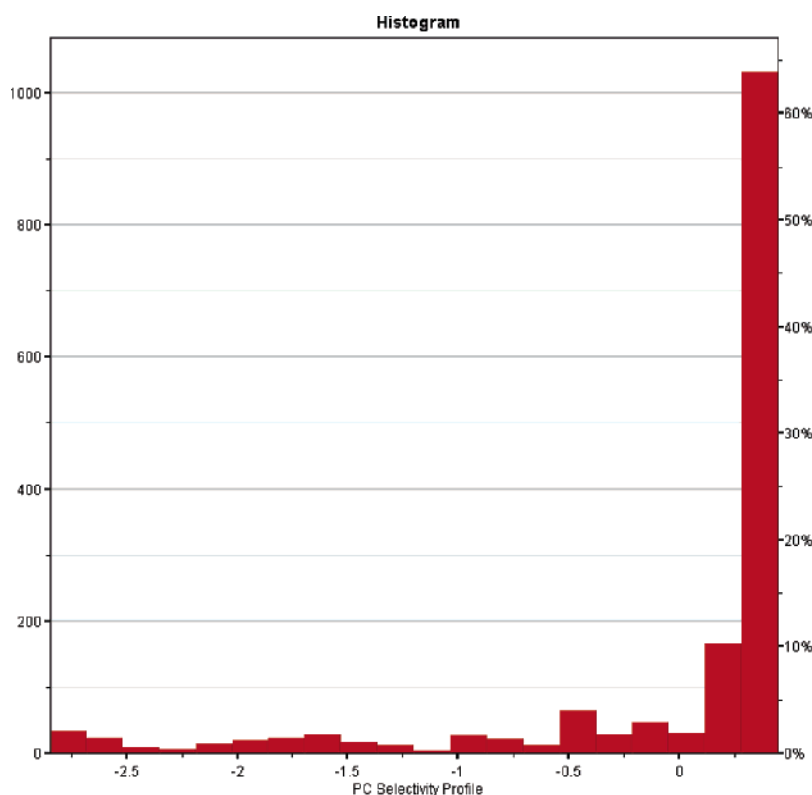


Figure 1. "PC Selectivity" profile of data set.

Table 3

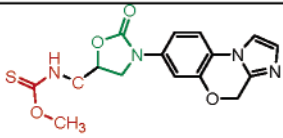
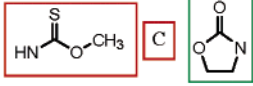
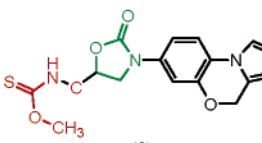
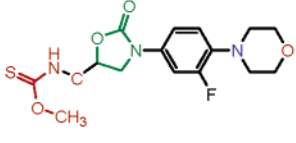
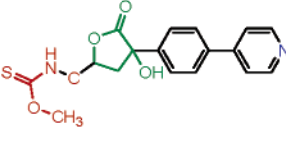
Rule	Example	Structural Units	Connectivity
R122	 (9)		A-B-wild

Table 4

Example 1 Oxazolidinone core	Example 2 Oxazolidinone core	Example 3 Lactonol core
 (9)	 (10)	 (11)

set. As has already been mentioned, to make a statistical evaluation by the algorithm possible, data sets should contain both active and less-active or inactive compounds, represented in this case by compounds with an undesirable selectivity profile.

The SUA analysis shown in Table 3 was carried out using the following parameters: rules with two to six units were allowed, with up to two wildcards in each, and an activity cutoff at 15% of the most "active" compounds was used. One of the six resulting rules (R122) was comprised of three structural units (two units colored red), the third being a wildcard (green). The average normalized activity against all pathogens of the five members of this rule was 6.7, all showing a better than average selectivity profile. Table 4 shows one series that was identified. Looking at the structures of these compounds revealed that the lactonol core (**11**) could substitute for the oxazolidinone ring (**9**) without sacrificing activity or selectivity.

Clustering based on substructural fingerprints does not indicate that compounds **9** and **11** are closely related, despite the fact that only minor changes at the core following the bioisosteric principle are present. In terms of intellectual property, the value of scaffold hopping from oxazolidinone to other cores is potentially very interesting and of considerable value.

CONCLUSIONS

SUA provides an important tool for the analysis of HTS data; the work presented here illustrates the effectiveness of SUA for series selection and for identifying alternative structural units or cores for scaffold hopping. Eight distinct structural classes were successfully identified from a sample of the NCI-H23 data set. The mean activity of each of these series is higher than the average activity of the data set. The ability to identify series with promising activity for further testing is extremely valuable to a scientist faced with the daunting task of compound selection from raw HTS results. Through the use of placeholder wildcard units, SUA provided a means to select potential substitutes for the oxazolidinone core of active compounds against RTI. The approach of sorting the data set by average activity and manual inspection might be an alternative route for small data sets with only a

few parameters to be optimized, but for the given set of 1600 compounds, it would have been prohibitively time-consuming or impossible to identify different promising cores in this way. Moreover, it would have been very difficult to determine their relevance in terms of activity or to correlate particular fragments with activity and selectivity. Other promising cores identified during these studies are currently under investigation at Bayer Healthcare. The heuristics of SUA addresses the need to provide an intelligent way to extract lead series from screens and to set future synthetic direction. Selecting series rather than individual compounds with above-average activity is an effective way to prioritize lead optimization efforts, whereas recognizing alternate cores opens new avenues to pursue for the medicinal chemist.

ACKNOWLEDGMENT

The authors thank Andreas Goeller, Jill Wood, and Michael Brands (Bayer HealthCare AG); Stephan Reiling (Aventis Pharmaceuticals); and Steven Burkett (Tuxedo Park Racing, formerly Tripos, Inc.) for their contributions to the successful development of SUA.

REFERENCES AND NOTES

- Reiling, S. U.S. patent (pending) US20040024531, 2004.
- Bocker, A.; Schneider, G.; Teckentrup, A. A Status of HTS Data Mining Approaches. *QSAR Comb. Sci.* **2004**, *23*, 207–213.
- Lengauer, T.; Lemmen, C.; Rarey, M.; Zimmermann, M. Novel Technologies for Virtual Screening. *Drug Discovery Today* **2004**, *9*, 27–34.
- Saeh, J. C.; Lyne, P. D.; Takasaki, B. K.; Cosgrove, D. A. Lead Hopping Using SVM and 3D Pharmacophore Fingerprints. *J. Chem. Inf. Model.* **2005**, *45*, 1122–1133.
- Rarey, M.; Dixon, J. S. Feature Trees: A New Molecular Similarity Measure Based on Tree Matching. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471–490.
- Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338–345.
- Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP—Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.
- Agrawal, R.; Imieliński, T.; Swami, A. Mining Association Rules between Sets of Items in Large Databases. *Proc. Conf. On Management of Data*; ACM Press: New York, 1993; pp 207–216.
- Benchmark HTS DataMiner is distributed by Tripos, Inc., 1699 S. Hanley Road, St. Louis, Missouri, 63144, U.S.A.

- (10) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, 23, 3–25.
- (11) Pharmacophore Informatics (PIx) platform. Bayer HealthCare AG, Pharma R&D, Discovery Research Europe, D-42096 Wuppertal, Germany.
- (12) Scott, W. J.; Weigand, S.; Nell, P. G.; Wirtz, S.; Lohrmann, E.; Brunne, R.; Mittendorf, J. Integrated Approaches to Informatics: Bayer Healthcare Pharmacophore Informatics Platform, Part 1: Document Handling, Project Support and Portfolio Management. Presented at the 229th ACS National Meeting, San Diego, CA, March 13–17, 2005, CINF 97.
- (13) Nell, P. G.; Haerter, M.; Brunne, R.-M.; Scott, W. J.; Mundt, S.; Goeller, A.; Wood, J.; Reiche, F.; Ruppelt, M.; Mittendorf, J. Integrated Approaches to Informatics: Bayer Healthcare Pharmacophore Informatics Platform, Part 2: Data Integration, Analysis and Visualization. Presented at the 229th ACS National Meeting, San Diego, CA, March 13–17, 2005, CINF 98.

CI050432Z