Molecular Property eXplorer: A Novel Approach to Visualizing SAR Using Tree-Maps and Heatmaps

Christopher Kibbey* and Alain Calvet

Pfizer Global Research and Development, Discovery Technologies, Michigan Laboratories,
2800 Plymouth Road, Ann Arbor, Michigan 48105

Received October 11, 2004

The tremendous increase in chemical structure and biological activity data brought about through combinatorial chemistry and high-throughput screening technologies has created the need for sophisticated graphical tools for visualizing and exploring structure−activity data. Visualization plays an important role in exploring and understanding relationships within such multidimensional data sets. Many chemoinformatics software applications apply standard clustering techniques to organize structure−activity data, but they differ significantly in their approaches to visualizing clustered data. Molecular Property eXplorer (MPX) is unique in its presentation of clustered data in the form of heatmaps and tree-maps. MPX employs agglomerative hierarchical clustering to organize data on the basis of the similarity between 2D chemical structures or similarity across a predefined profile of biological assay values. Visualization of hierarchical clusters as tree-maps and heatmaps provides simultaneous representation of cluster members along with their associated assay values. Tree-maps convey both the spatial relationship among cluster members and the value of a single property (activity) associated with each member. Heatmaps provide visualization of the cluster members across an activity profile. Unlike a tree-map, however, a heatmap does not convey the spatial relationship between cluster members. MPX seamlessly integrates tree-maps and heatmaps to represent multidimensional structure−activity data in a visually intuitive manner. In addition, MPX provides tools for clustering data on the basis of chemical structure or activity profile, displaying 2D chemical structures, and querying the data based over a specified activity range, or set of chemical structure criteria (e.g., Tanimoto similarity, substructure match, and “R-group” analysis).

INTRODUCTION

The development of software for visualizing multidimensional structure−activity data remains a significant challenge. Medicinal chemists demand that such software be intuitive, provide tools to interact with both chemical structure and biological data, and support the organization and visualization of information-rich data. A number of software tools have been developed during the past 10 years to help chemists understand structure−activity relationships present in their data. While this work is not meant to be a thorough review of the literature, a number of representative applications are described to illustrate the variety of approaches that have been applied to the analysis of multidimensional structure−activity data sets.

Navigator[1] was developed as a molecular database visualization tool in 1995. Navigator relies on a maximal common subgraph algorithm to determine neighboring relationships among chemical structures. This approach to data organization is intuitive to most chemists, because it facilitates comparison between related compounds. Two compounds are considered related if more than half their structure is identical and if one molecule can be transformed into the other by breaking a single bond and replacing the substituent at this position. MPX provides a similar means of identifying related compounds in a data set in a process called “R-group” analysis. In addition, Navigator, like MPX, provides tools for selecting and sorting data on the basis of their biological assay or molecular property values.

VisualiSAR[2] is a Web-based program that employs modal fingerprints along with Stigmata coloring of atoms to highlight common and unique structural features among compounds at various levels within a hierarchical cluster. VisualiSAR employs Daylight[3] fingerprints and Wards[4] clustering to organize chemical structures on the basis of their Tanimoto similarity. In addition, the software provides navigation tools for selecting among the various levels of the cluster hierarchy and displaying the chemical structures of cluster members. While VisualiSAR provides a useful means for visualizing chemical structures within specific clusters and cluster levels, it suffers from a common problem associated with the visual representation and navigation of hierarchical data. That is, VisualiSAR does not adequately convey the spatial relationships between clusters and cluster members among the various levels of the hierarchy.

An alternative to hierarchical agglomerative clustering is Optimizable K-Dissimilarity Selection[5] (OptiSim), which Tripos, Inc. employs in their SARNavigator[6] product. SARNavigator presents structure−activity data in a “landscape view,” wherein structurally similar compounds or clusters are plotted as circles of varying size within the central region of the landscape and dissimilar compounds

* Corresponding author phone: (734)622-5248; fax: (734)622-2782; e-mail: christopher.kibbey@pfizer.com.

10.1021/ci0496954 CCC: $30.25 © 2005 American Chemical Society
Published on Web 02/17/2005

are placed along the perimeter. The size of a circle represents the boundary of the cluster in structure space, and the circles are colored to correspond with specific activity data. While the "landscape view" is effective at providing a unifying visualization of structure and activity data, the relationship between compounds plotted in the central region of the landscape and those along the perimeter is lost.

Analysis of chemical substructure provides additional insight into structure−activity relationships. The program SLASH[7] generates a set of functional groups from an input file and then analyzes the distribution of these groups among the active compounds in the input data. LeadScope[8] keeps track of the number of compounds that possess specific functional groups, aromatics and heterocycles. Users may exclude structures from consideration by setting limits on the range of specific structure (e.g., molecular weight, logP, and number of rotatable bonds) and activity data. LeadScope relies heavily on the use of histograms, and scatter-plots, neither of which are well suited to visualizing SAR.

The challenge in presenting multidimensional data lies in the mapping of these data onto a two- or three-dimensional space. A common approach to reducing the dimensionality of a data set is nonlinear mapping, and this is usually achieved through principal-component analysis (PCA) or multidimensional scaling (MDS). An alternative to PCA and MDS is the use of Kohonen neural networks to construct Self-Organizing Maps.[9] Gedeck and Willett[10] elaborated on the application of these techniques to visualizing structure−activity relationships in high-throughput screening data in a recent review article. When applying nonlinear mapping to structure−activity data, there is always a tradeoff between choosing structure or activity as the primary means of representing the data. Organizing structure−activity data on the basis of chemical structure often interferes with the presentation of the corresponding activity data. Similarly, multidimensional activity data represented in 2D or 3D plots make it difficult for a chemist to grasp underlying correlations between chemical structure and activity. Data visualization programs, such as Spotfire's DecisionSite,[11] are effective in their approach to plotting multidimensional activity data in two or three dimensions with additional dimensionality represented by the shape, size and color of plot symbols. However, such plots are often used to segregate data, rather than correlate data. Data segregation is more effectively achieved using clustering techniques. Furthermore, DecisionSite does not provide native support for chemical structure data, and this limits its appeal as a visualization tool for structure−activity data. While a chemical structure plug-in is available for DecisionSite, the plug-in does not support substructure searching, similarity searching, or R-group analysis.

Hierarchical clusters often are represented as dendrograms, which can be difficult to navigate, especially when applied to large data sets. In 1992, Shneiderman[12] created tree-maps as an alternate visualization for hierarchical data structures containing several thousand members. A tree-map is a 2D space-filling approach in which each leaf of a tree is represented as a rectangle whose size and fill color correspond with specific attributes in the data being represented. The tree-map algorithm originally was applied to depict hierarchical computer file systems in an efficient manner. Three advantages of the tree-map algorithm are that it can render a tree of any size and depth in a predefined rectangular region, it maintains spatial relationships among items within the cluster, and it executes quickly.

Recently, the tree-map algorithm was applied to aid the visualization of microarray gene expression data mapped onto the Gene Ontology.[13] The Gene Ontology may be represented as a hierarchy; hence it is well suited to visualization of gene expression data in the form of tree-maps. One also may envision a hierarchical cluster of chemical structures depicted as a tree-map. However, in contrast to the fixed hierarchy imposed by the Gene Ontology, Molecular Property eXplorer dynamically organizes chemical structure data through agglomerative hierarchical clustering and renders the information contained in the resulting hierarchy as a tree-map. To our knowledge this work represents the first reported application of tree-maps to the visualization of chemical structure and biological activity data. In addition, we describe the tools incorporated into the software for partitioning a data set into smaller subsets on the basis of either activity or chemical structure criteria; clustering data on the basis of chemical structure or biological activity profile; and adding new structures to a data set for predictive purposes.

TREE-MAPS AND HEATMAPS

The tree-map algorithm traverses the branches of a hierarchical cluster recursively beginning with the root node. The algorithm begins with a rectangular region corresponding to the root of the hierarchical cluster. As each branch is visited, a rectangular region of the tree-map is split evenly along alternating vertical and horizontal centers. Upon reaching a terminal node in the hierarchical cluster, the corresponding rectangular region of the tree-map is associated with the terminal node of the cluster. Consequently, each rectangle of the tree-map is uniquely associated with a single node in the cluster. An example of a tree-map generated from a hierarchical cluster of 10 objects is shown in Figure 1A. The perimeter of the tree-map corresponds to the root of the hierarchical cluster shown on the left side of Figure 1A. The root node is split into a left and right branch, and this is reflected in the tree-map by dividing the rectangular region vertically into two equal halves. The left half of the tree-map corresponds to the left branch from the root of the cluster, and the right half of the tree-map corresponds to the right branch. The left branch of the cluster also possesses a left and right branch. The left half of the tree-map is split horizontally into two equal halves. The upper half represents the left sub-branch, while the lower half corresponds to the right sub-branch. Continuing with the left sub-branch in the cluster, the upper quarter of the tree-map is split vertically, and the left half of this region is assigned to node 1 in the cluster. Traversing the remaining branches of the hierarchical cluster results in the tree-map shown on the right side of Figure 1A. Each rectangular region of the tree-map is filled with a color corresponding to the value of a secondary property (e.g., biological assay) associated with that cluster node.

When created in the manner described, a tree-map provides immediate visualization of the spatial relationship between items and among subclusters of the hierarchy. In other words, a tree-map displays every subcluster in the hierarchy simultaneously. The size of a rectangle in a tree-map correlates with the depth of the corresponding node in the

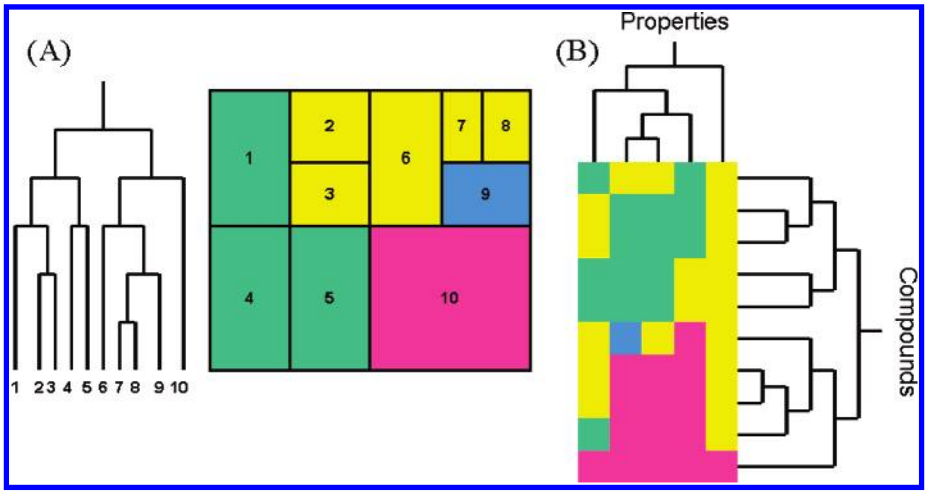


Figure 1. Representation of a hierarchical cluster of 10 compounds as (A) a tree-map and (B) a heatmap.

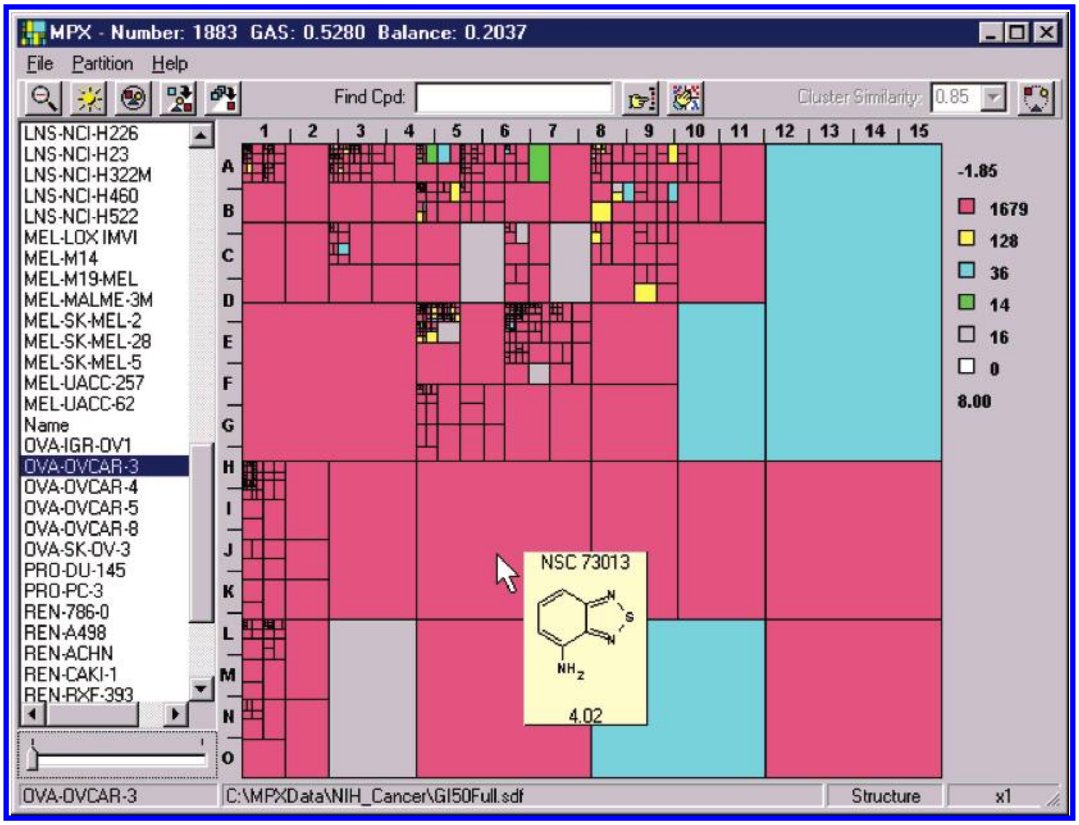
hierarchical cluster. Compounds present at the same cluster level are depicted as rectangles of equal size in the tree-map. Compounds within a subcluster are depicted by a smaller tree-map bounded by a rectangular region within the main tree-map. A clustered set of structurally related and diverse compounds would result in a tree-map characterized by regions of densely packed rectangles interspersed with more sparse regions. The sparse regions of a tree-map correspond to compounds belonging to subclusters that lie closer to the root of the hierarchical cluster.

A heatmap depicts a hierarchical cluster of items along its *y*-axis together with a hierarchical clustering of property data along its *x*-axis. An example of a heatmap representation of a cluster of 10 compounds across five properties is shown in Figure 1B. The left most terminal node in the cluster of compounds corresponds to the first row of the heatmap, and the right most terminal node corresponds to the last row. Likewise, the left most item in the cluster of properties corresponds to the first column of the heatmap, and the right most item in the cluster corresponds to the last column. At each row-column intersection, a rectangle is drawn and shaded to represent the value corresponding to that particular compound-property pair. Each rectangle of the heatmap is the same size.

Heatmaps and tree-maps provide complementary visualizations of clustered structure−activity data. A tree-map conveys both the topology of the corresponding hierarchical cluster and secondary information associated with each item of the cluster. In addition, the space-filling characteristics of the tree-map algorithm enable every subcluster within a dendrogram to be viewed simultaneously. For example, there are four subclusters at a depth of 2 from the root in the dendrogram of Figure 1A. These four clusters are represented by the four quadrants of the corresponding tree-map: the upper-left quadrant is cluster (1, 2, 3); the lower-left quadrant is cluster (4, 5); the upper-right quadrant is cluster (6, 7, 8, 9); and the lower-right quadrant is cluster (10). A tree-map is limited to depicting only one property associated with items of the cluster. In contrast, a heatmap provides a visualization of cluster nodes across multiple property values. A heatmap, however, does not depict the hierarchy that exists between nodes within the cluster. When applied to a common hierarchical cluster of data, a tree-map may be regarded as a more detailed representation of a columnar cross-section of a heatmap. The complementary relationship between heatmaps and tree-maps forms the basis on which structure−activity data is visualized in MPX.

Molecular Property eXplorer − MPX. The MPX graphical user interface consists of four major components (see Figure 2). The menu bar provides access to commands for opening a data set, modifying the graphical representation of the data, partitioning the data into smaller subsets on the basis of property or chemical structure criteria, and accessing on-line help. Below the menu bar is a tool bar that contains buttons for scaling the display region, toggling between heatmap and tree-map visualizations, clustering the data set, adding compounds to the data set, searching for compounds by name or by structure, and cycling through the display of property data over a predefined animation interval. The heatmap/tree-map display region occupies most of the application window. The tree-map visualization is annotated with a virtual map grid with letters along the left and numbers along the top edge of the tree-map. This grid provides visual reference to the absolute location of a zoomed region of the tree-map. To the right of the tree-map is a legend that defines the color assigned to each rectangle of the tree-map over a linear range for the selected property. The list of properties associated with each chemical structure is located to the left of the display region. Single or multiple properties may be selected from the list, and the software automatically updates the display region to reflect the current selection. When multiple properties are selected from the property list, the slider below this list becomes active and may be used to choose among the set of selected properties. As the mouse cursor passes over the rectangular regions of the tree-map, the name, chemical structure, and property value of the corresponding compound are displayed in a tool tip.

The tree-map algorithm tends to overemphasize singleton clusters lying close to the root of the hierarchical cluster. For example, the right half of the tree-map of Figure 2 is dominated by a number of large rectangular regions, and this reduces the amount of available space in which to represent the more populous subclusters. To compensate, MPX provides a tool for removing unwanted compounds from the cluster. The user selects one or more rectangular regions from the tree-map, and the software highlights the regions marked for removal. The software then deletes the corresponding compounds from the data set, reclusters the


MPX - Number: 1883 GAS: 0.5280 Balance: 0.2037

Cluster Similarity: 0.85

OVA-OVCAR-3 C:\MPXData\NIH_Cancer\GI50Full.sdf Structure x1

Figure 2. A tree-map of 1883 compounds from the NCI GI_{50} diversity data set. The compounds are clustered according to the similarity between their 2D chemical fingerprints. The rectangles of the tree-map are colored according to each compound's GI_{50} in the OVCAR-3 cell line. The figure illustrates the display of chemical structure and property data as the mouse pointer is passed over the tree-map.

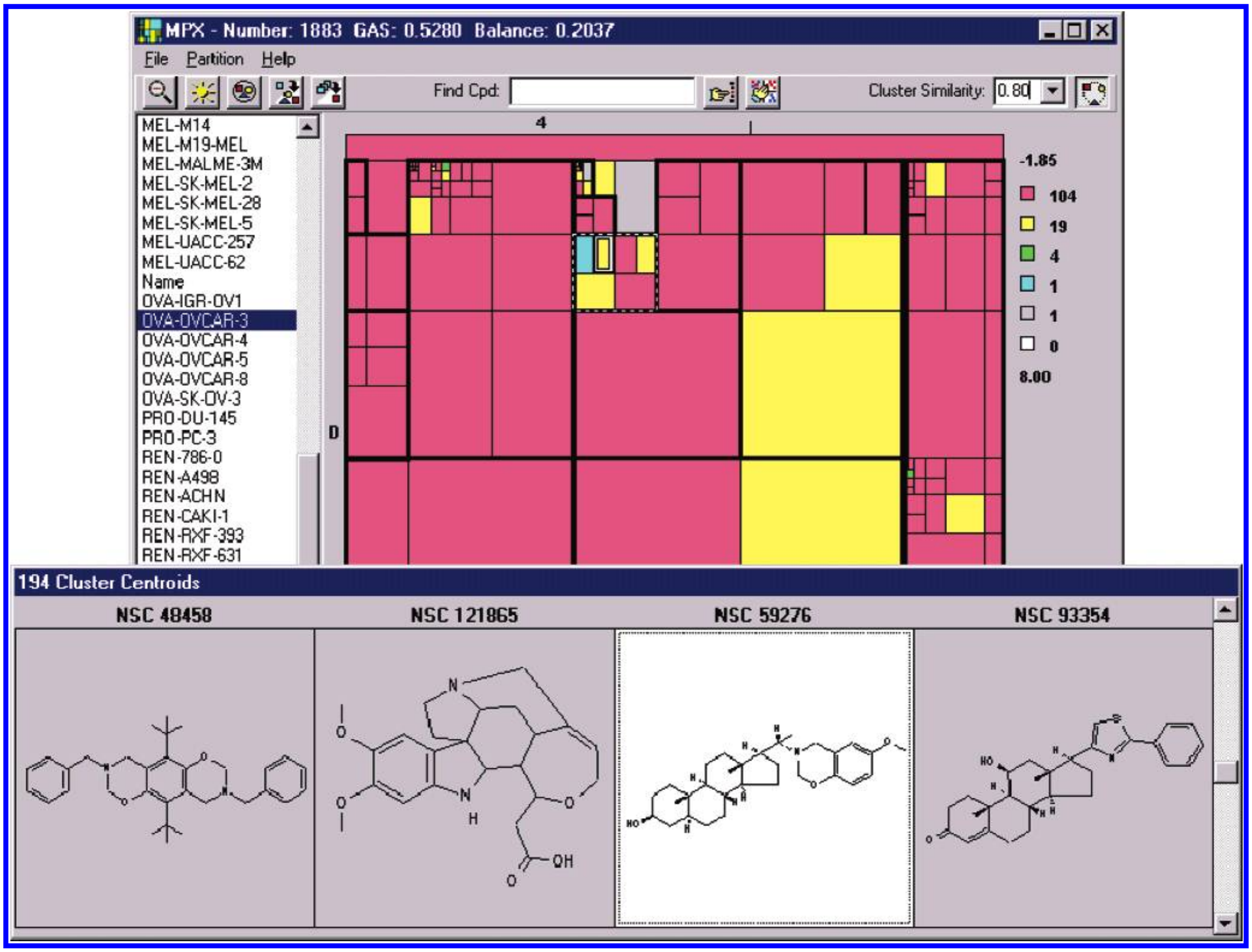
remaining compounds, and displays the result as a new tree-map. This approach allows the user to quickly identify singleton clusters and remove them as desired.

Chemical structures must first be organized into a hierarchical cluster prior to visualization as a tree-map or heatmap. MPX can cluster a data set on the basis of 2D chemical structure or a set of properties defining a profile. When clustering by chemical structure, MPX relies on the Accord Chemistry SDK[14] to generate 2D fingerprints from chemical structures. The Accord Chemistry SDK uses an approach similar to the Daylight[3] method of computing fingerprints from 2D structures. MPX uses this fingerprint data to populate a lower-triangular matrix with the Tanimoto similarity between pairs of compounds in the data set. The MPX software uses this similarity matrix to cluster compounds, compute centroids of subclusters, and compute a group average similarity (see below) for the data set.

Clustering is achieved using a reciprocal nearest neighbors (RNN) algorithm[15] and consists of two primary steps: computation of the distance between all items in the data set, followed by an agglomeration process in which subcluster hierarchies are formed. When clustering by chemical structure similarity, distances between pairs of compounds are computed as 1-Tamimoto. However, when clustering by property profile the distance between pairs of data may be computed in one of four ways: Canberra, cosine, Euclidean, and 1-Tanimoto. Five linkage algorithms are supported within MPX: single, complete, unweighted arithmetic average, weighted arithmetic average, and Ward's.

The MPX software provides a tool for highlighting significant subcluster regions of the tree-map and displaying the centroid structure within each identified subcluster (see Figure 3). The button on the far right of the toolbar activates/deactivates the highlighting of subcluster regions, and the drop-down list to the left of this button controls the threshold sensitivity for identifying significant subclusters. When the subcluster highlighting tool is activated, the software searches each subcluster of the hierarchy originating from the root. The average Tanimoto similarity for pairwise comparisons of all compounds within a subcluster (i.e., mean intercluster similarity) is compared with the similarity threshold specified in the toolbar. If the average Tanimoto similarity is greater than or equal to the specified cluster similarity, the software highlights the rectangular region in the tree-map corresponding to the subcluster. In addition, the software computes and displays the centroid structure for each highlighted subcluster of the tree-map. The centroid structures are mapped to their corresponding subcluster regions in the tree-map. Clicking on a centroid structure will outline the corresponding subcluster region of the tree-map with a dashed white line, and the rectangle corresponding to the centroid structure within this subcluster is outlined with a solid white line. The software automatically scales the tree-map in order to bring selected subcluster regions into view.

The MPX software displays three metrics in the title of the application window to aid interpretation of the corresponding hierarchical cluster. These are the number of compounds in the cluster, the group average similarity (GAS) between compounds, and the balance of the hierarchical cluster. The group average similarity is computed as the mean of the average Tanimoto similarity across rows of the similarity matrix, ignoring self-similarity. GAS is a qualitative measure of the similarity of a compound in the data set to all other compounds in the data set and ranges from 0 to



MPX - Number: 1883 GAS: 0.5280 Balance: 0.2037

Find Cpd:

Cluster Similarity: 0.80

Figure 3. Illustration of the subcluster highlighting tool in MPX. The centroid structures corresponding to four of the 194 largest subclusters possessing an average Tanimoto similarity of at least 0.80 are shown in the bottom panel. The centroid structure NSC 59276 is selected, and the corresponding subcluster region and centroid rectangle are outlined in the tree-map.

1. GAS differs from the more familiar mean intracluster similarity, the latter being the mean of N(N-1)/2 pairwise comparisons between compounds across clusters. While a rigorous treatment of GAS is beyond the scope of this paper, we find data sets consisting of diverse chemical structures produce a GAS of approximately 0.5, while more focused data sets (e.g., combinatorial libraries) yield a GAS between 0.8 and 0.85.

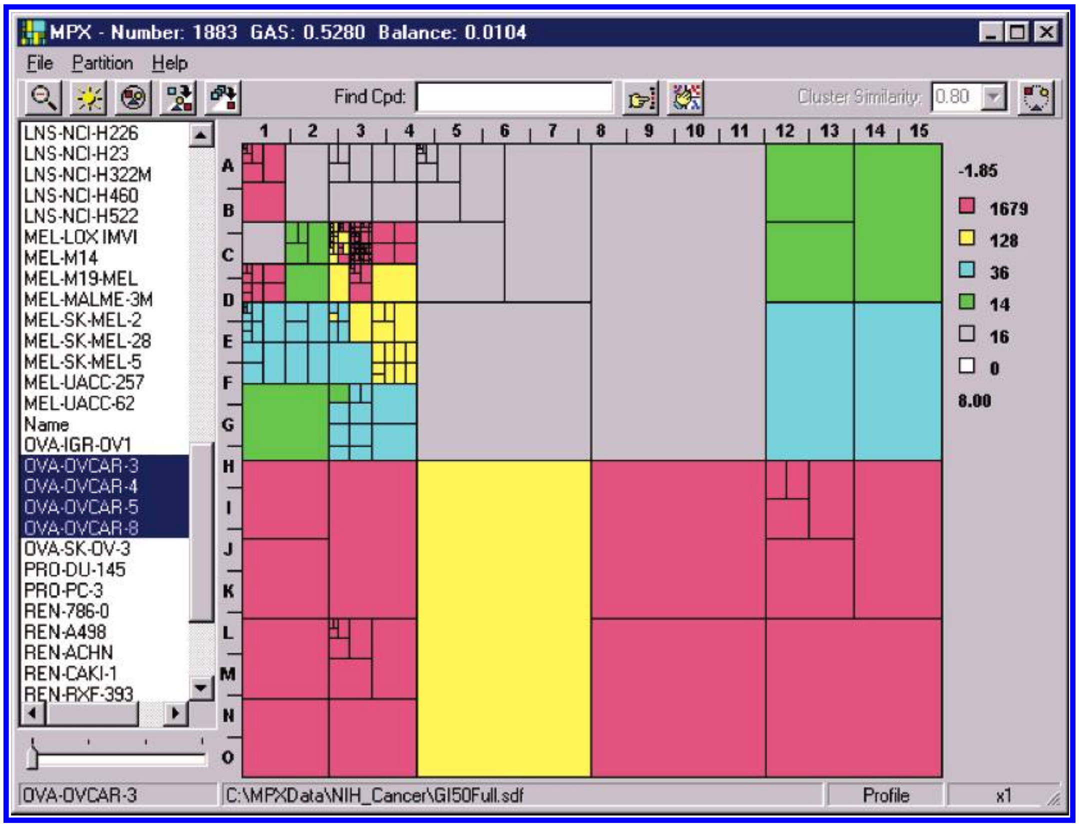
Balance is a measure of the depth of a hierarchical cluster relative to the minimum depth of a binary tree consisting of equal capacity. Equation 1 defines balance of a hierarchical cluster

$$\text{Balance} = \frac{\text{ceil}(\log_2(N))}{D} \qquad (1)$$

where ceil is the ceiling function, N is the number of compounds in the cluster, and D is the actual depth of the hierarchical cluster. The numerator in Equation 1 defines the minimum depth of a binary tree consisting of N nodes. Consequently, a balance of 1 indicates a hierarchical cluster with minimum depth, whereas a balance close to zero describes a hierarchical cluster with excessively long branches. The degree to which a hierarchical cluster is balanced also may be inferred from the prevalence of densely packed rectangular regions within the tree-map.

Each rectangular region of the heatmap/tree-map represents a single compound in the data set. Clicking and dragging with the left mouse button over a region of the heatmap or tree-map displays the corresponding 2D structures within a separate Structure Viewer window. Selected regions of the heatmap/tree-map are outlined in black. Clicking on a selected rectangle will outline it in red and highlight its corresponding structure in the Structure Viewer with a white background. A user may export selected structure−activity data in SD or tab-delimited text format from the Structure Viewer dialogue.

The MPX software presents all of the data within a data set within the display region at once. However, a user may be interested in visualizing specific subsets independent of the larger data set. The MPX software provides four means of partitioning a data set into smaller subsets for independent visualization. The data in such subsets are represented as heatmaps/tree-maps in dialogue windows separate from the main application window. A data set may be partitioned on the basis of property or 2D structure criteria. Partitioning on the basis of property criteria involves specifying discrete ranges for a set of properties. Only compounds whose property values lie within each specified range are included in the subset.



MPX - Number: 1883 GAS: 0.5280 Balance: 0.0104

Find Cpd:

Cluster Similarity: 0.80

OVA-OVCAR-3 C:\MPXData\NIH_Cancer\GI50Full.sdf Profile x1

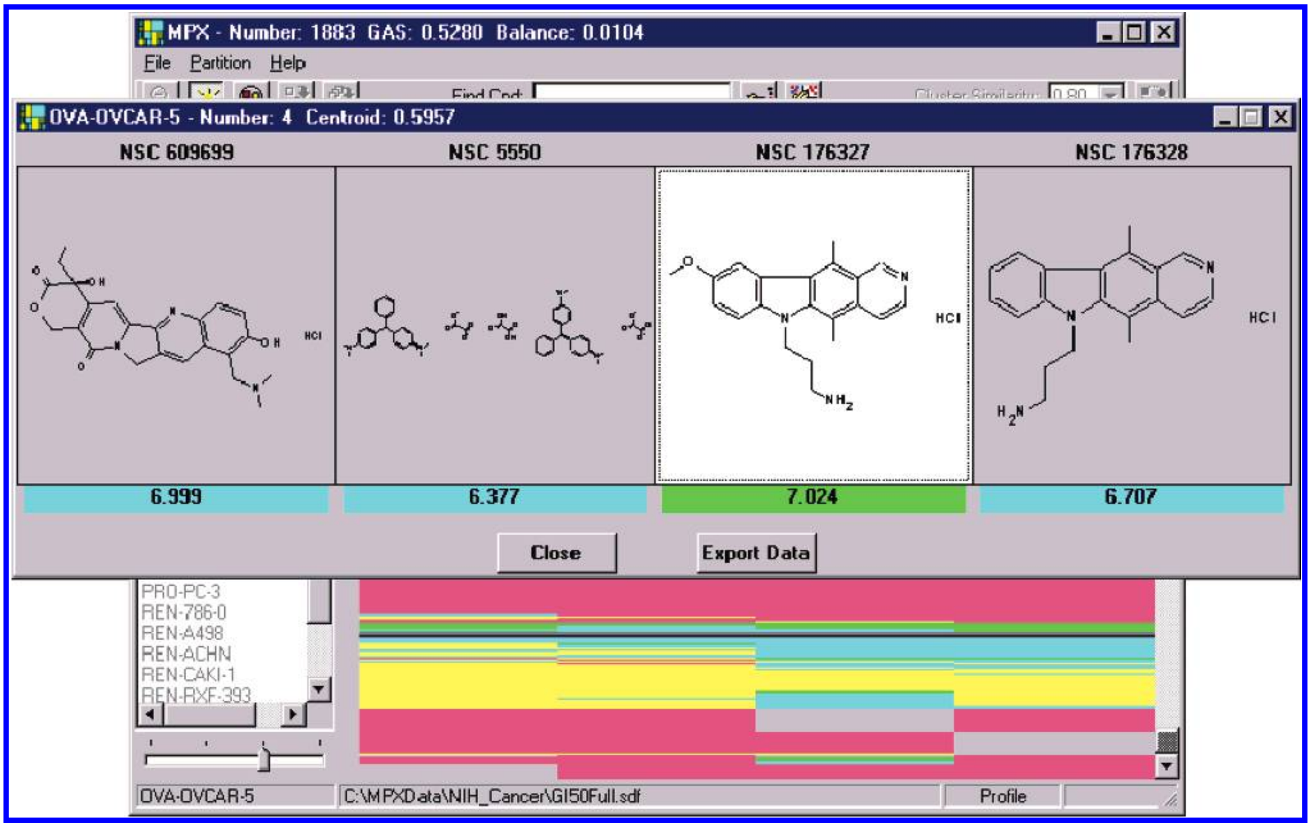
Figure 4. A tree-map of 1883 compounds from the NCI GI_{50} diversity data set clustered on the basis of their GI_{50} profile across the OVCAR-3, -4, -5, and -8 cell lines. The rectangles of the tree-map are shaded according to each compound's GI_{50} value in the OVCAR-3 cell line.

MPX offers three methods for partitioning a data set on the basis of 2D chemical structure criteria: substructure match, similarity match, and R-group analysis. The subsets generated from such partitions provide insight into the influence of specific structural features on SAR. The software partitions a data set on the basis of substructure criteria by identifying compounds in the data set that contain specified substructure(s). The user may define multiple substructure criteria and specify whether a matching compound must contain all substructures or at least one substructure. A partition based on chemical similarity identifies those compounds of the data set that meet or exceed a minimum Tanimoto similarity to a set of query compounds. Last, the MPX software can perform an R-group analysis of a set of compounds possessing a common core structure. The user draws a core substructure with designated R-groups, and the MPX software clips fragment groups at the specified R-group locations from each structure in the data set that possesses the core. These fragments are clustered and presented as separate tree-maps, one for each R-group. The rectangular regions of the R-group tree-maps are colored according to the property value of the parent compound. The R-group tree-maps may represent multiple instances of identical fragments clipped from different parent structures. Identical fragments appear in the same subcluster and are readily identified in the tree-map. The coloring of the rectangular regions corresponding to identical fragments allows one to visualize qualitative correlations between fragments and properties of the parent structures. If a fragment correlates strongly with a property, then each rectangle representing that fragment in the tree-map will be colored the same.

EXAMPLES OF USE

Two examples illustrate the use of the MPX software to visualize multidimensional structure−activity data sets. The first example employs the GI_{50} diversity set obtained from the National Cancer Institute's Developmental Therapeutics Program.[16] The assay values in this data set are reported as the negative log of the concentration of compound required to inhibit the growth of a tumor cell line by fifty percent. The second example consists of an estrogenic receptor binding data set obtained from the National Center for Toxicological Research Estrogen Receptor (NCTRER) binding database.[17]

The tree-map of Figure 2 represents a clustering by 2D chemical structure of 1883 compounds in the National Cancer Institute's GI_{50} diversity data set. The group average similarity and balance of the hierarchical cluster are shown in the title bar of the MPX application window. The group average similarity for this set of structures is 0.5280, and this value is consistent with a set of structurally diverse compounds. The balance of the hierarchical cluster of the compounds is 0.2037 and suggests the underlying hierarchical cluster possess branches of considerable depth. Indeed, this is apparent from the regions of densely packed rectangles in the upper-left quadrant of the tree-map. The property names that appear in the list to the left of the tree-map were created by concatenating the panel and cell-lines fields present in the GI_{50} data set. The selected property corresponds to the OVCAR-3 cell line from the OVA panel. The GI_{50} data for the OVA-OVCAR-3 property was truncated to the range 4.0−8.0 within the MPX software. Compounds with an OVA-OVCAR-3 GI_{50} below 4.0 were assigned the value 4.0,



MPX - Number: 1883 GAS: 0.5280 Balance: 0.0104

OVA-OVCAR-5 - Number: 4 Centroid: 0.5957

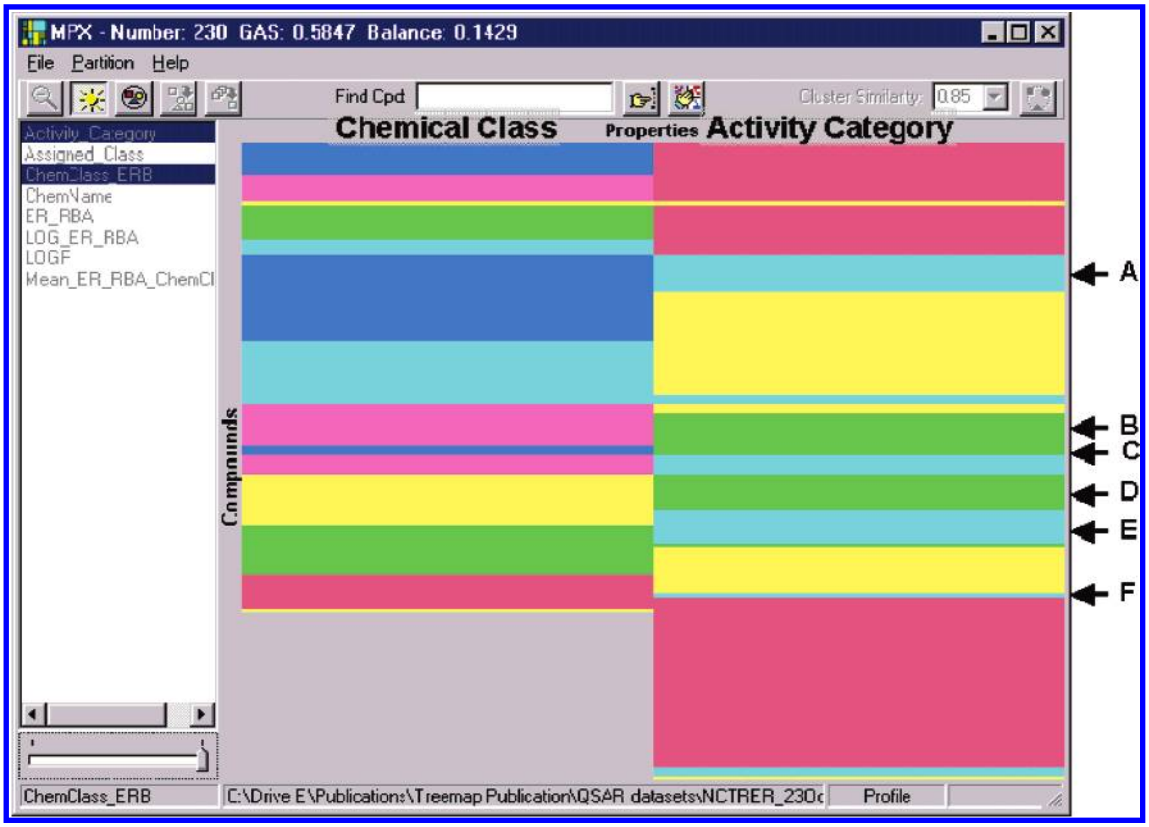
Figure 5. A heatmap representation of the tree-map of Figure 4. Compounds with high GI_{50} values across the OVCAR-3, -4, -5, and -8 cell lines are visible in the lower portion of the heatmap. Four compounds from this region of the heatmap are selected, and their structures are shown in the dialogue above the heatmap.

and compounds with a GI_{50} above 8.0 were assigned the value 8.0. The legend to the right of the tree-map indicates that 1689 of the 1883 compounds have a GI_{50} between 4 and 5, 128 compounds have a GI_{50} between 5 and 6, 36 compounds have a GI_{50} between 6 and 7, and 14 compounds have a GI_{50} between 7 and 8. In addition, there are 16 compounds with unreported GI_{50}. Compound identification, structure and property value are displayed in a tool tip as the mouse passes over the rectangular regions within the tree-map.

The data set consists of a variety of diverse compound classes, clustered by chemical structure. Consequently, there are multiple regions of SAR scattered throughout the tree-map of Figure 2. For example, compounds belonging to the camptothecin and ellipticine classes, both of which are known potent inhibitors of tumor growth,[18,19] are located within grid A-6 of the tree-map. Suppose one was interested in visualizing all compounds active against the OVCAR-3, OVCAR-4, OVCAR-5, and OVCAR-8 cell lines independent of their chemical class. Such visualization is achieved in MPX by clustering the data on the basis of a property profile as illustrated in Figure 4. The OVCAR-3, OVCAR-4, OVCAR-5, and OVCAR-8 cell lines were selected from the property list, and the data was reclustered (employing Euclidean distance and complete linkage) using the tool bar's cluster button. Two distinct regions of densely packed rectangles characterize the tree-map of the reclustered data. These regions correspond to compounds that are either potent or impotent inhibitors of tumor cell growth across the selected cell lines. Compounds with intermediate profiles separate these two regions. The most potent inhibitors of tumor cell growth are located within grids E−1, E−2 and G-3 of the tree-map.

A tree-map can represent only one property at a time. A heatmap allows visualization of multiple properties simultaneously. A heatmap of the GI50 data set clustered by profile across the OVCAR-3, OVCAR-4, OVCAR-5, and OVCAR-8 cell lines is shown in Figure 5. The potent inhibitors within grids E−1 and E−2 of the tree-map of Figure 4 are represented in the lower region of the heatmap. Four compounds within this region have been selected and their structures are shown in Figure 5. Note that compound NSC 5550 contains multiple fragments and salts. MPX does not remove salts or fragment structures from compounds in the data set. Removal of such items must be performed prior to loading a data set into MPX. The compounds within the selected region are structurally diverse as indicated by a mean (centroid) similarity of 0.5957 computed for the four compounds. The black line in the lower third of the heatmap in Figure 5 corresponds to compound NSC 176327 highlighted in the structure panel. Were one to toggle the display back to tree-map, the rectangle representing this compound in the tree-map also would be highlighted. The selected compound's nearest neighbors in the cluster hierarchy are easily identified from the tree-map.

Many commercial software applications that provide support for heatmaps do so only in the context of two-way clustering. That is, the criteria used to cluster across rows of the heatmap must be used to cluster across the columns. Heatmaps created within MPX allow two different sets of criteria to be used to cluster across row and columns. The advantages gained from treating rows and columns of the



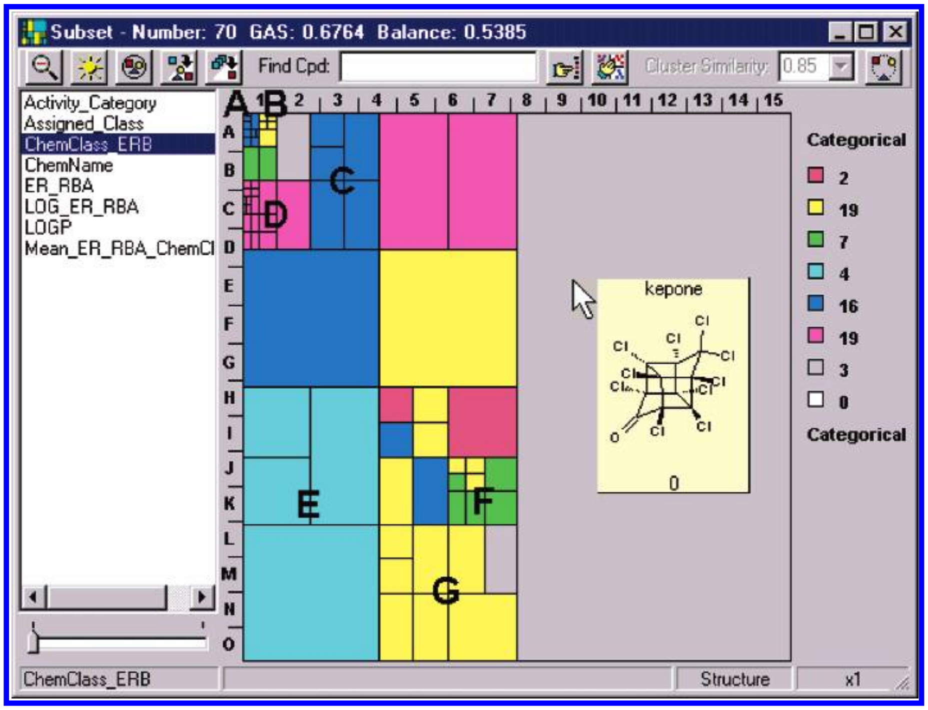
MPX - Number: 230 GAS: 0.5847 Balance: 0.1429

Find Cpd:

Cluster Similarity: 0.85

ChemClass_ERB | C:\Drive E\Publications\Treemap Publication\QSAR datasets\NCTRER_230 | Profile

Figure 6. A heatmap of the NCTRER estrogen receptor binding data set. The compounds in the data set are clustered by activity category and chemical class. A relationship between structure and activity is apparent from the overlap of activity category and assigned chemical class for the active estrogen receptor binding compounds in the data set: (A) Phytoestrogens-Flavones/Isoflavones, (B) Steroids with aromatic A-ring, (C) Phytoestrogens-Mycoestrogens, (D) DES Triphenylethylenes, (E) Diphenylmethanes, and (F) Biphenyls PCBs.



Subset - Number: 70 GAS: 0.6764 Balance: 0.5385

Find Cpd:

Cluster Similarity: 0.85

ChemClass_ERB | Structure | x1

Figure 7. A tree-map of the most active estrogen receptor binding compounds in the NCTRER data set. Due to their clustering on the basis of structural similarity, compounds belonging to the same chemical class lie adjacent to one another in the tree-map. The regions labeled correspond to the following: (A) Phytoestrogens-Flavones/Isoflavones, (B) DES Triphenylethylenes, (C) Phytoestrogens-Mycoestrogens, (D) Steroids with aromatic A-ring, (E) Phenols, (F) Diphenylmethanes-DDTs, and (G) DES Hexestrol derivatives. The gray region in the right half of the tree-map corresponds to kepone, whose structure is most dissimilar to those compounds depicted in the left half of the tree-map.

heatmap as independent clusters are illustrated in the analysis of the NCTRER data set that follows.

The NCTRER data set consists of 230 compounds representing a variety of chemical classes. The data set

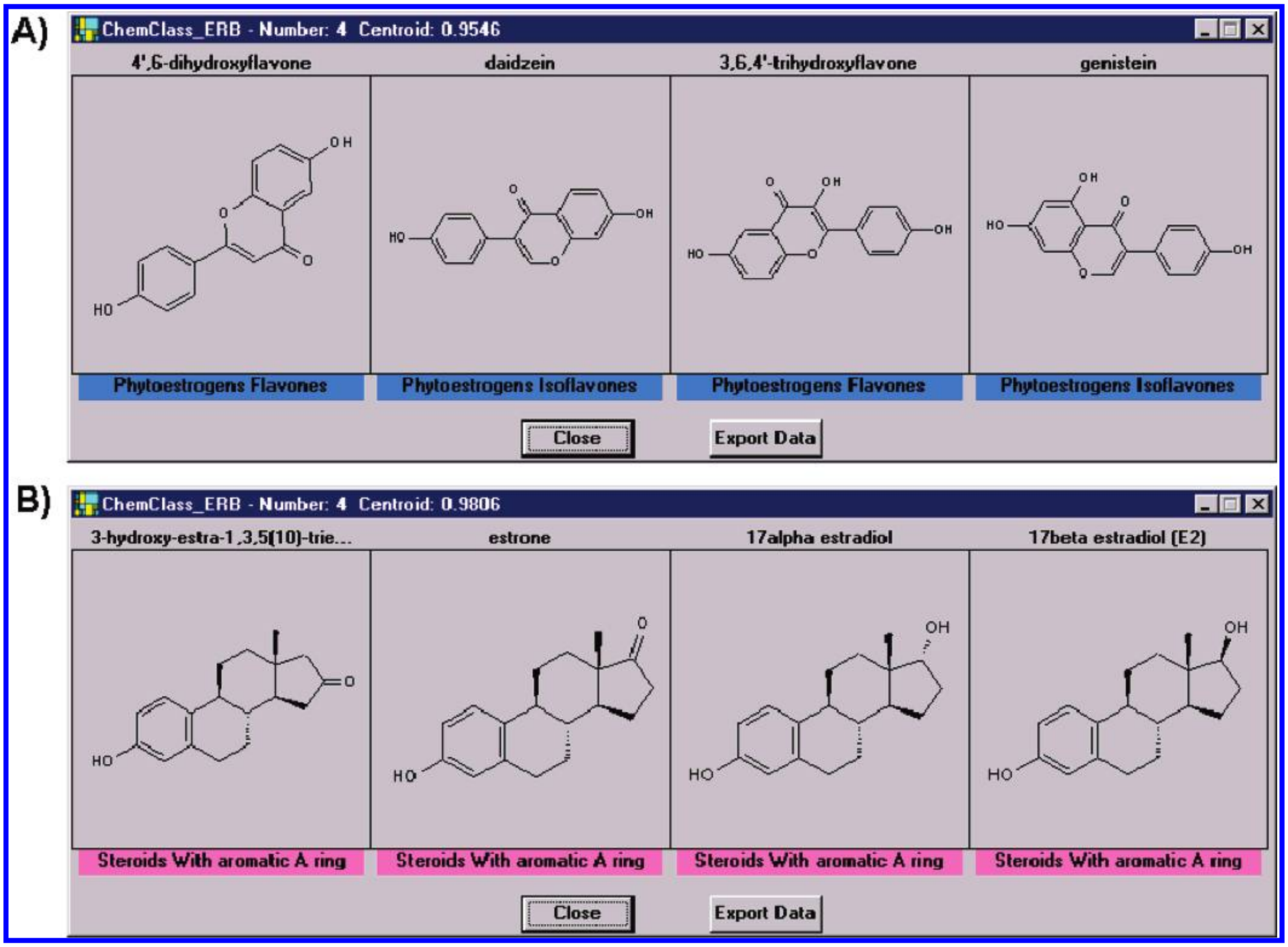

A)
ChemClass_ERB - Number: 4 Centroid: 0.9546
4',6-dihydroxyflavone
daidzein
3,6,4'-trihydroxyflavone
genistein
Phytoestrogens Flavones
Phytoestrogens Isoflavones
Phytoestrogens Flavones
Phytoestrogens Isoflavones
Close
Export Data
B)
ChemClass_ERB - Number: 4 Centroid: 0.9806
3-hydroxy-estra-1,3,5(10)-trie...
estrone
17alpha estradiol
17beta estradiol (E2)
Steroids With aromatic A ring
Steroids With aromatic A ring
Steroids With aromatic A ring
Steroids With aromatic A ring
Close
Export Data

Figure 8. Structures of genistein and three related phytoestrogen flavones/isoflavones (A) and four steroids possessing aromatic A-rings (B) selected from the upper left region of the tree-map in Figure 7.

includes the properties: Activity Category ER_RBA and ChemClass ERB. Activity Category ER_RBA classifies the estrogen receptor binding strength of each compound as one of the following: inactive, slight binder, active weak, active medium, and active strong. ChemClass ERB assigns each compound to one of six broad chemical classes: miscellaneous, biphenyls, diethylstilbestrol (DES), diphenylmethanes, phenols, phytoestrogens, and steroids. Subtypes are used to further define compounds within these classes. For example, phytoestrogen compounds fall into one of the following subclasses: flavones, isoflavones, and mycoestrogens. The MPX software is compatible with both continuous numeric and categorical text data. Hence, the text values associated with Activity Category ER_RBA and ChemClass ERB properties did not have to be numerically encoded prior to analysis.

The compounds in the NCTRER data set were clustered by a profile consisting of Activity Category and ChemClass ERB employing euclidean distance and complete linkage. The heatmap of the clustered compounds and the selected properties (Activity Category ER_RBA and ChemClass ERB) is shown in Figure 6. The group average similarity for these compounds is 0.5847, and the balance of the hierarchical cluster is 0.1429. Compounds assigned to the “inactive” Activity Category are colored orange, “active weak” compounds are colored yellow, “active medium” compounds are colored blue, and “active strong” compounds are colored green. Likewise, the assignment of compounds within ChemClass ERB are represented in the heatmap by color-coded rectangles. Relationships between chemical structure and estrogen binding receptor activity are readily apparent from the heatmap of Figure 6. The compounds classified as having “active medium” and “active-strong” estrogen receptor binding affinity are identified on the right-hand side of the heatmap. The chemical class assignments for the most active compounds are depicted in the left column of the heatmap. Together, the chemical class and activity category columns of the heatmap allow one to readily identify the most active compounds (by chemical class) present in the data set.

The NCTRER data set was partitioned into a subset on the basis of Activity Category being “active medium” or “active strong”. The compounds in this subset were clustered on the basis of the similarity of their 2D chemical structures. The compounds in the subset clustered primarily by chemical class as is evident from the tree-map of Figure 7. ChemClass ERB assignment is used to shade each rectangle of the tree-map, and regions occupied by compounds belonging to the various chemical classes have been added to the figure. The large rectangle occupying the right half of the tree-map corresponds to the compound kepone, an unusual estrogen receptor binder assigned to the “miscellaneous” chemical class and is structurally dissimilar to every other compound in the subset.

The structures of four phytoestrogen isoflavones from region A-1 and four steroids from region C-1 of the tree-map in Figure 7 are shown in Figure 8A and 8B, respectively. The interpretation of the SAR within the two sets of structures is straightforward. Isoflavones become potent binders to estrogenic receptors when hydroxyl groups in the 7 and 4′ positions mimic 4, 4′ OH positions in diethylstilbestrol, as illustrated by genistein. Steroids possessing 3-hydroxy substituted, phenolic, A-rings bind to estrogen receptors, and the strength of this binding is increased when an oxygen atom is present at the 17-position.

DISCUSSION

The MPX software offers a number of potential advantages beyond those already described. The software may be used qualitatively to predict the properties of new compounds. A button on the tool bar provides mechanism for adding new chemical structures to a data set. As new compounds are added, the data set is reclustered and the tree-map is redrawn to reflect the placement of the new compounds within the cluster. Assuming a sufficient number of compounds defining an SAR exits within the original data set, the activity of a new compound may be inferred from that of its nearest neighbors within the cluster. Assessing the structural similarity of the compound in question with its nearest neighbors may also substantiate the validity of such qualitative comparisons. If a new compound clusters within a dense region of the tree-map, and this region has well defined SAR and the structure of the new compound compares favorably with its nearest neighbors, then the activity of the new compound may be inferred from the activity of its nearest neighbors.

The combination of hierarchical clustering based on 2D chemical structure and visualization as a tree-map is a novel approach to representing the topology of the chemical space for a set of chemical structures. Properties other than those relating to biological activity may be mapped onto this topology. For example, the date on which a compound was synthesized and the name of the corresponding therapeutic program could be incorporated into the data set. Such information would allow one to visualize the discovery process within and across therapeutic projects. Shading the rectangles of the tree-map by date provides a historic representation of the various medicinal chemistry strategies applied within a project. A tree-map encoding the name of the therapeutic project for which a compound was synthesized might be used to identify compounds applicable to other therapeutic programs.

CONCLUSION

Development of software capable of representing multidimensional structure−activity data in a straightforward and intuitive manner remains a challenge. Representation of clustered data as heatmaps and tree-maps is a novel means of visualizing SAR. We have combined these two powerful visualizations with a set of data-mining tools into a software application well suited to exploring and understanding multidimensional, structure−activity data sets. The MPX software may be used to identify regions of structural similarity and dissimilarity within a data of compounds, segregate compounds into distinct regions on the basis of a defined activity profile, and visualize relationships between structure and activity. The MPX software is best applied to data sets of up to ∼10,000 compounds. Larger data sets will take significantly longer to cluster and may not be well represented as heatmaps and tree-maps. Feedback from the medicinal and computational chemists within Pfizer suggests MPX is a convenient tool for visualizing and exploring qualitative SAR within multidimensional data sets.

REFERENCES AND NOTES

(1) Chapman, D.; Harris, N.; Park, J.; Critchlow, R. E., Jr. Navigator: Tools for informal structure−activity relationship discovery. *J. Mol. Graphics* **1995**, *13*, 242−249.

(2) Wild, D. J.; Blankley, C. J. VisualiSAR: A Web-based application for clustering, structure browsing, and structure−activity relationship study. *J. Mol. Graphics* **1999**, *17*, 85−89.

(3) The guide to Daylight theory is available online: *http://www.daylight.com/dayhtml/doc/theory/theory.finger.html*.

(4) Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236−244.

(5) Clark, R. D. OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181−1188.

(6) SARNavigator is available from Tripos, Inc. *http://www.tripos.com/*.

(7) Cosgrove, D. A.; Willett, P. SLASH: A program for analyzing the functional groups in molecules. *J. Mol. Graphics* **1998**, *16*, 19−32.

(8) 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.

(9) Kohonen, T. In *Self-Organizing Maps*, 2nd Edition, Springer-Verlag: Berlin, 1997.

(10) Gedeck, P.; Willett, P. Visual and computational analysis of structure−activity relationships in high-throughput screening data. *Curr. Opin. Chem. Bio.* **2001**, *5(4)*, 389−395.

(11) Spotfire DecisionSite is available from Spotfire, Inc. *http://www.spotfire.com/*.

(12) Shneiderman, B. Tree visualization with Tree-maps: a 2-d space-filling approach. *ACM Trans. Graphics* **1992**, *11*, 92−99. A history of tree-map research is available online: *http://www.cs.umd.edu/hcil/treemap-history/index.shtml*.

(13) Baehrecke, E. H.; Dang, N.; Babaria, K.; Shneiderman, B. Visualization and analysis of microarray and gene ontology data with treemaps. *BMC Bioinformatics* **2004**, *5(1)*, 84−96.

(14) The Accord SDK is available from Accelrys, Inc. *http://www.accelrys.com/*.

(15) Murtagh, F. A Survey of Recent Advances in Hierarchical Clustering Algorithms. *Comput. J.* **1983**, *26(4)*, 354−359.

(16) Screening data and 2D structures of compounds in the NCI GI50 diversity data set and available online: *http://dtp.nci.nih.gov/webdata.html.*

(17) Structures and estrogenic receptor binding data are available online: *http://www.epa.gov/nheerl/dsstox/sdf_nctrer.html.*

(18) Ohashi, M.; Oki, T. Ellipticine and related anticancer agents. *Expert Opin. Ther. Pat.* **1996**, *6*, 1285−1294.

(19) Shi, L. M.; Myers, T. G.; Fan, Y.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. Mining the National Cancer Institute Anticancer Drug Discovery Database: Cluster Analysis of Ellipticine Analogues with p53-Inverse and Central Nervous System-Selective Patterns of Activity. *Mol. Pharm.* **1998**, *53*, 241−251.

CI0496954