ELSEVIER

VisualiSAR: A Web-based application for clustering, structure browsing, and structure–activity relationship study

David J. Wild and C. John Blankley

Parke-Davis Pharmaceutical Research Division, Warner-Lambert Company, Ann Arbor, Michigan, USA

VisualiSAR is a program designed to display chemical structures, find similarities and differences between them, and highlight relationships that might exist. The program integrates cluster analysis for the grouping of structurally related compounds, modal analysis of molecular fingerprints for the sorting and highlighting of chemical features, and a Web-based interface for flexibility and ease of use. VisualiSAR has proved useful for a number of applications including the discernment of structure–activity relationships (SAR) of high-volume screening data, and general structure browsing. This article discusses the design of the tool and illustrates two applications. © 2000 by Elsevier Science Inc.

Keywords: World Wide Web, SAR, molecular similarity, cluster analysis, modal fingerprints

INTRODUCTION

VisualiSAR is a program that has been developed in response to a need for a general-purpose structure browsing and SAR analysis tool that is accessible to a wider range of scientists than just specialist computational chemists. It has been designed for application in a number of areas of chemistry, but especially in the structure–activity analysis of biological screening results, particularly those derived from our high-volume screening protocols. VisualiSAR brings together a number of established technologies, including cluster analysis for the grouping of structures, fingerprinting and modal analysis for the scoring and sorting of compounds, and a method known as Stigmata for using colors to highlight areas of commonality and difference on the 2D structure diagrams of compounds.

Previous publications have demonstrated the effectiveness of modal analysis and the Stigmata coloring scheme for highlighting commonality, diversity assessment, and database searching.[1,2] In this article, we show how these technologies can be used for visually identifying structure–activity relationships, and how in combination with cluster analysis they are appropriate for use with data sets containing structures from different structural series. By bringing the technologies together in a simple-to-use application, and by projecting SAR information onto the structures themselves (as opposed to presenting equations or statistical plots), we have been able to facilitate SAR analysis for a wider range of chemists than was previously possible. We describe here the technologies of VisualiSAR, make some comments about the design of the application, and give two examples of how this kind of tool can be used on pharmaceutical data sets.

Color Plates for this article are on pages 120–125.

Corresponding author: John Blankley, Parke-Davis Pharmaceutical Research Division, Warner-Lambert Company, 2800 Plymouth Road, Ann Arbor, MI 48105, USA.

VISUALISAR TECHNOLOGIES

Much of the processing in VisualiSAR is based on the manipulation of molecular *fingerprints.* A fingerprint is a collection of descriptors that represent information (usually structural) about a single molecule. In the simplest case, a fingerprint can be a string of binary digits (bits), where each bit represents the presence (1) or absence (0) of a particular structural feature (this simplest case is often referred to as a set of *structural keys*). In a more complex case, a fingerprint may represent many more features than there are bits in the fingerprint by hashing information onto a fixed set of bits, or may represent frequency information using nonbinary descriptors. Currently, VisualiSAR uses Daylight fingerprints,[3] which are of the hashed variety, although its modular design allows easy incorporation of other binary fingerprint types. Specifically, the fingerprints we used are 2048 bits long, and encode information about the atom types and paths of length 1–7 bonds present in compounds.

An extension of the fingerprint concept used in VisualiSAR is the *modal fingerprint.*[1] A modal fingerprint may be generated from the fingerprints for a collection of compounds. A given bit position is set if a given percentage (known as the *modal threshold*) of the fingerprints in the collection have that bit set. The modal fingerprint of a set of compounds thus

Journal of Molecular Graphics and Modelling 17, 85–89, 1999

655 Avenue of the Americas, New York, NY 10010

1093-3263/99/$–see front matter
PII S1093-3263(99)00026-1

represents the structural features that are present in at least a given fraction of the compounds. It is then possible to calculate a whole range of measures relating individual compound fingerprints to the modal fingerprint of the collection of compounds. For example, *MSIM* is a measure of similarity between the fingerprint of an individual compound and a modal fingerprint (using the *Tanimoto coefficient*[4] of similarity). An MSIM of 1.0 means that the compound's fingerprint is identical to the modal, and an MSIM of 0.0 means it has nothing in common with it. High values of MSIM indicate that a compound is fairly representative of the set of compounds from which the modal is derived. *MODP* is the fraction of bits set in the modal that are also set in the individual compound's fingerprint. A MODP of 1.0 means that the compound's structure contains all of the fragments represented by the modal, while a MODP of 0.0 means that the structure contains none of what is common to the set. In VisualiSAR, these measures are used for sorting compounds in a set or within a cluster. If MSIM is used as the primary sort key, then the compounds near the top of the ranking are those most representative of the set or cluster. If MODP is used, then the compounds ranked highest are those that contain most of what is common to the set.

Since the modal fingerprint need not represent a real structure, there is no obvious way of depicting the structural information present in it. The Stigmata[1] method enables coloring of the atoms in chemical structure depictions based on their frequency of occurrence in a modal fingerprint. Atoms that are not part of any of the structural features represented in a modal are colored red, while atoms that are part of all of the structural features present in the modal are colored white. In between is a 10-scale or 4-scale color gradation based on temperature, from red, through orange, yellow, green, and blue to white. Stigmata thus provides an intuitive method of highlighting common and unusual structural features in a set of compounds. Other approaches are possible, for example, the highlighting of highest-scoring common substructures as illustrated by Sheridan and Miller,[5] but we have thus far used Stigmata only in the VisualiSAR program.

Cluster analysis has long been used for grouping chemical structures.[4,6,7] Clustering methods take a set of descriptors for each point in a data set, and then attempt to group points together that have similar values of the descriptors. These methods generally fall into one of two categories: *hierarchical*, in which a hierarchy of clusterings is produced, from a single cluster at the top of the hierarchy to each point in its own cluster at the bottom, and *nonhierarchical*, which produce a single clustering based on an input set of parameters. We have used a hierarchical cluster analysis technique in VisualiSAR called *Ward's clustering*[8] for grouping compounds into structurally similar groups, using the fingerprint bits as the descriptors. Ward's clustering has been found to be effective in a number of chemical structure-related applications, including separating known actives from inactives,[9] property prediction,[10] and the selection of a diverse set of compounds.[11] Since Ward's clustering produces a hierarchy instead of a single clustering, the user must manually navigate the hierarchy unless some automated way can be found to select a cluster level. We use such a method, suggested by Kelley *et. al.*[12] and used in a commercially available chemical structure clustering package[13]; it attempts to minimize both the number of clusters and the mean interpoint distance within clusters. This method has served well in studies to date in providing a good default display choice that accords well with chemical expectation, with reasonable computational requirements. A comprehensive study of this and some other measures for level selection has led to this choice, and the results of this study will be reported separately.[19]

PROGRAM DESIGN

The VisualiSAR program resides entirely on a Web server, and is accessed using a browser such as Netscape or Internet Explorer. The application is written mainly in Perl,[14] a scripting language commonly used for Web applications, which communicates with the browser through a CGI interface. The Perl code in turn uses a set of C programs that implement much of the functionality of VisualiSAR. These modular subprograms share a common file protocol for the exchange of information, and utilize a toolkit of functions for manipulating molecular information. Some of these subprograms also use the commercially available Daylight Toolkit,[3] most notably those relating to the generation of fingerprints and the depiction and coloring of 2D chemical structures.

Our philosophy in creating the user interface to VisualiSAR was to employ good interface and information design, and then to implement this design using the simplest technology possible. Computer interfaces can easily overemphasize computer administrative elements (such as dialog boxes, menus, and navigation buttons) that have little to do with the underlying information content and that can detract from it.[15] An overuse of technology can result in inefficient or unstable programs. We tried to employ a design that maximizes chemical information content and minimizes "clutter." We were able to implement such a design using simple HTML elements such as frames, and thus VisualiSAR works efficiently and robustly across different platforms and browsers.

The VisualiSAR opening screen (shown in Color Plate 1) permits several input options. SMILES[16] or MDL SD files[17] can be read in directly, or SMILES can be pasted into the paste box (e.g., from a spreadsheet). If SMILES are used, columns of associated data (e.g., compound names and biological activities) may be supplied also. Structures can be input directly from our corporate database using another in-house application, *WebRead.*

Once the structures have been read, the main VisualiSAR screen is displayed, the features of which are illustrated in Color Plate 2. On the left is a toolbar where all of the options for display, clustering, and sorting of molecules are listed. The user may make selections from the toolbar, and then click on the *GO* button at the bottom to effect them. The main part of the screen shows, initially, a grid containing an equally spaced sample of structures sorted by MSIM for the set. The structure at the top left of the grid is the highest-ranked (and therefore most representative) structure, and that at the bottom right is the lowest-ranked (and therefore least representative) structure. Each structure depiction has a title (which by default contains the name of the compound and optionally one of the supplementary data fields such as biological activity, which may have been supplied). Below the depictions are the values of the criteria on which the structures were sorted (by default MSIM and MODP). Beneath the grid are some modal statistics, and also links to a number of files that can be downloaded for further use, for example, the SMILES of the compounds displayed, or a file containing the modal fingerprint of the set.

The *molecules to display* section at the top of the toolbar allows the selection of how many and which compounds to display out of the set (or each cluster, if the compounds have been clustered). If the data set is fairly small, the user may wish to see all of the compounds. Otherwise, a sample of compounds can be displayed (giving a spread over the sorted list of compounds), or the top-ranked and bottom-ranked compounds may be displayed. The ranking by default is by MSIM, with MODP as the secondary key, although these sort keys can be changed using the menus near the bottom of the toolbar. For example, the compounds could be sorted by a biological activity value, and then the most and least active compounds displayed. The *coloring of fragments* section controls whether the compounds are colored according to the Stigmata coloring scheme, and allows the modal threshold to be set to any value (we have, however, found modal thresholds in the range of 50–80% to be most effective). A color bar is displayed beneath the grid when the coloring is turned on, to give an easy reference to the color scheme. Further options in the *other options* section allow control over the width of the depiction grid, and the size of the depictions. One feature of using HTML frames is the easy implementation of a *split screen* option, which allows a second instantiation of VisualiSAR to be run in the lower half of a split screen. This is useful for comparing data sets, an application that is illustrated later in this article.

Clustering is effected by selecting the *show clusters* checkbox. An indication is given of the time required to cluster the compounds, and the minimum cluster size (below which compounds are considered "unclustered") may be specified. After the compounds have been clustered, the main screen contains one grid for each cluster, and the clusters may be navigated using the scroll bar, or by clicking on links in the toolbar. Normally, unclustered molecules are shown together in a separate grid. An example of the VisualiSAR screen after clustering (and coloring) is given in Color Plate 3. By default, the clustering shown is that selected by the Kelley method as being the best. However, this will not always be the clustering that is most useful, and therefore options are made available for navigating the cluster hierarchy, either by a simple *try different clusters* selection (which moves to Kelley's next best level), or by specifying a specific cluster level. Files containing hierarchy information (such as the number of singletons and clusters and their sizes at each level) are made available so that full use can be made of the hierarchy if desired. A third option allows a level to be selected with a maximum cluster size, which is useful if the selected level has a large cluster that needs to be subdivided.

EXAMPLES OF USE

Finding relationships between structure and a continuous variable in a small data set

We frequently need to determine relationships between structure and a measured value such as percent inhibition, IC_{50}, or bioavailability, in sets of a few hundred compounds. When handling this kind of data set, we have found it useful to use the following strategy with VisualiSAR:

1. Cluster all of the compounds.
2. Sort each cluster by the measured value or MSIM.
3. Use the Stigmata coloring to highlight similarities and differences within the clusters.
4. Look for key differences in clusters that cause a large change in the measured value.

For example, we took a set of 651 compounds from a variety of sources, each with a measured percentage oral bioavailability, and clustered them. The selected cluster level contained 67 clusters, identifying distinct series present: for example, a set of 14 penicillin-derived compounds. A sample of nine of these compounds is shown in Color Plate 4, with the names and bioavailabilities shown above the structures. The compounds are sorted from top left to bottom right by activity. The compounds are also colored using the Stigmata coloring scheme, so that common (or "stable") regions of the compounds are highlighted in white and blue, and the less common (or "variable") regions are highlighted in green, yellow, and orange. The coloring shows that variability between the compounds occurs adjacent to the side-chain carbonyl, with the exception of amdinocillin, which contains a different side chain altogether. The three most bioavailable compounds (sorted to the top) all contain α-amino substitutions, and the presence of polar hydroxy and azido groups in amoxicillin and azidocillin appear to significantly improve bioavailability over ampicillin. On the next row, penicillin-V and nafcillin may be compared with penicillin-G, showing that the presence of an ether on the side chain is good for bioavailability. Finally, it can be seen that the amdinocillin side chain and the side chain α-carboxyl substitution in carbenicillin appear to render the compounds nonbioavailable. Note that ticarcillin has no bioavailability value and is thus sorted to the end. We might assume that the presence of the carboxylic acid and absence of an α-amino substitution in ticarcillin would mean that it is likely to have a low bioavailability.

Sorting the compounds by MSIM instead of bioavailability (Color Plate 5) conveys the same information in a different way. In this rendering, the compound at the top left is that which is most similar to what is common to the entire set (i.e., the most representative compound, in this case penicillin-G). Subsequent compounds diverge from what is common by increasing amounts. Thus, penicillin-G, ampicillin, carbenicillin, and amoxicillin are sorted to adjacent positions, highlighting the minor ways in which they differ, but also that these small changes greatly affect bioavailability, shown in the title line.

Analysis of a large HVS-like data set

When a large number of structures need to be analyzed (more than ~1000), such as with a high-volume screening (HVS) data set, we employ a different strategy. Clustering such data sets tends to be time-consuming (especially for extremely large data sets of tens of thousands of compounds), and also produces more clusters than can be easily analyzed visually. We instead employ the following procedure.

1. Separate compounds into those considered "active" and those considered "inactive" (possibly leaving some low and moderate activity compounds out).
2. Cluster the actives.
3. Examine clusters for interesting chemical series.
4. Save interesting clusters and their modal fingerprints for further examination.
5. Search inactives with the modal fingerprint of each cluster to retrieve additional compounds tested from the same series. Score and rank these structures by MODP.

6. Re-analyze all the members of the individual series and use Stigmata coloring to look for common or unique features relating to activity or lack thereof.

Our example is drawn from experiments we did using the NCI anti-HIV data set (May 1997 version), downloaded from the NCI's internet site.[18] The data set contains 32,110 compounds screened for anti-HIV type 1 activity in a cell-based assay, and classified into active (100% inhibition in the assay), moderately active (at least 50% inhibition), and inactive (less than 50% inhibition). It is similar to a typical corporate data set HVS screen in that it is diverse, but contains distinct chemical series, and has nominal rather than continuous activity data. After conversion to SMILES (which was not possible for a few of the structures), we were left with 198 active compounds, 380 moderately active conpounds, and 29,114 inactive compounds. For these experiments, we considered both active and moderately active compounds to be "active."

These 578 compounds were clustered by VisualiSAR and a level containing 39 clusters and 151 singletons was selected. One of these clusters contained 17 thiazolobenzimidazoles, 4 of which were classified as active, the rest being moderately active. We used the split-screen facility of VisualiSAR to compare the compounds in this cluster with the most similar inactives retrieved by a modal search of the inactive set, using the 50% modal of the active cluster as a query, and sorted by decreasing MODP and MSIM. This can be seen in Color Plate 6. In the top section, a sample of 10 of the active compounds are colored by the 100% modal, to highlight the template (in this case the phenylthiazolobenzimidazole). In the lower section, the inactive compounds ranked most similar to the active cluster are shown. The inactive compounds are also colored using the 50% modal of the actives, thus highlighting in green, yellow, and red the features in the inactives that are not common with the actives, and which therefore may be associated with inactivity. For example, the position of halogen substitutions on the phenyl is clearly important: the three highest ranked inactives all have difluoro substitutions like some of the actives, but the coloring highlights the fact that the actives do not contain ortho or para substitutions. Looking further down the inactive list, substitutions on the benzimidazole are clearly not common in the actives (although some do retain activity with these substitutions). The thione form is not found in any of the actives.

CONCLUSIONS

We have shown that modal analysis and Stigmata coloring enable the discernment of the relationships between chemical features and biological activity or other properties in a way that can be interpreted using 2D structure diagrams, and without the need for specialist statistical or computational knowledge. Clustering compounds into structurally related groups before applying these techniques enables them to be applied to data sets that are too large and diverse to be assimilated by manual inspection of the structures, such as those coming through high-volume screening experiments. By bringing these methods together in a simple-to-use application, we have been able to make SAR analysis accessible to a wide group of people at Parke-Davis. At the time of writing, VisualiSAR is being widely used in our computational group and is starting to be adopted by bench chemists for organizing and visualizing their SAR data. The program has found use beyond the chemistry department, particularly in the analysis of relationships between structure and pharmacokinetic data. Also, we have found that the Stigmata coloring has proved particularly helpful in interdisciplinary projects where there is a need to convey SAR information to nonchemistry colleagues. We would like to compile statistics on the use of VisualiSAR in these different areas, although the program has not been available long enough to do this at the present time. We are currently assessing ways in which the interface can be improved, as well as developing VisualiSAR to incorporate other types of fingerprint, and integrating it into other applications that require or are enhanced by a structure-browsing capability.

For proprietary and practical reasons we have not made the full code for VisualiSAR available, although the code for generating Stigmata colorings is available as contributed code from Daylight.[3]

ACKNOWLEDGMENTS

We would like to thank George Cowan and T. J. O'Donnell for technical advice; Alain Calvet and Christine Humblet for supporting the work; Norah MacCuish and Jeremy Yang of Daylight CIS Inc., for assistance in integrating the Stigmata coloring into the structure depictions; and Alain Calvet, Eric Gifford and George Cowan for helpful comments on the first draft of this paper.

REFERENCES

1 Shemetulskis, N.E., Weininger, D., Blankley, C.J., Yang, J.J., and Humblet, C. Stigmata: An algorithm to determine structural commonalities in diverse datasets. *J. Chem. Inf. Comput. Sci.* 1996, **36,** 862–871
2 Blankley, C.J. measuring molecular diversity: Evaluation of alternative subsets selected from reagent building block libraries for combinatorial chemistry. *Pharm. Pharmacol. Commun.* 1998, **4,** 139–146
3 Software and documentation available from: Daylight Chemical Information, Inc., 18500 Von Karman #450, Irvine, CA. E-mail: info@daylight.com
4 Willett, P. *Similarity and Clustering in Chemical Information Systems*. Research Studies Press, Letchworth, 1987
5 Sheridan, R.P., Miller, M.D. A method for visualizing recurrent topological substructures in sets of active molecules. *J. Chem. Inf. Comput. Sci*. 1998, **38,** 915–924
6 Barnard, J.M., and Downs, G.M. Clustering of chemical structures on the basis of two-dimensional similarity measures. *J. Chem. Inf. Comput. Sci.* 1992, **32,** 644–649
7 Willett, P., Barnard, J.M., and Downs, G.M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* 1998, **38,** 983–996
8 Ward, J.H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc*. 1963, **58,** 236–244
9 Brown, R.D., and Martin, Y.C. Use of structure–activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* 1996, **36,** 572–584
10 Downs, G.M., and Willett, P. The use of similarity and

clustering techniques for the prediction of molecular properties. In: *Applied Multivariate Analysis in SAR and Environmental Studies* (Devillers, J., and Karcher, W., eds.). ECSC, The Netherlands, 1991, pp. 247–279
11 Bayada, D.M., Hamersma, H., and van Geerestein, V.J. Molecular diversity and representativity in chemical databases. *J. Chem. Inf. Comput. Sci.* 1999, **39,** 1–10
12 Kelley, L.A., Gardner, S.P., and Sutcliffe, M.J. An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally-related subfamilies. *Protein Eng.* 1996, **9,** 1063–1065
13 Software and documentation available from: Barnard Chemical Information, Ltd., 46 Uppergate Road, Sheffield S6 6BX, UK. E-mail: barnard@bci1.demon.co.uk
14 Schwartz, R.L., and Christiansen, T. *Learning Perl*, 2nd Ed. O'Reilly, Sebastopol, California, 1997
15 Tufte, E.R. *Visual Explanations*. Graphics Press, Cheshire, CT, 1997
16 Information about SMILES is available on Daylight's Web server, www.daylight.com
17 Software and documentation available from: Molecular Design, Ltd., 14600 Catalina St., San Leandro, California 94577
18 The data set may be downloaded from http://epnws1.ncifcrf.gov:2345/dis3d/aids−screen/aidspub.html
19 Wild, D.J. and Blankley, C.J. A comparison of 2D fingerprint types and hierarchy selection methods for structural grouping using Ward's Clustering. *J. Chem. Inf. Comput. Sci.* 2000, **40,** (in press)