

HierS: Hierarchical Scaffold Clustering Using Topological Chemical Graphs

Steven J. Wilkens,* Jeff Janes, and Andrew I. Su

Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, California 92121

Received November 30, 2004

An exhaustive ring-based algorithm, HierS, has been developed in order to provide an intuitive approach to compound clustering for analyzing high-throughput screening results. The recursive algorithm rapidly identifies all possible ring-delimited substructures within a set of compounds. Molecules are grouped by shared ring substructures (scaffolds) so that common scaffolds obtain higher membership. Once all of the scaffolds for a set of compounds are identified, the hierarchical structural relationships between the scaffold structures are established. The complex network of hierarchical relationships is then utilized to navigate compounds in a structurally directed fashion. When the scaffold hierarchy is traversed, over-represented structural features can be rapidly identified so that excess compounds that contain them can be removed without significantly impacting the structural diversity landscape of the compound set. Furthermore, the removed compounds can provide the opportunity to follow-up on active compounds that had previously been discarded because of practical limitations on follow-up capacity. A Web-based interface has been developed that incorporates this algorithm in order to allow for an interactive analysis. In addition, biological data are coupled to scaffolds by the inclusion of activity histograms, which indicate how the compounds in each scaffold class performed in previous high-throughput screening campaigns.

Introduction

The recognition and classification of shared chemical features present in the large and diverse sets of compounds identified as hits by high-throughput screening (HTS) are important and difficult tasks. Effective feature identification facilitates the decisions that affect the allocation of resources throughout the lead discovery process. To this end, a number of approaches have been developed to structurally classify compound hit lists from HTS campaigns.^{1–4} The identification of structurally related biologically active compounds enables scientists to focus follow-up efforts on representatives from each set, which can maximize the number of candidate scaffolds classes available for optimization and minimize the chances that a potentially desirable scaffold is overlooked. In addition, effective compound clustering can also be used to identify biologically promiscuous chemical features or experimental artifacts such as fluorescent chemical groups or reactive impurities from combinatorial chemistry.

Judging the relative performance of two chemical clustering methods presents significant challenges because the criteria for evaluation are context-dependent and often subjective. Results are ultimately judged by how well the compounds chosen for follow-up experiments perform throughout the lead discovery process. Experience has shown that medicinal chemists prefer clustering methods whose results are readily interpretable and simple to explain. Therefore, the utility of a clustering method is measured, at least in part, by how well its results agree with chemical intuition. Clustering methods often use abstract molecular representations, such as molecular fingerprints,^{5,6} which reduce chemical structures to a bit string of descriptors. However, these types of approaches can result in chemists spending

more time trying to understand why compounds were clustered in a particular way rather than using the clustering data to identify and understand molecular patterns in their data. Unfortunately, intuitive chemical concepts⁷ may be extremely difficult to implement or computationally intensive. Conversely, programmatically amenable concepts, such as clustering methods that utilize molecular fingerprints, are not as useful for the conceptual understanding of clustering data.

In this paper, we present the HierS package, which employs a fast and straightforward algorithm for clustering compounds by their explicit topological chemical graphs and a Web-based user interface that allows for the visualization and navigation of clustering results. More specifically, HierS employs an unsupervised algorithm that constructs hierarchical relationships between ring features. The ring system “scaffolds” that are generated provide a highly relevant means by which to visualize chemical classes because ring-based linkages are central structural features in most drug molecules. This was shown by Lewell et al.,⁸ who recently noted that of the 10 000 development compounds in PJB’s Pharmaprojects,⁹ 96% of the compounds contain rings. Of those, 56% of the molecular weight is accounted for by atoms in rings.

The HierS algorithm is related to other applications that have been developed to group compound data found to be active in HTS experiments. The method presented by Roberts et al. differs from the HierS algorithm in that it uses a predefined hierarchy of over 27 000 chemical features to classify compounds and so may not be optimally adapted for novel structural motifs.¹ Alternatively, the algorithm used by Tamura et al. identifies maximum common substructures that appear to be significant in conferring activity to a group of molecules, which leads to a flat categorization as opposed to a navigable hierarchy.³ The algorithm presented by Miller identifies ring-system scaffolds consisting of two to four rings, which are then used to build a predictive recur-

* To whom correspondence should be addressed. Kalypsys, Inc., 10420 Wateridge Circle, San Diego, CA 92121. Phone: 858-754-3453. Fax: 858-754-3301. E-mail: swilkens@kalypsys.com.

sive partitioning model that consists of a statistically rather than structurally determined hierarchy.⁴ HierS also differs from all three programs in that it employs a Web-based user interface.

At later stages of the lead discovery process the large hit lists produced by HTS experiments can burden biologists with large numbers of compounds for follow-up experiments. Because the goal of HTS is to produce candidate molecular scaffolds for lead optimization, it is often unnecessary and inefficient to perform follow-up experiments on molecules whose scaffold structure is already well represented in the group of hits. Also, practical considerations may limit the number of compounds that can be used in follow-up experiments. The efficiency of the lead discovery process can be improved by identifying compounds in scaffold classes that are over-represented or deemed to be undesirable by medicinal chemists. However, automatic flagging of compounds with problematic chemistry provides a significant challenge because the judgment of the chemists can vary from project to project.⁷ Considerations for advancement of compounds through the drug discovery pipeline include the biology and value of the target, existing intellectual property, existing preliminary structure–activity relationships (SAR), and the number of molecules considered to have good chemistry potential that were identified as hits in a given screen. To help facilitate the compound selection process, HierS was designed to identify significant structural features and to couple them to historical biological activity, which allows biologists and medicinal chemists to rapidly filter through hit lists consisting of thousands of compounds in order to identify patterns in their data.

Methods

Scaffold Building Algorithm. The ring-based structural analysis procedure described here is based on previously developed concepts.^{10,11} Molecules are composed of three components: ring systems (ring), side chain bonds and atoms (chain), and linking bonds and atoms (linker). Atoms that are external to a ring but are bonded to a ring atom with a bond order greater than 1 are considered to be part of the ring system because they modify the nature of the ring. For the special case where a molecule does not contain a ring, chain bonds and atoms are trimmed until a double or triple bond is encountered. Atoms that are double-bonded to linker atoms are also considered to be part of the linker because they can modify the nature of the linker (e.g., the carbonyl in a peptide linkage significantly increases the rotational barrier of the C–N bond).

The set of basis scaffolds for a given molecule is defined as the structures that result from the removal of all linkers and chains. In other words, the basis scaffolds for a molecule are the set of all unique ring systems in the molecule, where a ring system is defined as one or more rings that share an internal bond. Ring systems consisting of a single benzene ring are not included in the set of basis scaffolds because they are too ubiquitous to be considered a discriminating feature. The superscaffold for a molecule is determined by deleting only the chains. In the special case where a molecule consists of zero or one ring system, the basis scaffold is the superscaffold. Figure 1A shows the p38 MAP kinase inhibitor BIRB 796,¹² along with its basis

scaffolds (Figure 1B) and superscaffold (Figure 1C). Generally, the basis and superscaffolds by themselves do not sample chemical space at sufficient granularity to be useful for clustering compounds. Therefore, for finding structural patterns in a set of molecules, a number of intermediate scaffolds are necessary to identify shared significant features within the set.

A recursive algorithm is used to elucidate all candidate scaffold structures including those derived from exhaustive combinations of the basis scaffolds. The process begins by trimming all chains to reveal the superscaffold (Figure 1C). If the scaffold is novel, it is added to the list of scaffolds. Next, HierS identifies all basis scaffolds contained in the superscaffold. If the number of basis scaffolds is less than or equal to 1, HierS continues to the next scaffold candidate because no smaller scaffolds exist for the fragment being processed. If the number of basis scaffolds is greater than 1, HierS generates new compound fragments by deleting each ring system present in the scaffold being processed. Then the resulting fragments are used as input for the first step and HierS continues looping until all possible ring combinations have been identified. In other words, HierS recursively removes each ring system from the superscaffold to generate fragments that contain all possible ring system combinations. Finally, the process is completed by adding the compound to the list of member compounds for each scaffold. Figure 1D shows the scaffolds consisting of all two-ring system combinations, and Figure 1E shows the scaffolds consisting of all possible three-ring system combinations.

After all the distinct basis and multiring system scaffolds for each molecule in the input list have been identified and added to the scaffold list, hierarchical structural relationships between the scaffold classes are established. Scaffolds that contain another scaffold as a substructure are said to be “derived” from that scaffold. To build the scaffold hierarchy, a superstructure search against all other scaffolds in the list is performed for each scaffold. If a target scaffold is found to be a substructure of the query scaffold, the query scaffold is added to the list of scaffolds that are derived from the target scaffold. In this way, all possible hierarchical relationships between scaffolds are identified. Figure 2 shows an example of the hierarchical relationships between the scaffolds of BIRB 796. Scaffolds that are lower in the hierarchy in Figure 2 are derived from scaffolds that are higher in the hierarchy as shown by the connecting arrows in Figure 2. Establishing a structural hierarchy in this fashion builds a framework for navigating the chemical features of a set of compounds in a structurally directed manner. Scaffolds are navigated from small general features to larger features, which are composed of various combinations of the smaller elements. Figure 3 shows the data flow diagram for both the scaffold identification and hierarchy building processes.

Once all scaffolds are identified and their hierarchical relationships are established, HierS can rapidly traverse the network of structural connections to determine scaffold class membership. That is, for a fixed set of input compounds, cluster membership is determined by inspecting the connections between scaffolds along the hierarchy. A compound belongs to all scaffold classes

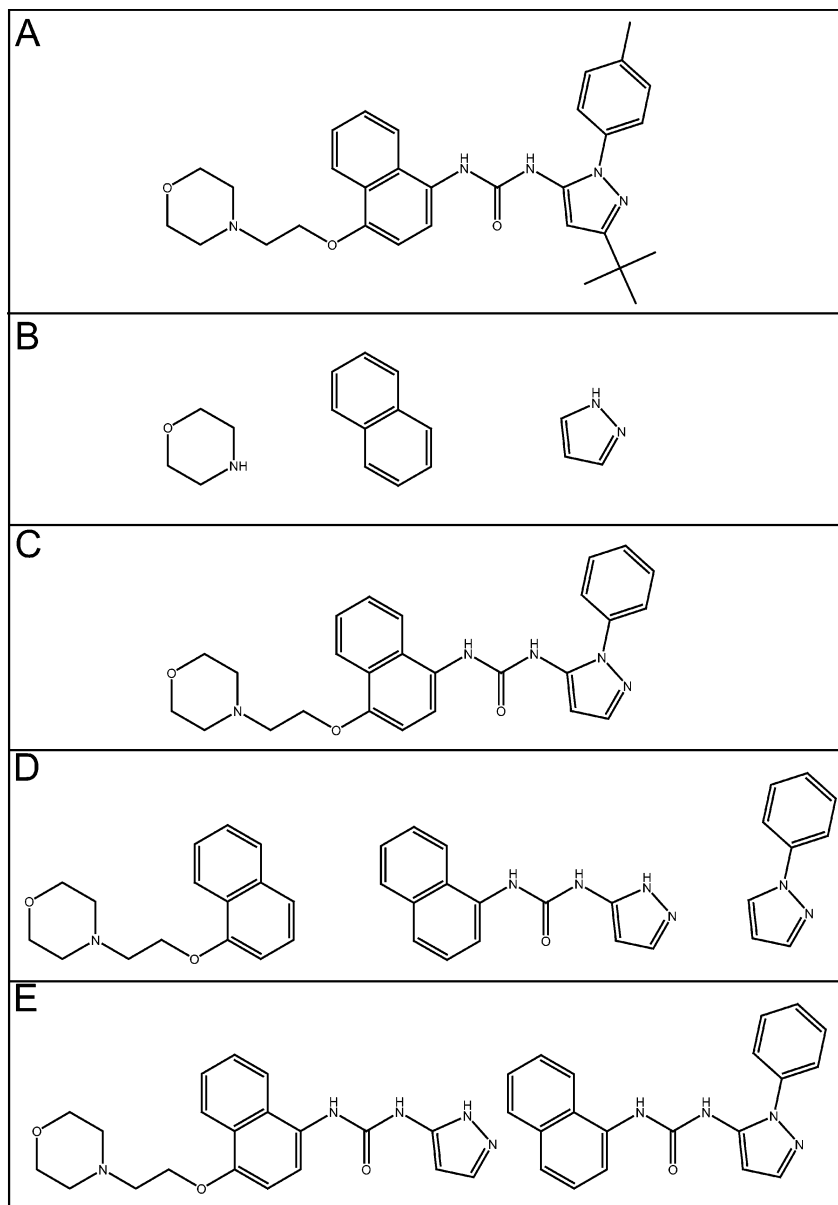


Figure 1. Scaffold structures for the p38 MAP kinase inhibitor BIRB 796:¹² (A) BIRB 796; (B) basis scaffold structures; (C) superscaffold structure; (D) all two ring system scaffolds; (E) all three ring system scaffolds.

that are substructures of it, which means that a given compound can belong to several scaffold classes. Moreover, if a compound belongs to a given scaffold class, then it must also belong to all the scaffold classes from which that scaffold was derived. If a scaffold has only a single member compound and its member compound is a member of one or more other scaffold classes, the scaffold is considered to be “redundant” and is removed from the list of scaffolds. Such scaffolds serve to increase the complexity of the scaffold network but provide no informational value to the overall set of scaffolds.

From Figures 1 and 2, it appears that the effect of scaffold identification and hierarchy building adds unnecessary complexity rather than simplifying structural analysis. This, of course, is the exact opposite of the desired outcome of a clustering analysis. While the number of unique scaffolds for a given set of compounds is often many times greater than the number of initial compounds, a significant portion of the scaffold structures are discarded because they are redundant. Furthermore, commonly occurring scaffold structures will

gain significant membership, which highlights their importance. Also, this hierarchical complexity is hidden behind a simple “drill-down”, which provides an intuitive and organized method for inspecting the scaffold landscape. The distribution of scaffold membership varies as a function of structural diversity of the input compound set. In addition, the number of distinct scaffolds is a function of the number of distinct ring systems and the linkers that connect them.

Average Pairwise Tanimoto (APT). While the scaffolds built by HierS provide an intuitive method for organizing diverse compound sets by shared topological features, they do not provide a means for assessing the overall structural similarity between the compounds in a given scaffold class. In other words, the scaffolds describe localized topological features that are shared in a set of compounds while the remaining features in the molecules are ignored. This is a significant issue given that a compound can have membership in a number of scaffold classes, which can lead to an arbitrary decision as to which scaffold or scaffolds provide

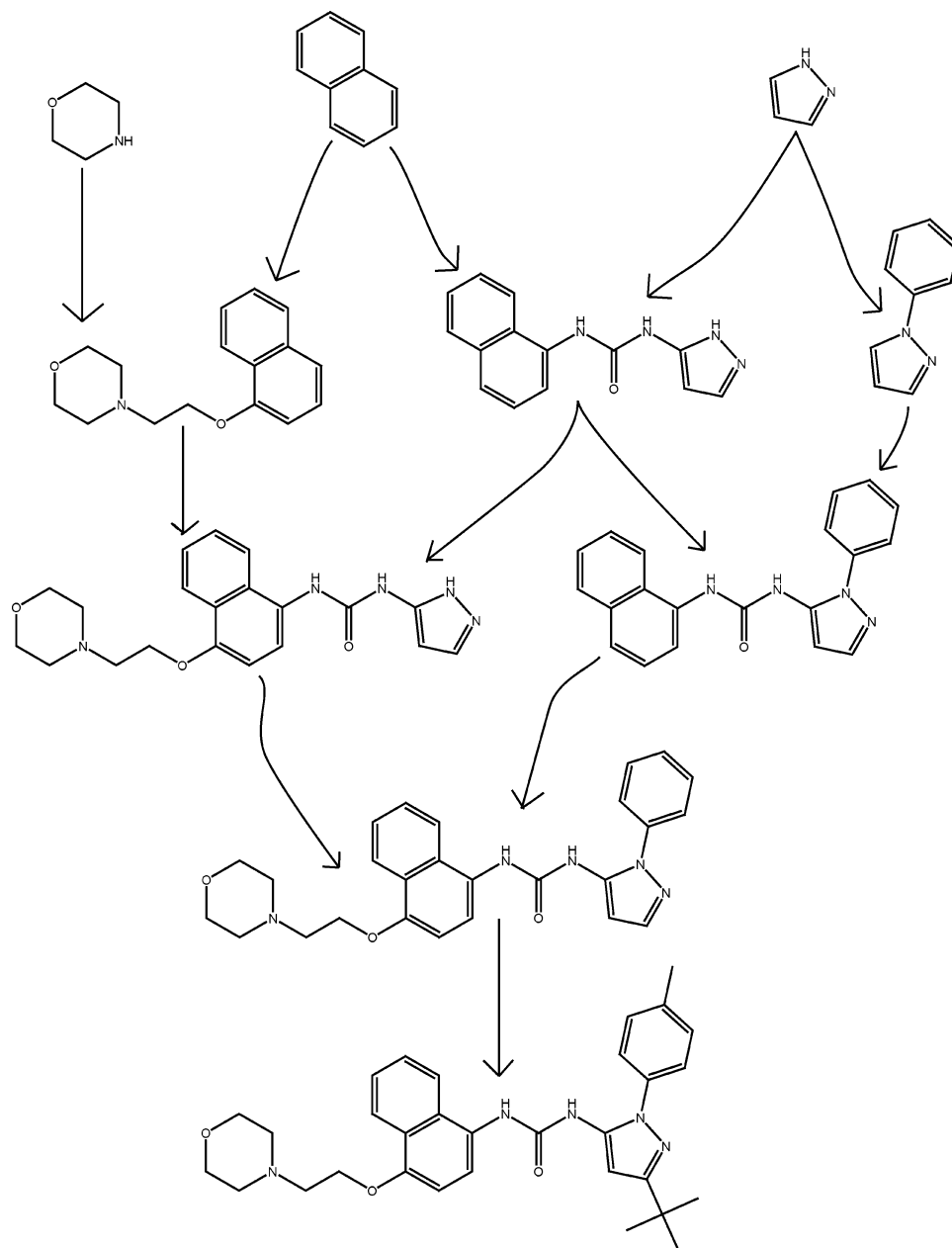


Figure 2. Hierarchy of scaffolds for the p38 MAP kinase inhibitor BIRB 796.¹²

the “best” overall representation of the compound set. To remedy this, an average pairwise Tanimoto coefficient (APT) is computed for each scaffold set using molecular fingerprints computed by JChem,¹³ which are closely related to Daylight fingerprints.¹⁴ The APT for a scaffold is determined by summing the computed Tanimoto coefficients¹⁵ between each pair of compounds in a scaffold group and dividing by the number of pairs of compounds in the group as shown in eq 1:

$$\text{APT}_i = \frac{1}{n(n-1)} \sum_{j \neq k}^n \frac{N_{jk}}{N_j + N_k - N_{jk}} \quad (1)$$

where n is the total number of member compounds in scaffold i , N_j is the number of “on” bits in compound j , N_k is the number of “on” bits in compound k , and N_{jk} is the number of “on” bits in common between compounds j and k . The APT provides a convenient means for approximating overall topological similarity because

many chemists are already familiar with the concept of a Tanimoto coefficient. In addition, it is a simple scalar metric by which scaffolds can be filtered and ranked.

The APT effectively provides a summary of how closely related the overall topological structures of the compounds in a given scaffold class are to each other. For example, if a given scaffold is small (e.g., a single pyrimidine), the compounds within that class may be quite diverse given that the scaffold accounts for a relatively small portion of each of the molecules in that scaffold group (assuming the compounds in the group are druglike in size). As a result, the APT coefficient for that scaffold class will likely be small. A low APT coefficient may also result if a scaffold class contains other more tightly clustered scaffold classes that are further down the structural hierarchy. In general, the APT increases as the scaffold size increases because larger scaffolds represent a higher proportion of the structural features of the molecules in that scaffold set.

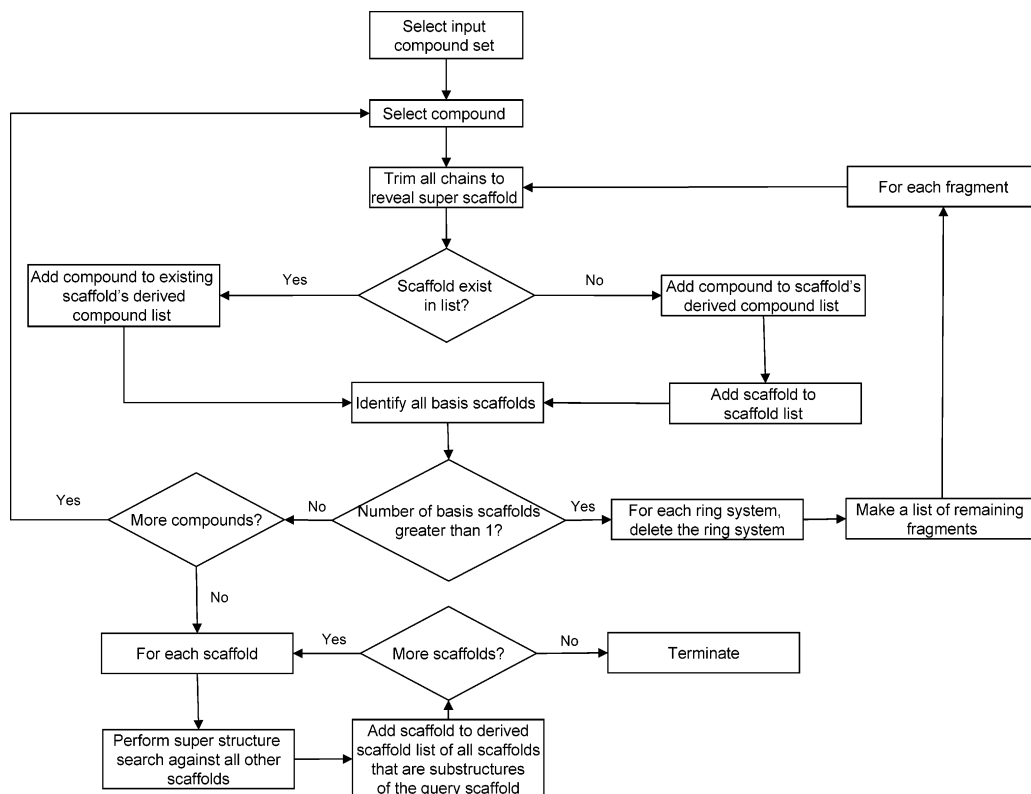


Figure 3. Flow diagram for identifying all scaffolds and for building the hierarchical structural relationships between the scaffolds.

As a result, APT values can be expected to increase as one proceeds down a given branch in the scaffold hierarchy. Nevertheless, APT values can be a useful metric for comparing scaffold classes at similar levels in the hierarchy in terms of the relatedness of their member compounds.

Automated Identification of Over-Represented Scaffolds. In addition to providing a simple method for approximating the overall topological similarity between compounds in a given scaffold class, the APT coefficient can also be used as a metric to gauge scaffold over-representation in a set of compounds. To identify over-represented scaffolds, HierS first builds a list of all scaffolds that exceed a user-defined APT criterion (e.g., 0.80). Next, HierS sorts the list by ascending molecular weight and then proceeds down the list and inspects each scaffold to see if it is derived from a scaffold that precedes it in the list. Any scaffold in the list of over-represented scaffolds that is found to be derived from a higher ranking (i.e., lower molecular weight) scaffold is removed because all of the compounds that have membership in such scaffolds are already accounted for by the higher ranking scaffold. In other words, the compounds that are members of a scaffold that is derived from an over-represented scaffold are already implicitly accounted for by the scaffold from which it is derived. Therefore, the derived scaffold can be removed from the list of over-represented scaffolds because it is redundant. The final list of scaffolds can be used as query structures to investigate scaffold enrichment in a given screen. Or compounds can be selected for removal from over-represented scaffold classes in order to reduce the size of a compound set while minimizing the loss of chemical diversity, as discussed below.

Implementation. This HierS algorithm has been implemented using version 1.4 of the Java language.¹⁶ Pro-

grammatic metaphors for chemical entities such as atoms, bonds, and molecules are provided by version 2.2.1 of the JChem package from ChemAxon.¹³ In addition, the atom by atom matching functionality implemented by JChem is used for both the scaffold and hierarchy building sections of the algorithm. The performance of the hierarchy building process is improved by several orders of magnitude by utilizing molecular fingerprint matching. HierS computes and caches 1024-bit fingerprints as needed by using the JChem fingerprinter. Because of this optimization, the hierarchy building process scales linearly to tens of thousands of input compounds. For a set of about 2000 compounds, the process of identifying all the scaffolds and building the hierarchy takes about 6 min on a 2.5 GHz Xeon processor. Because HierS is implemented as a Web application, it is immediately available on all networked computers.

HierS also makes use of three open source components. The Apache Jakarta Tomcat¹⁷ (version 5.0) servlet container and the Struts¹⁸ Model View Controller package (version 1.1) provide the Web application framework. In addition, the JFreeChart¹⁹ package (version 0.9.16) was used to generate chart images. HierS is available for downloading at <http://www.gnf.org/publications/hiers/>.

Results and Discussion

Test Data Preparation. To illustrate the utility of the application, we use the September 2003 release of the Developmental Therapeutics Program (DTP) Human Tumor Cell Line Screen set from the National Cancer Institute (NCI) and National Institutes of Health (NIH).²⁰ The 42 000 compounds in this set were first structurally standardized, and salts were removed. In addition, compounds with invalid structures, compounds containing metals, and compounds with a molecular

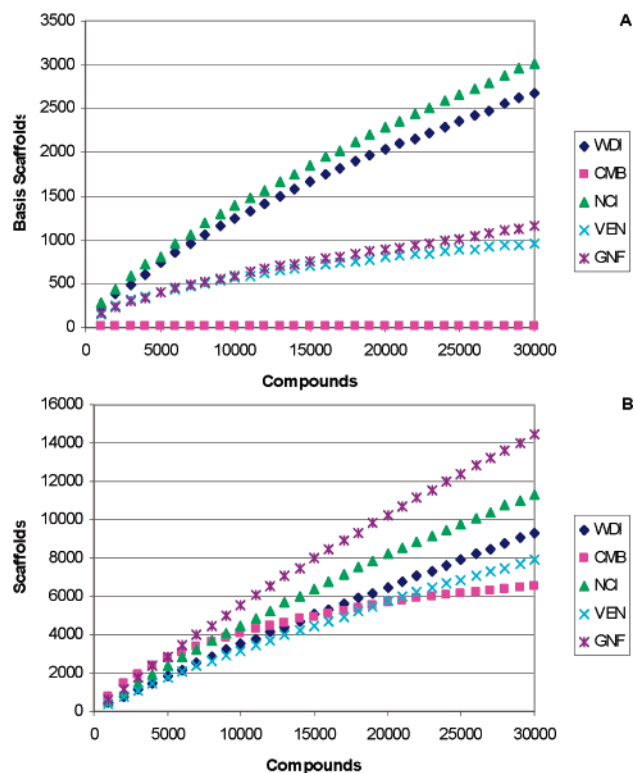


Figure 4. Accumulation of distinct scaffolds as a function of compounds selected at random in groups of 1000: (A) accumulation of basis scaffolds; (B) accumulation of all scaffolds (basis, super, and all intermediate scaffolds).

weight greater than 800 Da were removed. About 32 000 compounds remained after filtering.

Other compound sets used include the 2003.2 edition of the World Drug Index (WDI), an in-house-developed combinatorial compound library (CMB), a vendor com-

pound collection (VEN), and the portion Genomics Institute of the Novartis Research Foundation (GNF) compound collection used for high-throughput screening (GNF). These compound sets were processed and filtered in a fashion similar to the NCI compounds.

Scaffold Accumulation. Figure 4 displays how distinct scaffolds accumulate as compounds are selected at random from the five different compound collections (NCI, WDI, CMB, VEN, and GNF). From each set of 30 000 input compounds (which were chosen at random from their respective sources), groups of 1000 compounds were selected at random and their scaffolds were computed. The list of scaffolds from the new set of compounds was then compared with the existing set, and the novel scaffolds were added to the existing list. In Figure 4A, the total unique basis scaffold count is plotted as a function of the number of compounds for each of the five sets. From Figure 4A, it is apparent that the NCI and WDI sets accumulate unique basis scaffolds much more quickly than the other three sets. After all compounds are sampled, they have accumulated about 3000 and 2600 basis scaffolds, respectively. This is because the compounds in the WDI and NCI sets tend to be composed of fewer large ring systems (e.g., natural products) rather than a number of small ring systems. In contrast, the CMB set is quickly saturated at 21 unique basis scaffolds in the first set of 1000 compounds. This is, of course, because the CMB compound set is a combinatorial library.

Figure 4B shows the accumulation of all scaffolds (i.e., basis, super, and all intermediate scaffolds) from each of the five compound sets. Unlike Figure 4A, the GNF compound set accumulates the most unique scaffolds (about 14 000), despite the fact that it accumulated far fewer basis scaffolds than the NCI and WDI sets. Although the GNF set has fewer basis scaffolds, it has

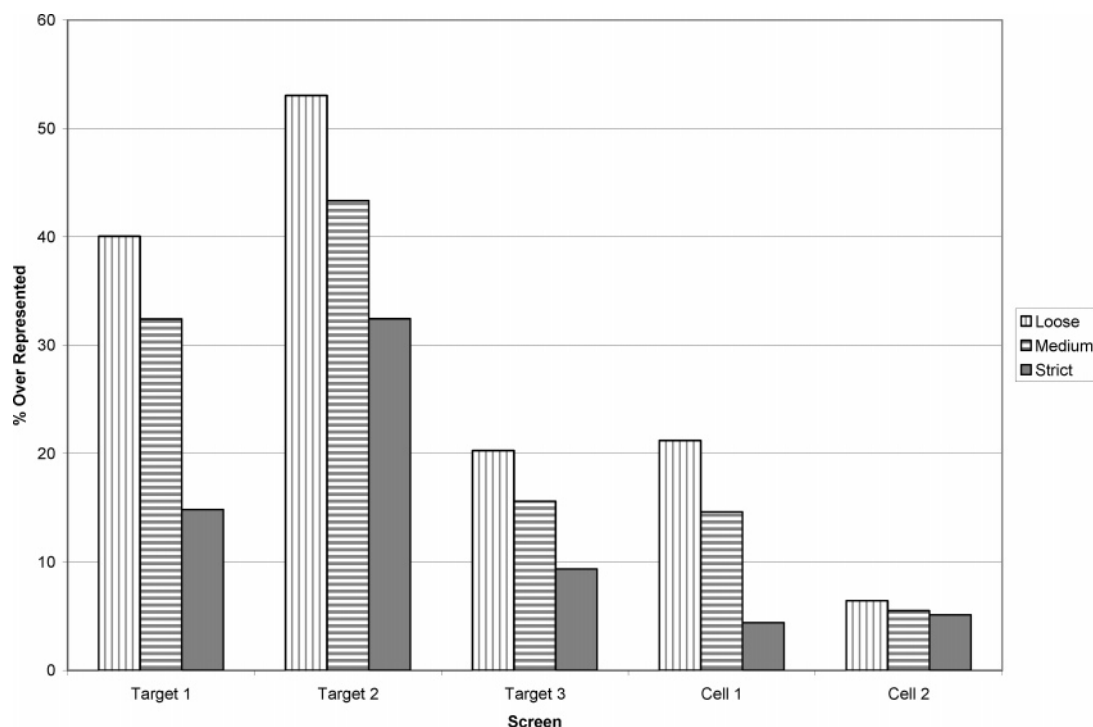


Figure 5. Percentage of compounds with confirmed activity in over-represented scaffold classes in five HTS campaigns carried out at GNF. Targets 1–3 correspond to biochemical target-based inhibition screens, and cells 1 and 2 refer to cellular antagonist screens.

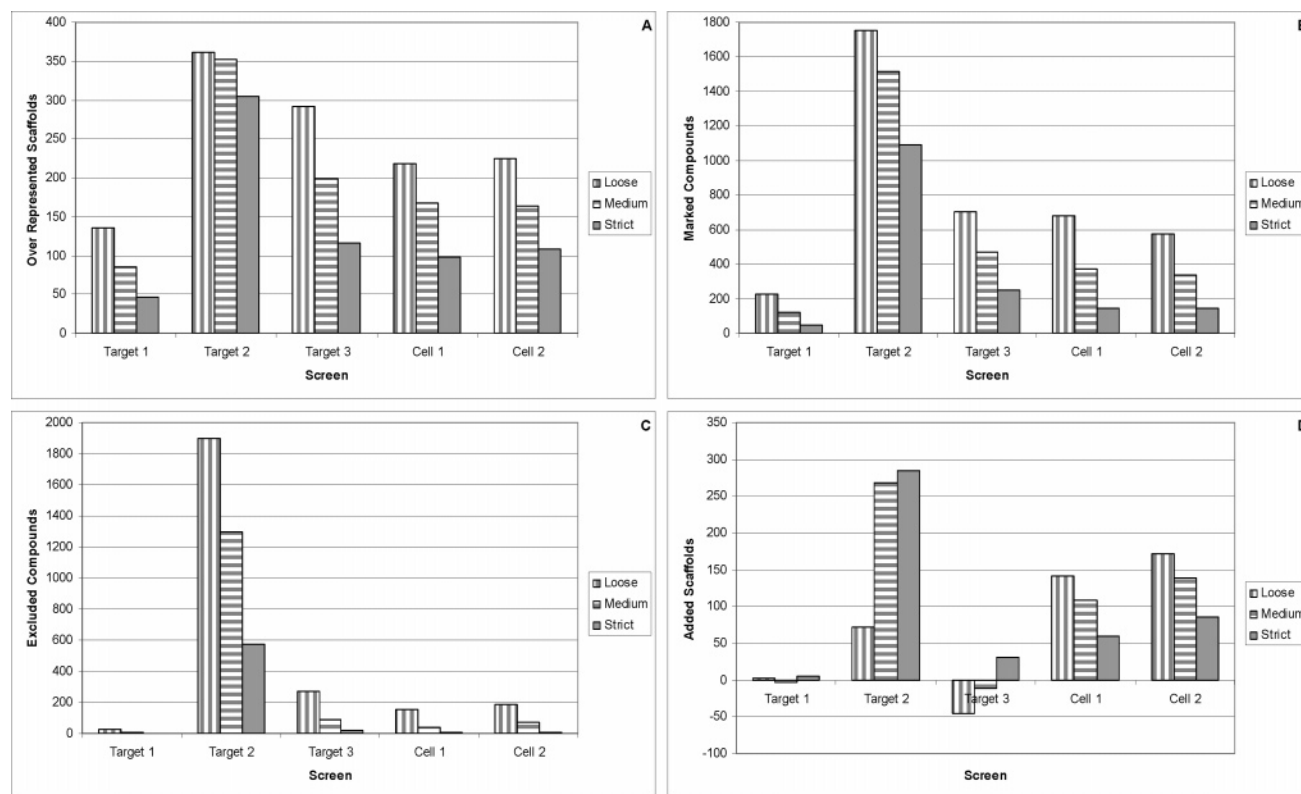



Figure 6. Automated hit pick optimization of 4000 compounds from five representative HTS campaigns. Targets 1–3 correspond to biochemical target-based inhibition screens, and cells 1 and 2 refer to cellular antagonist screens: (A) count of over-represented scaffold classes; (B) count of marked compounds in over-represented scaffold classes; (C) compounds excluded from addition to the hit pick list because they contained over-represented scaffolds; (D) the difference in the unique scaffold count after removing over-represented compounds and adding compounds from new scaffold classes.

a greater diversity of linkers than the other sets, which results in the greatest number of total unique scaffolds. Another interesting observation is that, despite having only 21 unique basis scaffolds, the CMB set still produced about 6500 unique scaffold structures from its 30 000 input compounds. In addition, the accumulation of scaffolds in the GNF set shown in Figure 4B presents an upper limit to what a set of HTS hits would produce. A successful HTS campaign would be expected to select structurally related compounds from the screening set, which would lower the number of total unique scaffolds in the set of “hits”. Previous experience has shown that a typical HTS campaign would be expected to select about two-thirds of the number of scaffolds that would be produced by a random selection of the equivalent number of compounds from the GNF screening set.

Automated Identification of Over-Represented Scaffold Classes in HTS Data. Often, compounds identified as hits in HTS campaigns contain groups of molecules that share significant structural similarity. Ideally, this is due to a common structural feature that confers activity against the particular target being screened. However, this may also be due to undesirable effects, such as aggregation²¹ or fluorescent chemical groups. Regardless, at the early stages of lead discovery, it is more beneficial to represent the maximum number of putatively active scaffolds than it is to have many compounds that represent only a few scaffold classes. This is especially important when deciding how to allocate resources for compound purification, analytical analysis, resynthesis, or reordering of larger quantities of compounds. Identifying over-represented scaffold

classes can be a tedious and error-prone process. To remedy this, HierS provides an algorithm for identifying and marking compounds in over-represented scaffold classes, as described in Methods.

Figure 5 illustrates the percentage of compounds that would be removed from actual HTS compound hit lists by the automated marking functionality in HierS using three APT thresholds for selecting over-representation. The thresholds Loose, Medium, and Strict correspond to 0.75, 0.80, and 0.85, respectively. These values were chosen by experience and by considering the work of Martin et al., who studied the relationship between Tanimoto similarity and biological similarity.²² The compound lists were taken from five representative lead discovery projects (three biochemical inhibition and two cell-based antagonist assays) and are composed of a few hundred to 1000 compounds. All five high-throughput screens were performed using almost the entire GNF screening collection, and so all have roughly the same set of compounds. The lists include all compounds with confirmed activity in their respective assay using unpurified material that had not undergone analytical verification. Ideally, purified versions of each compound that have also passed quality control tests would be obtained for further experiments. However, this goal is often difficult to achieve because of the required time and resources. It is at this stage that the automated marking functionality can help to prune and prioritize compounds in a way that maintains explicit scaffold representation. As Figure 5 shows, the percentage of compounds in over-represented scaffold classes varies significantly between screening projects. Of the five com-

HierS *hierarchical scaffold clustering* 

Compound Input **Compound Summary** **Scaffold Filter** **Scaffold Main** **Download**

Scaffold Filter

- Number of Member Compounds: Minimum: 5 Maximum:
- Number of Rotatable Bonds: Minimum: Maximum:
- Number of Rings: Minimum: 2 Maximum:
- Number of Aromatic Rings: Minimum: Maximum:
- Number of Atoms: Minimum: Maximum:
- Molecular Weight: Minimum: Maximum:
- APT: Minimum: Maximum:

Scaffolds are built

Auto Mark Options

- Mark all but 2 in over-represented scaffold classes
- Use a threshold of Loose

Figure 7. Scaffold filter page.

pond lists chosen, it appears that the lists from the biochemical screens tend to have a higher degree of over-representation than the cellular screens. This is likely due to a tendency for in-house HTS campaigns involving biochemical targets to be less noisy than campaigns involving cellular assays. Although biasing toward known active chemical features can produce a large number of hits for a given screen, the compounds that are identified as hits can be undesirable lead candidates because of intellectual property hurdles. Therefore, marking and removing compounds from over-represented scaffolds can reduce the amount of resources expended on compounds that are unlikely to be selected for development by medicinal chemists. This is especially important in cases where a medicinal chemist has not had the opportunity to analyze the data or become familiar with the patent landscape around a particular target.

Optimizing HTS Hit Lists for Diversity. The automated marking feature provided by HierS can also be used to optimize the HTS hit selection process. When HTS hits are selected, it is not uncommon to have resources for additional experiments on a fixed number of compounds, where the number of “active” compounds may be more or less than the upper limit on capacity. In other words, biologists may wish to select the top N compounds for follow-up experiments, regardless of whether N is greater than or less than the number of compounds that met the hit criteria. Typically, activity is the only criterion used to prioritize compounds for follow-up experiments. However, if HTS is viewed as a means to provide lead candidates for medicinal chemistry, scaffold diversity is an equally important criterion. To demonstrate this approach, the top 10 000 compounds from each of the HTS campaigns used above were labeled as biologically interesting hits and ranked by potency in their respective screen. For demonstration purposes, a follow-up capacity of 4000 was chosen (i.e., $N = 4000$), which means that the top 4000 compounds from each list of 10 000 were treated as the starting hit

list for each screen. To begin the process of optimizing the hit lists for diversity, the scaffolds for the top 4000 compounds in each set were computed and the over-represented compounds were marked using the three APT thresholds. In this case, at least two compounds were left unmarked to serve as representatives for the rest of the compounds in each over-represented scaffold class. The representatives were selected by first ranking the compounds in the scaffold class in descending order based on the number of neighbors within a scaffold class that have a Tanimoto similarity greater than or equal to 0.85, then selecting two compounds with the most neighbors. Next, the marked compounds were removed from the lists in each case. To fill the vacancies left by the removed compounds, HierS proceeded down the ranked lists of the remaining 6000 compounds in each set and added each new compound that did not contain any of the over-represented scaffold substructures until the hit lists reached the upper limit of 4000. Finally, the scaffolds for each new set of compounds were computed to determine the difference in the total nonredundant scaffolds between the initial and final hit lists. The results from this exercise are summarized in Figure 6.

The total number of over-represented scaffolds for each APT threshold in each data set is shown in Figure 6A. The number of over-represented scaffolds varies from about 50 to 350, where the target 1 screen has the fewest, target 2 has the most, and the remaining screens have about the same number of over-represented scaffolds. From these results, there does not appear to be a significant trend in scaffold selectivity between cell-based and biochemical screens for the particular data sets used in this study.

Figure 6B shows the number of compounds in over-represented scaffold classes that were removed from the initial 4000-compound list for each screen for each of the APT thresholds. Clearly, the target 2 screen exhibits far more over-representation than the other four screens. This result indicates that either the screen

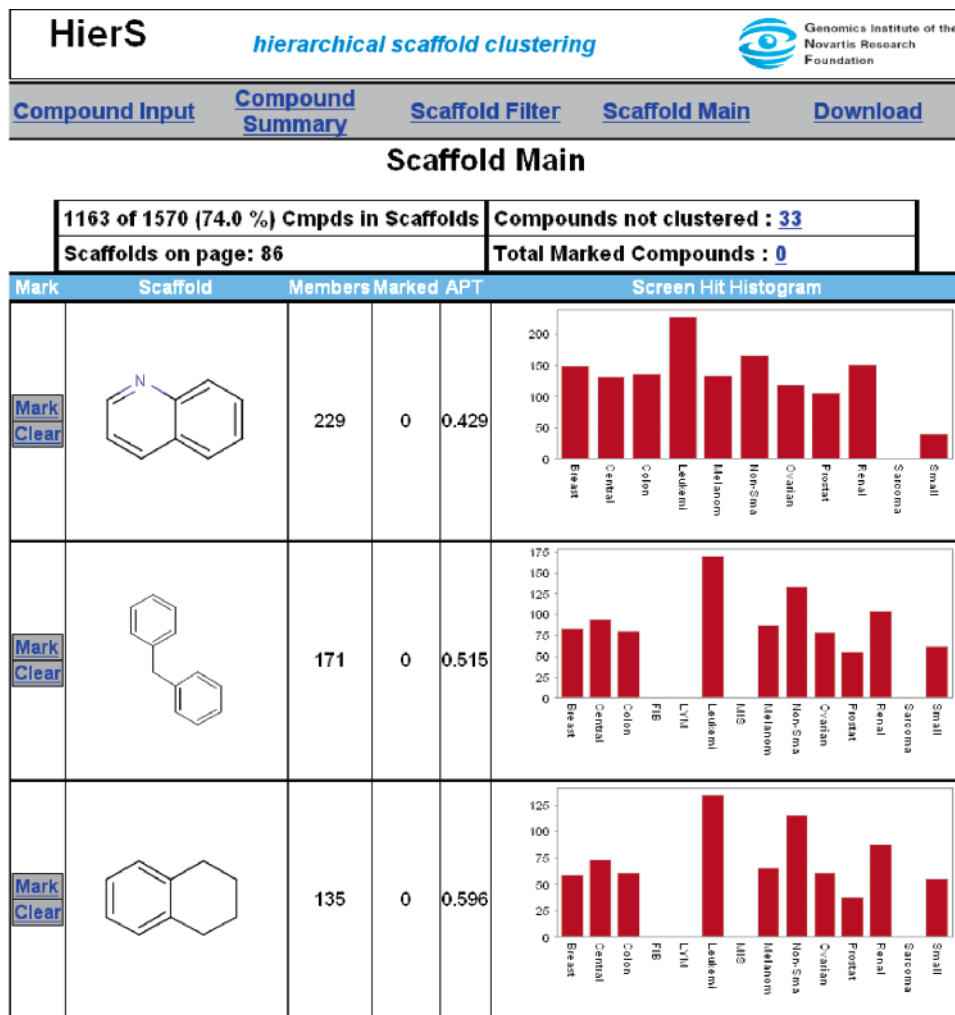


Figure 8. Scaffold main page.

is particularly selective for a few scaffold classes or there were experimental artifacts associated with compounds of a certain structure, such as reactive side products that were left over from combinatorial synthesis. Figure 6C shows the number of compounds from the remaining set of 6000 compounds that were disqualified from inclusion into the final hit lists because they also contained one or more of the over-represented scaffold structures. As with parts A and B of Figure 6, target 2 has the largest number of excluded compounds. This further indicates a strong bias toward compounds containing the scaffolds that were identified as being over-represented in the top 4000 compounds. Figure 6D shows the difference between the number of nonredundant scaffolds in the initial and final hit lists. For target 2, cell 1, and cell 2 the number of distinct scaffolds in the final set was greater than that in the initial set for all three APT thresholds. However, target 1 and target 3 displayed mixed results. In these cases HierS does not provide a significant improvement in scaffold diversity, which is not surprising given that chemical diversity of active compounds will vary from screen to screen. The ability to perform this analysis before the execution of the hit pick enables one to determine the best option before allocating resources. In this way, HierS can be used to filter and add compounds to facilitate the identification of novel scaffolds from the HTS data that might have otherwise been overlooked.

Web-Based Clustering with HierS: Case Study of DTP Cancer Data. The GI_{50} data²⁰ from the NCI compound set were used as test data for HierS analysis. These data were deemed to be the most suitable because they most closely resemble the hit criteria used by biologists at GNF. For this data set, $GI_{50} = 100$ nM in a given screen was considered a hit. This strict hit criterion was chosen because the data set consisted of many compounds that show significant potency in several cell lines. For simplicity, the DTP data were grouped by panel as defined by the NCI, where each panel can be made up of several cell lines from related tissues. The 14 panels used consisted of non-small (lung), small (lung), colon, ovarian, leukemia, renal, melanoma, central, breast, prostate, sarcoma, LYM (lymphoma), MIS (Mullerian inhibiting substance), and FIB (fibroblast) cell lines. A “hit” in one or more of a given panel’s cell lines counted as one hit for the panel.

The process of analyzing scaffolds begins with the selection of compounds. We considered the set of 1603 hits from the leukemia panel for this analysis. The next step in the process involves specifying the scaffold filter criteria, as shown in Figure 7. These filter criteria define the entry point for traversing the complex network of substructures created by the scaffold building process. The filters provided for selecting scaffolds are upper and lower bounds for the number of member compounds, rotatable bonds, total rings, aromatic rings, and atoms.

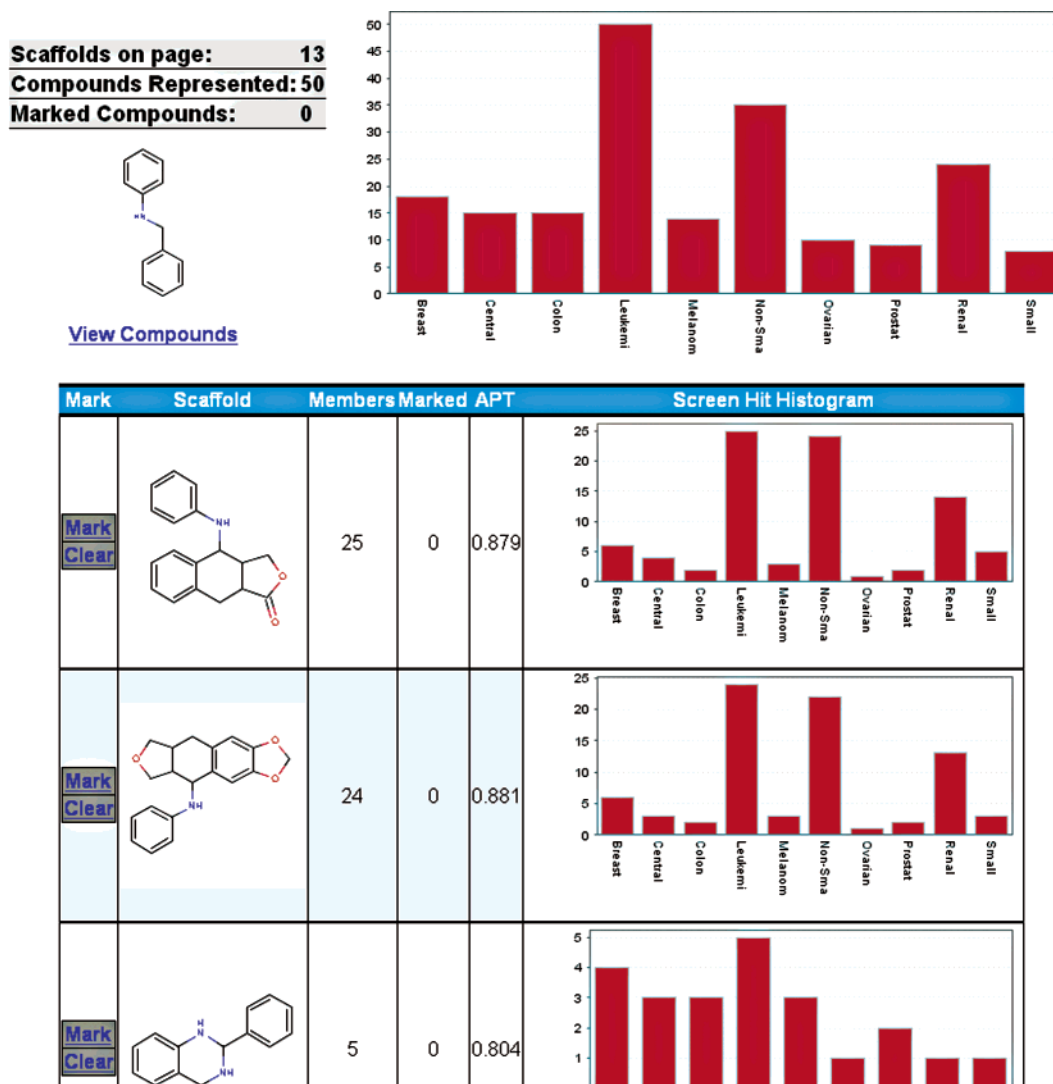


Figure 9. Scaffold drill down page. Although the scaffolds listed on the scaffold drill down page cannot be directly reduced to the *N*-benzylbenzenamine scaffold, they are still derived from the *N*-benzylbenzenamine scaffold because each scaffold contains *N*-benzylbenzenamine as a substructure.

In addition, filters specifying the maximum and minimum molecular weight and APT are also included. Unspecified filters are ignored.

Once the scaffold filter form is submitted, HierS builds the set of scaffolds for the compound set and computes their hierarchical relationships. Once the scaffold set is constructed and filtered, the list is sorted in descending order according to their member compound count. If two or more scaffolds have the same number of derived compounds, they are sorted in ascending order on the basis of the atom count in each scaffold. Following this, any scaffold that is structurally derived from another scaffold that precedes it in the list is removed. This ensures that no scaffold on a given page is a substructure of another scaffold on that page. In other words, each scaffold represents a separate branch in the hierarchical scaffold network. Finally, the scaffolds are displayed as shown in Figure 8; 1163 of 1570 compounds (74%) are represented in the scaffold classes that satisfied the filter requirements, and 33 compounds were excluded in analysis, 28 of which were rejected because they contain only one ring system, which is a benzene ring. The remaining five compounds were rejected because they contain large and flexible

single ring structures (e.g., cyclosporine), which often require an excessive amount of processing time. From left to right, the data displayed are the scaffold structure, number of member compounds, number of marked compounds, APT, and the screen hit histogram of the compounds in the scaffold class. Scaffolds are sorted by the number of member compounds so that the most common scaffolds are shown first. The screen hit histogram displays the count of the compounds in the scaffold class that were considered a hit in any prior assay. Assays where none of the compounds were considered hits are not displayed. In this way, core structural features are coupled to historical biological data to allow both chemists and biologists to rapidly assess the structural desirability and specificity of a given compound class. All of the compounds within a particular scaffold class can be “marked” by clicking the Mark link. This feature is typically used to partition undesirable compounds from the rest of the set.

As stated above, the filter parameters specify the entry point for viewing the scaffolds. Strategies for viewing different portions of the scaffold set can be developed by tuning the filter parameters. In general, the default of selecting for scaffolds with a minimum of

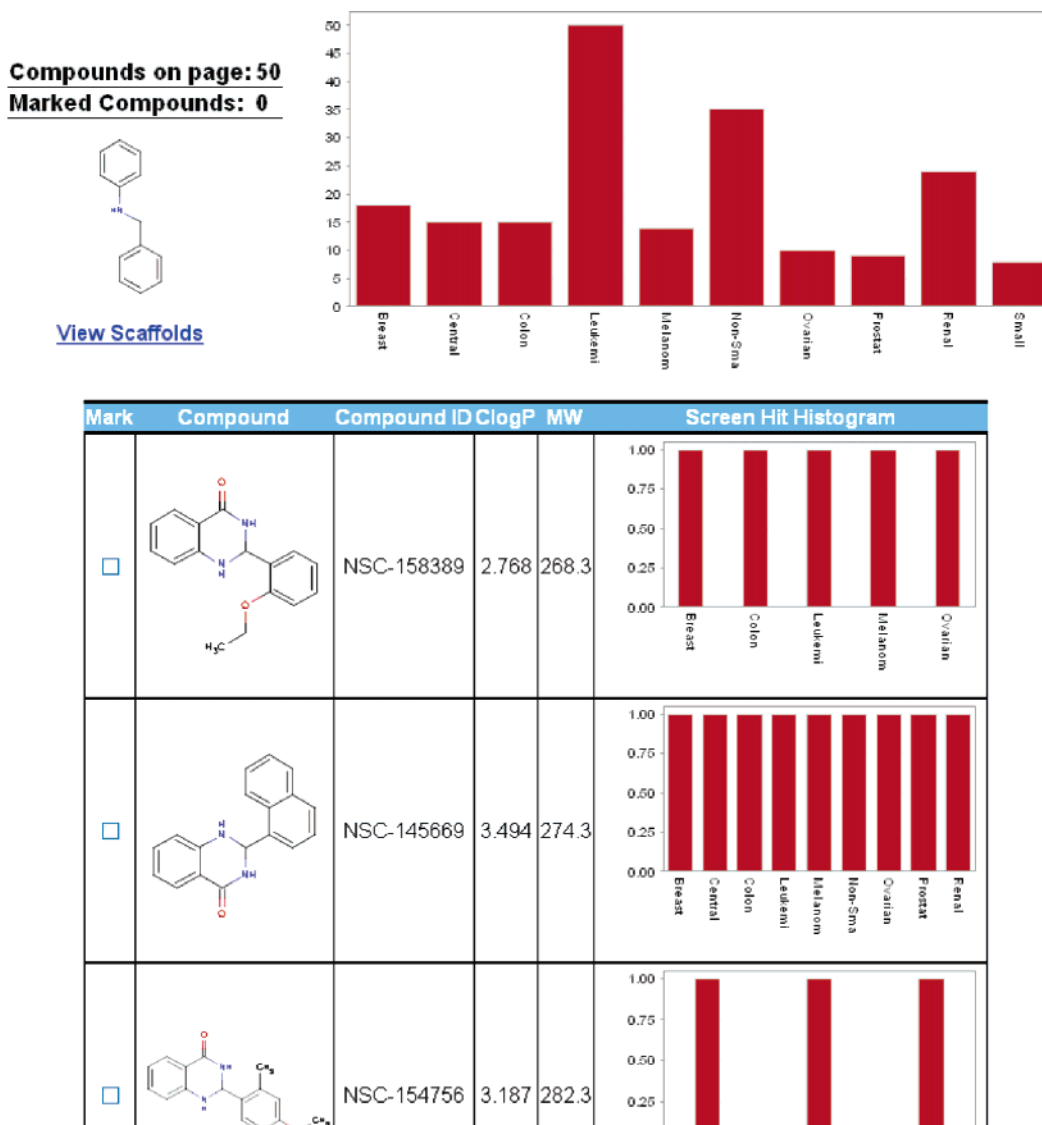


Figure 10. Compound drill down page.

five member compounds and two rings serves as a good starting point. For example, if one is interested in viewing only scaffolds that are kinase-inhibitor-like (e.g., BI-RB 796¹²), the scaffolds can be refiltered to select scaffolds that have at least three rings and at least two or three rotatable bonds. To view all the scaffolds that contain compounds that share a large degree of overall similarity, the minimum APT filter can be set to 0.80. Because the scaffolds are computed only once, several filtering operations can be performed in rapid succession.

One can drill down into a scaffold class by clicking on the scaffold image. This is equivalent to taking a single step down the scaffold hierarchy. For example, Figure 9 displays the result of clicking on the *N*-benzylbenzenamine scaffold, which is listed further down on the scaffold main page (not shown). All of the compounds that contain the *N*-benzylbenzenamine scaffold can be viewed by clicking the “View Compounds” link below the scaffold structure, as shown in Figure 10. These pages provide a larger picture of the screen hit histogram, which is followed by a list of all the scaffolds that are structurally derived from the scaffold that was selected or all the compounds in the selected scaffold,

respectively. The screen hit histograms in Figures 9 and 10 provide selectivity information for the compounds in the *N*-benzylbenzenamine scaffold class. In this example, the compounds that contain the top two scaffolds that are derived from *N*-benzylbenzenamine are selective for the leukemia, non-small lung, and renal cancer cell lines. This information can be critical for prioritizing scaffolds based on the selectivity profile, which may be indicators of potential toxicities or off-target effects.

As with the scaffold main page, scaffolds in scaffold drill down page (Figure 9) are sorted by the number of member compounds. In addition, more details on the scaffolds in the list can be obtained by clicking on a scaffold structure image. In this way, the entire structural hierarchy can be quickly navigated. Also, it is interesting to note that in Figure 9 there is stated a total of 50 compounds represented by the scaffolds on the page. In addition, there are 15 scaffolds on the page that are derived from the *N*-benzylbenzenamine scaffold. Just as with the scaffold main page, each *N*-benzylbenzenamine-based scaffold represents a distinct branch of the structural hierarchy in the scaffold network. Clicking on any *N*-benzylbenzenamine-based scaffold will cause HierS to take an additional step down

the scaffold hierarchy. This process can be repeated a number of times until the particular branch scaffold network is exhausted (which is usually between three and five clicks). Following this process, the chemical landscape in a given set of compounds can be quickly traversed in a structurally directed fashion.

After all the undesirable compounds have been marked for removal, the edited list can be downloaded and stored. This allows for collaboration between medicinal chemists and biologists during the hit selection process so that a large hit list can be reduced in size in order to remove compounds from over-represented and promiscuous scaffold classes or compounds from scaffold classes that are believed to have structural features that lead to artifacts in the data.

The automated process for identifying over-represented scaffold classes can also be controlled from the Web application interface provided in HierS. This is done in the "Auto Mark Options" dialogue on the scaffold filter page (Figure 7). Three choices of APT threshold are provided ("Loose", "Medium", and "Strict") by a dropdown box, in addition to a dropdown box for specifying the number of compounds to keep from each over-represented scaffold class. Compounds can be selected to be left marked according to their potency or by how well a given compound structurally represents the other compounds in the scaffold class (as defined by how many other compounds in the scaffold class are within a Tanimoto distance of 0.80).

When the automated process of marking compounds in over-represented scaffold classes is completed, the marked compounds can be inspected on the scaffold main page by either clicking on the number of marked compounds (see Figure 8) or navigating the hierarchy of scaffolds in the usual way. The automated marking of compounds serves as a starting point from which more compounds can be manually marked or unmarked. In this way, the user always has the final decision in what is removed and what is not. In addition, compounds can be marked or unmarked on the basis of their potency. This ensures that a user can retain the most potent compounds, regardless of their scaffold classification.

Conclusion

The topological graph-based approach for clustering that is implemented in HierS provides an efficient and straightforward mechanism for visualizing and editing the diverse compound sets. By clustering compounds by their explicit topological structure, HierS provides a readily interpretable ordering of compound data. HierS is particularly useful for analyzing hit lists produced by high-throughput screening because chemical scaffolds are coupled with historical biological data to enable a rapid correlation between scaffold features and biological activity. In addition, the ability to automark compounds in structurally over-represented scaffold classes in HierS provides a simple way to reduce the number of compounds in given set while minimizing the loss of scaffold diversity. Removing compounds by this process improves the efficiency of the follow-up process in the lead discovery phase by ensuring that all scaffold classes are represented while at the same time minimizing the

resources that are expended on compounds that are eventually found to be unsuitable for medicinal chemistry.

Because HierS groups compounds according to a single shared topological structure, it provides a simple means for extending the analysis to include further automated statistical or predictive modeling. For example, the scaffolds identified by HierS could prove to be useful in providing preliminary structure-activity relationships (SAR) within scaffold classes. Predictive models for R-group substitutions could be generated in an automated fashion using previously presented methodologies.^{23,24} In addition, comparing the number of active versus inactive compounds that contain a given (over-represented) scaffold can provide statistical insight into the enrichment of the scaffold in a HTS campaign.

References

- Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E. LeadScope: Software for exploring large sets of screening data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302-1314.
- Cross, K. P.; Myatt, G.; Yang, C.; Fligner, M. A.; Verducci, J. S.; et al. Finding discriminating structural features by reassembling common building blocks. *J. Med. Chem.* **2003**, *46*, 4770-4775.
- Tamura, S. Y.; Bacha, P. A.; Gruver, H. S.; Nutt, R. F. Data analysis of high-throughput screening results: Application of multidomain clustering to the NCI anti-HIV data set. *J. Med. Chem.* **2002**, *45*, 3082-3093.
- Miller, D. W. A chemical class-based approach to predictive model generation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 568-578.
- Craig, J. A.; Weininger, D.; Delany, J. *Daylight Theory Manual*; Daylight Chemical Systems: Mission Viejo, CA, 2004.
- Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. **2001**, *41*, 233-245.
- Takaoka, Y.; Endo, Y.; Yamanobe, S.; Kakinuma, H.; Okubo, T.; et al. Development of a method for evaluating drug-likeness and ease of synthesis using a data set in which compounds are assigned scores based on chemists' intuition. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1269-1275.
- Lewell, X. Q.; Jones, A. C.; Bruce, C. L.; Harper, G.; Jones, M. M.; et al. Drug rings database with Web interface. A tool for identifying alternative chemical rings in lead discovery programs. *J. Med. Chem.* **2003**, *46*, 3257-3274.
- Pharmaprojects*; PJB Publications Ltd.: London.
- Xu, J. A new approach to finding natural chemical structure classes. *J. Med. Chem.* **2002**, *45*, 5311-5320.
- Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887-2893.
- Pargellis, C.; Tong, L.; Churchill, L.; Cirillo, P. F.; Gilmore, T.; et al. Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. *Nat. Struct. Biol.* **2002**, *9*, 268-272.
- JChem*, version 2.2.1; ChemAxon: Budapest, Hungary.
- Daylight Chemical Systems, Inc., Mission Viejo, CA.
- Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983-996.
- Web site: <http://java.sun.com/>.
- Web site: <http://jakarta.apache.org/>.
- Web site: <http://jakarta.apache.org/struts>.
- Web site: <http://www.jfree.org/jfreechart/>.
- NCI/NIH DTP Human Tumor Cell Line Screen, 2003.
- McGovern, S. L.; Shoichet, B. K. Kinase inhibitors: Not just for kinases anymore. *J. Med. Chem.* **2003**, *46*, 1478-1483.
- Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350-4358.
- Holliday, J. D.; Jelfs, S. P.; Willett, P.; Gedeck, P. Calculation of intersubstituent similarity using R-group descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 406-411.
- Ertl, P. World Wide Web-based system for the calculation of substituent parameters and substituent similarity searches. *J. Mol. Graphics Modell.* **1998**, *16*, 11-13.