General Purpose Interactive Physico-Chemical Property Exploration

Andrew Smellie*

ArQule Inc., 19 Presidential Way, Woburn, Massachusetts 01801

Received February 15, 2006

There have been many tools and methods developed to investigate structure−activity (SAR) and structure−property (SPR) relationships. Many of these tools are fully or almost fully automated and attempt to predict various properties of molecules. Even with these tools, there is still an urgent need to provide a simple and useful methodology so a consumer of these various models, such as a medicinal chemist or molecular modeler, can learn how various properties of small molecules relate to their structure. This paper describes a new viewing and navigation paradigm called the Spiral View, which can be used to facilitate the interpretation of structure−property relationships. Examples of how this tool proves useful in the *human* interpretation of these relationships are drawn from the fields of experimental toxicity, LogP, and biological activity data drawn from various sources.

INTRODUCTION

There is a very large number of methods available for the prediction and rationalization of chemical structure−activity[1−5] (SAR) and structure−property[6−9] (SPR) relationships. For the purposes of this paper, the term SPR will be used to cover both cases of SAR and SPR. It is more common to find methods that attempt prediction of properties than to find methods that facilitate the *understanding* of the subtleties of changes in chemical structure as it relates to change in the property (or properties) of interest. Often, a consumer of property data is trying to decide which compounds to make next to boost or diminish some targeted property. For example, the goal might be to maximize the biological activity of a small molecule as measured by IC50,[10] or to minimize the toxicity as measured by LD50.[11] A tool that simply predicts properties does not necessarily provide insight into why a particular molecule is active and, hence, makes the decision of what to make next very difficult.

There are a few techniques in existence that help to understand SPRs, and each has their advantages and disadvantages.

The *substituent spreadsheet*[12−14] is a common method where molecules are presented in a 2D spreadsheet. Each cell of the spreadsheet is a molecule that is often color-coded by the value of the property being studied. The rows and columns of the spreadsheet are substituents of the molecules in the cells. This representation is very versatile as the rows and columns of the spreadsheet can be sorted in an attempt to group molecules with similar properties together. However, it is often not easy to identify key relationships in a large data set, and the method is only really useful if the molecules share a common core.

The LeadScope[15] approach is a very powerful methodology that combines model building with a spreadsheet representation and attempts to directly identify key features that correlate with the property of interest. The identified features are color-coded for easy identification. The method works best when robust models can be derived from the data, which is not always possible for data sets with diverse structures.

The SAR Tree[16] is a newly disclosed method for visualizing very large SPR data sets from high-content screening. It is a fractal-like, recursive drill down view of color-coded cells, where each cell is a tested compound. Grouping behavior is often observed when active compounds are discovered and the actives share a common structure.

The Spiral View is complementary to all of the above methods and has proved valuable in the interpretation of SPR data. It take a microscopic view of the SPR data and permits the easy navigation of sequences of "interesting" pair-wise property relationships. Using a set of simple and well-documented SPR relationships, it will be shown that the Spiral View is a facile way of discovering structure−property relationships in a large and diverse data set.

THE SPIRAL VIEW

The Spiral View originated from an empirical observation: the most useful human-interpretable pair-wise comparisons are those where there is a large change in the property of interest but a small change in structure. A large number of such comparisons can reveal trends in structure−property relationships. The Spiral View is a method of simply identifying and displaying such relationships.

The prototypical spiral view is shown in Figure 1. There is always a *central molecule C*, defined as the current molecule of interest. Molecules that are "closest" to the central molecule are arranged in a clockwise spiral with the closest neighbor of C at the 12 o'clock position, and the next closest neighbors arranged in a clockwise spiral ordered by increasing distance. This is the first visual cue: the closest molecule is always in the same place with the next closest neighbors arranged in a consistent orientation. For the purposes of this paper, the distance is measured as the simple Tanimoto distance using Pipeline Pilot[17] fingerprints. The View can support any arbitrary user-defined distance. Each molecule is rendered in a cell that can display the values

* Author e-mail: asmellie@arqule.com.

10.1021/ci060052t CCC: $37.00
Published on Web 05/08/2007

(labels) of up to two properties (one above and one below) for each molecule, depicted as label 1 and label 2, respectively.

The length of the lines drawn between the central molecule and its closest neighbors is proportional to the distance between the pairs of molecules. Because the molecules are arranged around the central molecule in order of increasing distance, the display forms a spiral, hence the name Spiral View. The width of the line rendered is proportional to the *difference* in the property value associated with the line, which is user-controllable. Large property differences result in thicker lines. This is the second visual cue: the user is quickly drawn to thick lines that denote large property differences.

The default view is useful but becomes especially valuable if sequences of views can be seen simultaneously. Left-clicking on any neighbor of the central molecule brings up a new Spiral View in a different window, with the selected neighbor being the new central molecule in the new view. Linking these together forms an *SPR chain*. These chains have proven valuable at uncovering SPRs in a way that is easy to interpret.

Figure 2 shows how the Spiral View can be adapted for different data sets. Each molecule is rendered in a cell with two properties displayed. Text boxes 1 and 2 are pull-down menus that allow different properties to be labeled. The central compound can be linked to each of its spiral neighbors by up to three different lines; thus, multiple property relationships can be studied. Text box 3 allows the user to select which property will be represented by the first line and the scaling factor to be applied to the measured property differences. This method can be used to display multiple properties that might be on very different scales. For example, if molecular weight (MW) is a property, the difference in MW could be on scale of 100s. Thus, scaling the difference by factors of 10−50 may be appropriate when determining the thickness of the line. Conversely, if LogP is the property, differences on the order of single values are expected, so these differences should be scaled by factors of 0.5−1.0.

The thickness of the lines is given by the simple equation:

$$T(n,c,i) = \mathrm{MIN}\left[M,\mathrm{MAX}\left(1,\frac{|p(c,x) - p(n,x)|}{F(i)}\right)\right] \quad (1)$$

where $T(n,c,i)$ is the thickness of the ith line (in pixels) joining the central compound C and its nth nearest neighbor N, $p(c,x)$ is the value of property X for C, $p(n,x)$ is the value of property X for N, $F(i)$ is the scaling factor to take into account that property differences are on different scales, and M is the maximum thickness in pixels (12).

Button **4** in Figure 2 allows the user to change the line color to aid in interpretation. Text boxes **5** and **7**, and buttons **6** and **9**, allow multiple lines to be specified, allowing multiple properties to be displayed simultaneously. Text box **10** allows the user to change the definition of the distance used to find the nearest neighbors of the central molecule. For this paper, the distance is the simple structural fingerprint distance but, in other studies, might be might be an alternate molecular similarity metric such as pharmacophore similarity,[18] feature tree similarity,[19] or shape superposition.[20]

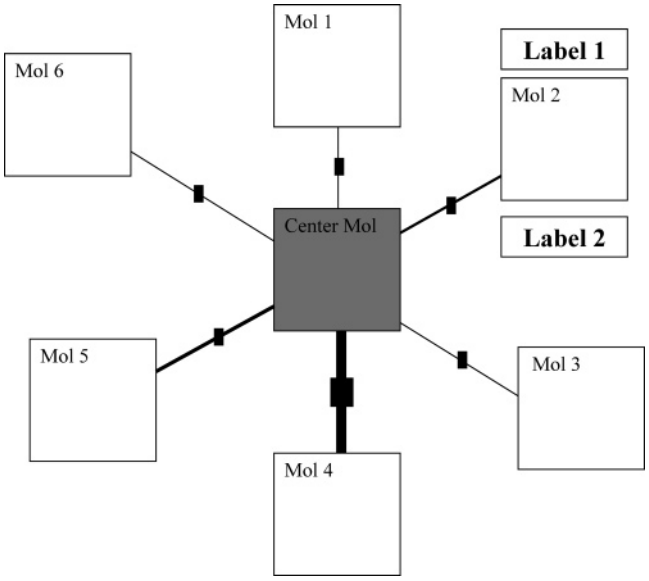

Figure 1. Spiral view prototype.

Case Study 1: Experimental Toxicity. For this example, the Fathead Minnow Acute Toxicity[21] data set was selected for analysis. It was selected because it presents a particular challenge of QSAR tools because of the high diversity of chemical structures in the data set. A random selection of the 617 structures is shown in Figure 3. From casual inspection, a large number of chemotypes and substituents can be observed. The raw SDF file downloaded from http://www.epa.gov/nheerl/dsstox/sdf_epafhm.html was processed with Pipeline Pilot to generate parameters suitable for a Spiral View analysis: namely, structural fingerprints, MW, and LogP.[22] There are many structure−property inferences that can be made from this data set, but as an illustrative example, a well-known toxicity phenomenon is investigated: that nitro groups are generally toxic to these fish.[23] A simple SPR chain was generated from a series of three spiral views, as shown in Figure 4, that explores the hypothesis that, in general, when more nitro groups are present in the molecule, it tends to be more toxic. In this chain, only the closest three neighbors of each central molecule are shown for clarity, but up to nine neighbors can be displayed. Starting from the molecule in cell 1 of Spiral View 1, it can quickly been seen that the LD50 (bottom number in the labels on the cells) appears to be low (i.e., more toxic) for molecules with two nitro groups as opposed to one. The Spiral View makes this very easy to see by following the thicker lines (that imply a greater difference between LD50 values). To further explore the hypothesis, the molecule in cell 2 with one nitro group is left-clicked to bring up Spiral View 2 with this molecule as the central molecule in cell 3. The hypothesis seems to hold that more nitro groups in a molecule makes it more toxic, but to explore further, the molecule in cell 4 is left-clicked to bring up Spiral View 3. In this view, a very compelling example is found by comparing cells 5 and 6. These pairs of molecules only differ by replacing nitro with hydroxyl, with an almost 100-fold decrease in toxicity. So from a short sequence of Spiral Views, a series of pair-wise comparisons was discovered that supports the hypothesis. This simple example illustrates how sequences of Spiral Views can be used to quickly guide a user toward key relationships in order to *understand* the data.

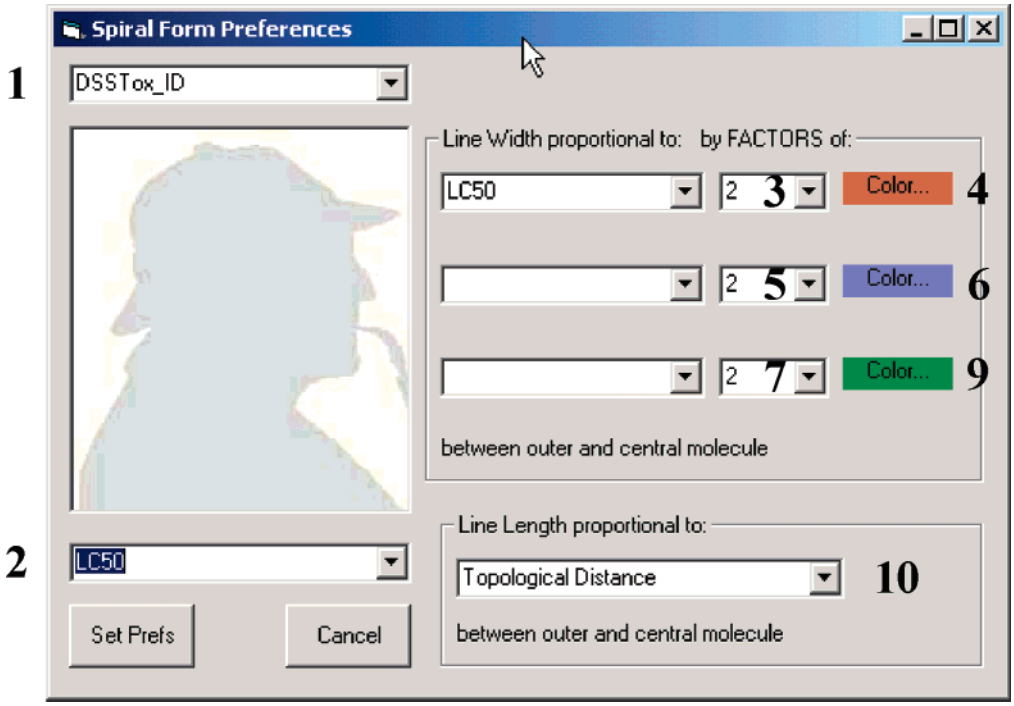

Figure 2. Changing the spiral view.

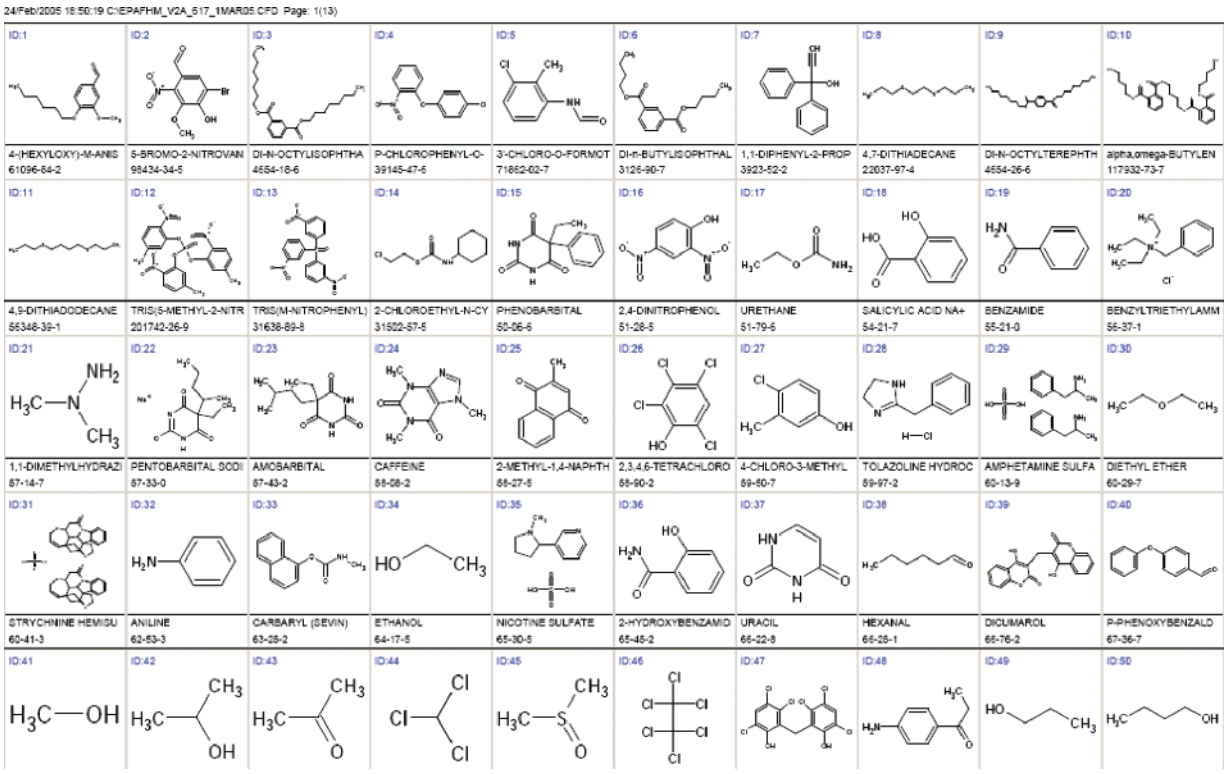

Figure 3. Representative molecules from the fathead minnow data set.

Case Study 2: Exploring the Relationship between AlogP and Molecular Weight. Using the same data set, the relationship between AlogP and molecular weight was explored through sequences of Spiral Views. This relationship is clearly understood and has been well-documented elsewhere[24] but is shown here as a clear example of how Spiral Views can be used to understand and rationalize two properties simultaneously. Figure 5 shows a couple of short sequences that clearly support the case that (a) AlogP decreases in the presence of more hydrophilic groups and (b) AlogP and molecular weight are positively correlated in typical druglike compounds. Comparing cell 1 and cell 2 in Spiral View 1 shows that increasing MW (the top number in each cell) and hydrophobicity (by adding Br atoms) greatly increases the AlogP (the bottom number in each cell). The thick lines act as an immediate visual cue to draw attention to that fact. Left-clicking on cell 2 brings up the new Spiral View 2 with the selected molecule in cell 3. The thick lines to cells 5 and 6 show that, as hydrophilicity increases, AlogP (and MW) falls. There are thinner lines to cell 4 that contain

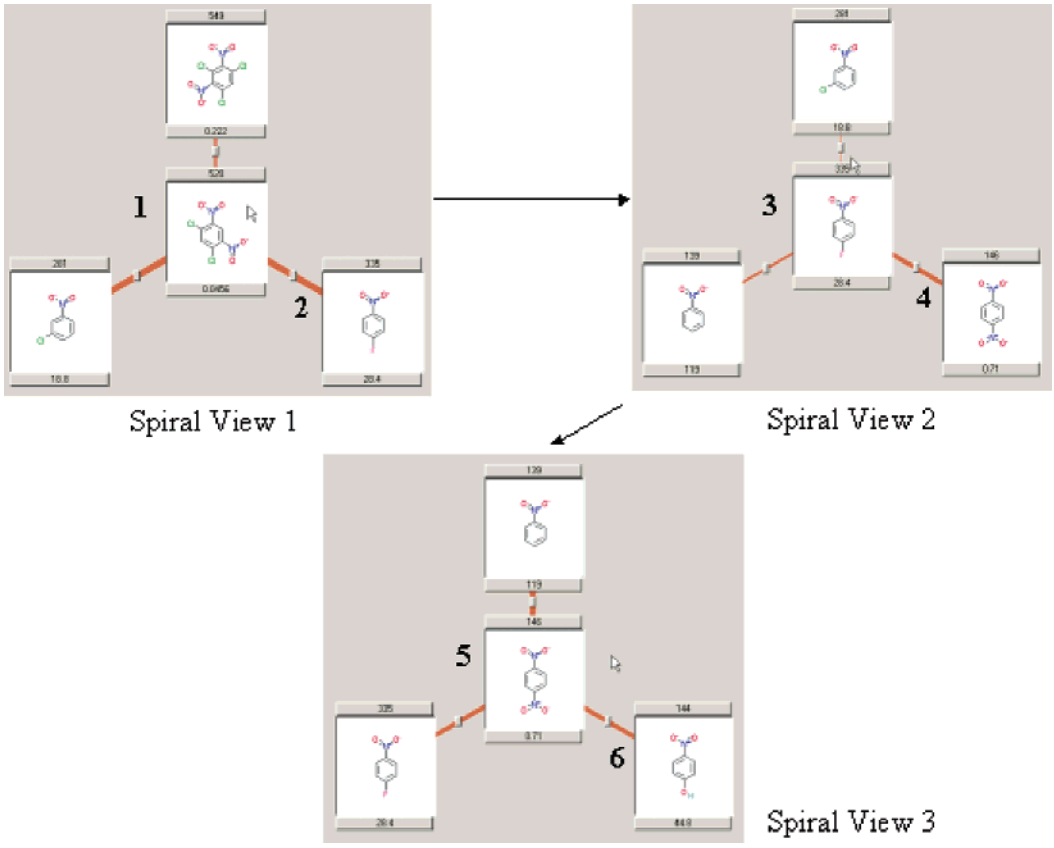

Figure 4. SPR chain exploring toxicity relationships.

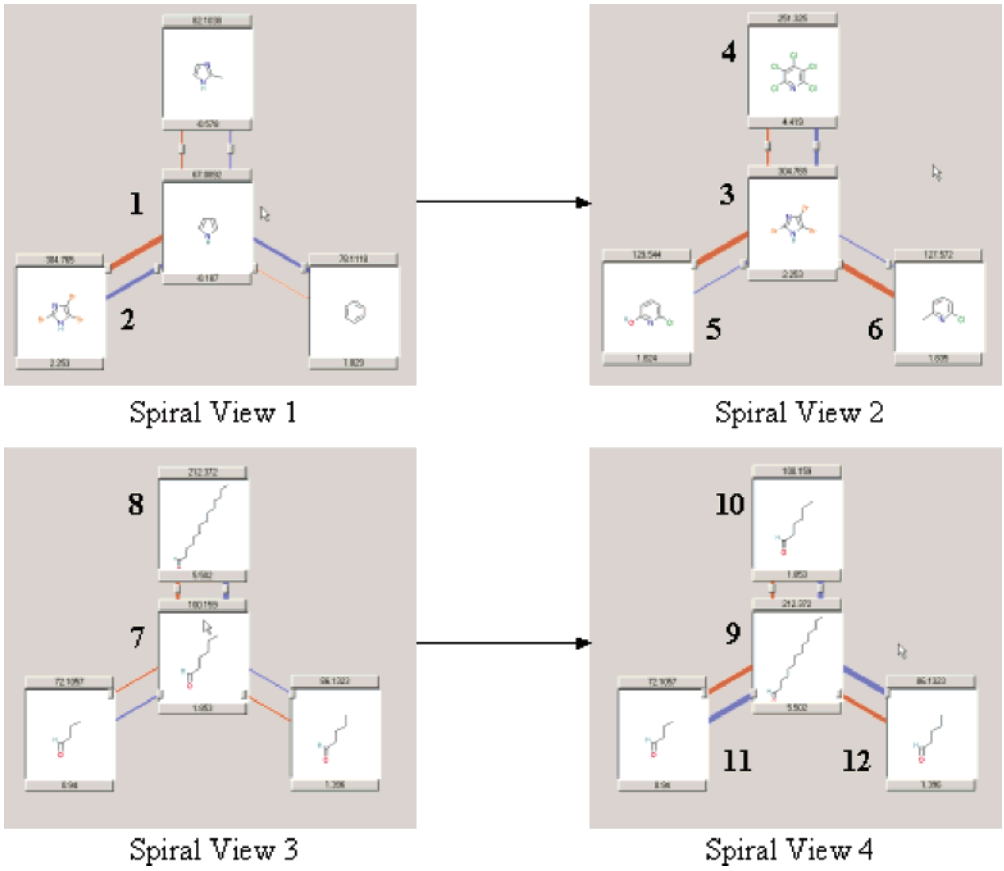

Figure 5. SPR chains to explore ALogP and molecular weight.

a hydrophobic molecule. A different sequence is shown starting with Spiral View 3. This sequence generally investigates the effect of carbon chain length on AlogP. The views show that, as chain length increases, AlogP increases

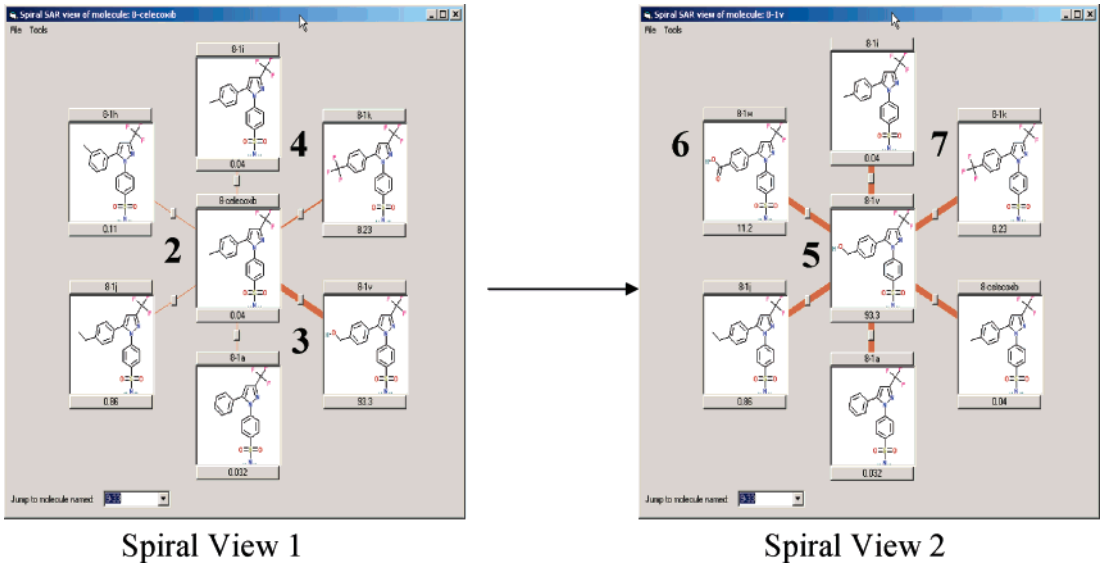

Figure 6. SPR chains to demonstrate key COX-2 activity.

(as does MW). Though a trivial result, it is striking how easily the relationships can be seen, especially in the pairings cell 9−10 and cell 9−11, which show clearly, via increasingly thick lines, that, as chain length increases, AlogP increases.

Case Study 3: Biological Activity. The Spiral View can also be used to investigate SAR relationships from biological screening. As a final illustrative example, various spiral views are shown from a COX-2 data set that quickly and efficiently demonstrates key structural properties of various COX-2 inhibitors. The SAR literature around selective COX-2 inhibitors is vast,[25−28] and it is beyond the scope of this paper to describe all SAR trends discernible from the data set.[29] The utility of Spiral Views is illustrated with a couple of examples. In Figure 6, Spiral View 1 suggests that the terminal phenyl group is important for activity and more specifically, potency is increased by the presence of hydrophobic groups at the para position of this phenyl ring. Compounds **3** and **4** have hydrophilic substituents, and the potency is decreased. Left-clicking on compound 3 brings up a Spiral View 2. In this view, the central compound (with a hydrophilic substituent on the terminal phenyl) is less active than all of its nearest neighbors (quickly identified by following the thick lines). As confirmation of the terminal hydrophobic hypothesis, all neighbors of compound **5** that have hydrophobic substituents are significantly more potent. Compounds **6** and **7** have hydrophilic substituents and are consequently not as potent as those with hydrophobic substituents. Thus, from a couple of views, a hypothesis of activity can be visually observed (in Spiral View 1) and tested (in Spiral View 2). This relationship has been identified in ref 25.

CONCLUSIONS

This paper has introduced a new viewing paradigm for SAR and SPR data: the Spiral View. Drawing on toxicity, physicochemical, and biological activity data, it has been shown that the Spiral View is a powerful and facile method of visually extracting and testing hypotheses from large data sets. The example shown in the paper were of known SPR relationships for the purposes of illustration, but the tool is in routine use internally where many new observations have been made on internal proprietary projects.

ACKNOWLEDGMENT

The author wishes to acknowledge Rocio Palma and Anton Filikov for many useful discussions and proof reading. He also wishes to thank Judy Blaine for aid in gathering test data and references.

REFERENCES AND NOTES

(1) Hansch, C. Quantitative Structure−Activity Relationships. In Drug Design; Ariens, E. J., Ed.; Academic Press: New York, 1972; Vol. I, 271−342.
(2) Cammarata, A. Interrelation of the Regression Models Used for Structure−Activity Analyses. *J. Med. Chem.* **1972**, *15*, 573.
(3) Rohrbaugh, R. H.; Jurs, P. C. Descriptions of Molecular Shape Applied in Studies of Structure/Activity and Structure/Property Relationships. *Anal. Chim. Acta.* **1987**, *199*, 99−109.
(4) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(5) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34* (4), 854−866.
(6) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J. Phys. Chem.* **1998**, *102*, 3762−3772.
(7) Micheli, A.; Portera, F.; Sperduti, A. QSAR/QSPR Studies by Kernel Machines, Recursive Neural Networks and Their Integration. *Neural Nets, WIRN VIETRI 2003*, Vietri sul Mare, Italy, June 4−7, 2003; *Lecture Notes in Computer Science, LNCS Vol. 2859*; Apolloni, B., Marinaro, M., Tagliaferri, R., Eds.; Springer-Verlag: Berlin, Heidelberg, 2003; pp 308−315; Revised Papers, ISBN: 3-540-20227-7
(8) Maw, H. H.; Hall, L. H. E-State Modeling of Corticosteroids Binding Affinity: Validation of Model for Small Data Set. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1248−1254.
(9) Rytting, E.; Lentz, K. A.; Chen, X.-Q.; Qian, F.; Venkatesh, S. A Quantitative Structure−Property Relationship for Predicting Drug Solubility in PEG 400/Water Cosolvent Systems. *Pharm. Res.* **2004**, *21* (2), 237−244.
(10) IC50: The concentration of an inhibitor that is required for 50% inhibition of an enzyme in vitro.
(11) LD50: A chemical dose lethal to 50 percent of a test population.
(12) As implemented in: *Sybyl*; Tripos, Inc.: St. Louis, MO 63144-2319.
(13) As implemented in: *MOE*; Chemical Computing Group: W. Montreal, Quebec, Canada H3A 2R7.
(14) As implemented in: *Cerius2*; Accelrys Inc.: San Diego, CA 92121.
(15) Leadscope Inc., 1393 Dublin Road, Columbus, Ohio 43215.

(16) Singh, J. *Technologies for Small Molecule Drug Design*; Bio IT World: Boston, MA, May 17−19, 2005.
(17) *Pipeline Pilot*; SciTegic: San Diego, CA 92123-1365.
(18) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem.* **1999**, *42* (17), 3251−64.
(19) Rarey, M.; Dixon, J. S. Feature Trees: A New Molecular Similarity Measure Based on Tree Matching. *J. Comput.-Aided Mol. Des.* **1998**, *5*, 471−90
(20) Hahn, M. Three-Dimensional Shape-Based Searching of Conformationally Flexible Compounds. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 80−86
(21) Russom, C. L.; Bradbury, S. P.; Broderius, S. J.; Hammermeister, D. E.; Drummond, R. A. Predicting Modes of Action from Chemical Structure: Acute Toxicity in the Fathead Minnow (Pimephales promelas). *Environ. Toxicol. Chem.* **1997**, *16* (5), 948−967.
(22) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure−Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci*. **1989**, *29*, 163−172.
(23) Bailey, H. C.; Spanggord, R. J. *Aquat. Toxicol. Hazard Assess.*; ASTM Spec Tech Publ 802: West Conshohocken, PA, 1983; pp 98−107.
(24) Huuskonen, J. Estimation of Aqueous Solubility in Drug Design. *Comb. Chem. High Throughput Screening* **2001**, *4* (3), 311−316.
(25) Kalgutkar, A. S.; Zhao, Z. Discovery and Design of Selective Cyclooxygenase-2 Inhibitors as Non-Ulcerogenic, Anti-Inflammatory Drugs with Potential Utility as Anti-Cancer Agents. *Curr. Drug Targets* **2001**, *2*, 79−106,
(26) Talley, J. J. *Prog. Med. Chem*. **1999**, *36*, 201−234.
(27) Carter, J. S. *Expert Opin. Ther. Pat*. **1998**, *8* (1), 21.
(28) Prasit, P.; Riendeau, D. *Annu. Rep. Med. Chem*. **1997**, *32*, 211−220.
(29) Sutherland J. J.; O'Brien. L. A.; Weaver, D. F. A Comparison of Methods for Modeling Quantitative Structure-Activity Relationships. *J. Med. Chem*, **2004**, *47* (22), 5541−5554.

CI060052T