BUTTERWORTH HEINEMANN

Navigator: Tools for informal structure–activity relationship discovery

David Chapman,* Nomi Harris, John Park,† and Roger E. Critchlow, Jr.‡

Arris Pharmaceutical Corporation, South San Francisco, California 94080

Navigator is a molecular database visualization system, designed to support exploratory data analysis and informal structure–activity relationship studies. In addition to the operations commonly found in chemical database systems, it provides new tools that facilitate substituent analysis and help elucidate the relationships among similar molecules and between related assays. Navigator's capabilities include two ways of displaying the relationships between analogs, mouse-sensitive charts of sets of molecules, mouse-sensitive plots of assay relationships, and access to a system for three-dimensional quantitative structure–activity relationship discovery. Navigator's mouse-based user interface provides a one-object/one-window paradigm that makes data manipulation easy even for inexperienced users. Navigator runs on Silicon Graphics workstations.

Keywords: Compass, drug discovery, graphical user interface, maximal common subgraph, molecular database, molecular similarity, Navigator, structure–activity relationships

INTRODUCTION

Modern computational tools for automatic analysis of molecular structure–activity relationships can often give both accurate quantitative predictions for the activities of proposed molecules and structural insight into the determinants of these activities.[1] However, for a variety of reasons, many drug discovery projects rely on informal studies of structure–activity relationships, using a chemical database as the only computational tool. Sophisticated, quantitative tools may have limited applicability, may be difficult and expensive to use, and generally require extensive user training.

We have developed a molecular database visualization system called Navigator. The system integrates a suite of simple, intuitive tools that bridge the gap between sophisticated quantitative tools and existing chemical database systems such as ChemFinder, Daylight, ISIS/Base, and UNITY.[2] Navigator's new tools are designed to support database visualization, exploratory data analysis, and informal structure–activity relationship studies, particularly substituent analysis.

We describe capabilities of Navigator that include two ways of accessing substituted analogs of a molecule; high-density, mouse-interactive display of sets of structures with selected properties; and interactive scatterplots that allow the user to choose a data point with the mouse and thereby view the corresponding molecule. Navigator provides, in addition to such new tools, operations already familiar to users of existing chemical database systems. Its simplicity makes it easy to use without any knowledge of computers (beyond the use of a keyboard and mouse). Although Navigator is intended primarily for use by medicinal chemists, we have also found it valuable in preparing to apply more sophisticated tools. Navigator also serves as a user interface to several such tools we have developed. All of Navigator's capabilities, novel and familiar, are integrated into a uniform framework on the basis of a set of design principles enumerated in the next section.

DESIGN PRINCIPLES

Navigator is conceptually organized around *objects* and *styles*. Every Navigator operation consists of displaying some object in some style. Displaying an object creates a window, which contains a visualization of the object. The one-window/one-object principle provides an intuitive paradigm for manipulating data. Navigator makes everything visible: the complete set of operations you can perform at

Color plates for this article are on pages 240 and 241.

*Address reprint requests to Dr. Chapman at Arris Pharmaceutical Corporation, 385 Oyster Point Boulevard, South San Francisco, California 94080.

†Currently in Section on Medical Informatics, Stanford University, Stanford, California.

‡Present address: 1109 Castro Street, San Francisco, California 94114.

Received 4 January 1995; revised 6 April 1995; accepted 21 April 1995

Journal of Molecular Graphics 13:242–249, 1995

655 Avenue of the Americas, New York, NY 10010

0263-7855/95/$10.00
SSDI 0263-7855(95)00034-4

any given time is always visually available as a set of buttons.

Each *type* of object can be displayed in any of several styles. For example, a molecule is a type of object; we may want to display a molecule either in two-dimensional or three-dimensional style. A set of molecules is also an object; we might display it as a scatterplot, showing assayed activities, or as a chart, showing the structure of each molecule.

Navigator evolved as a series of responses to specific user needs. However, in producing the final product, we have put much effort into the human factors design, to ensure a coherent whole. The next section, which describes the general user interface, illustrates this point. The following sections explain the unique features of Navigator; a final section provides implementation details.

DATABASE ACCESS: BASIC DISPLAY STYLES

In this section, we describe our approach to operations familiar from conventional chemical databases. We take many of our examples in this section, and throughout the article, from a set of steroids assayed for corticosteroid-binding globulin (CBG) and testosterone-binding globulin (TeBG) activity (Figure 1), which has been the subject of several quantitative structure–activity relationship (QSAR) studies.[3]

The Navigator Main Panel (Figure 2) provides access to all system operations; it is the only window that does not itself display an object. The Main Panel lets you choose an object to display and a style to display it in. For example, to display a molecule, we click on Molecule in the Object Type menu in the center of the left column. This updates the Display Style menu at the bottom of the left column to show the styles in which molecules can be displayed. Then we select a molecule either by choosing from the menu of molecule names on the right or by typing its name into the text box just below the menu. This pops up a new display window showing the molecule in the selected style. For instance, Figure 3 shows molecule **6** (corticosterone) in the 2D style. This style shows the two-dimensional structure of the molecule, together with all the information known about it displayed textually. This information can include the assayed activities of the molecule, physical properties such as melting point, and synthetic history.

By providing a separate window for each object displayed, Navigator allows the user to rearrange windows with the mouse in whatever spatial layout is most natural for the task at hand. (This contrasts with previous database access systems, in which molecules are displayed in fixed locations within a complex program frame.) For example, in examining a series of molecules with substitutions at two positions, one can organize them in rows by one substitution position and in columns by the other, with a given substitution corresponding to each row and column. Navigator allows an essentially unlimited number of display windows to exist simultaneously.

Every display window has buttons at the top for all the operations you can perform on it. By attaching the controls for an object to its window, Navigator encourages exploration and makes the system easy to learn. In the case of the window in Figure 3, only three operations are available: you can Print out a hardcopy (black and white or color) of the display, you can make the window go away (Close), or you can redisplay the molecule in a different style. When you click on Change Style, a menu appears, within the window, from which you can select alternative display styles. For example, three-dimensional structure is key to understanding the biological activity of a molecule, so one might wish to view it in any of several available three-dimensional styles. Color Plate 1 shows molecule **6** redisplayed as a CPK (space-filling) model. Such three-dimensional displays can be scaled, translated, and rotated with the three mouse buttons.

Navigator considers a set of objects to be itself an object that can be displayed and manipulated like any other object. (In alternative systems, "hit lists" are not treated uniformly with other objects.) The simplest way to display a set is as a list of the names of its members. If the user displays a set of molecules in the List style, a display window lists the name of each molecule. The molecule names in a List display are mouse sensitive: clicking on a name displays the corresponding molecule.

Another display style for sets is Sequential. In Sequential style, the window displays one member of the set at a time, and you can click through them like a slide show using the Next and Previous buttons at the top of the window. The Jump button lets the user jump to a specified member of the set.

Navigator has a facility for naming molecule sets and saving them in disk files. If Molecule Set in the Main Panel is selected as the type of object to display, a menu of names of sets appears in the center right box. If, however, the user wishes to create his own set, Navigator allows the construction of sets of molecules on the basis of their properties. The bottom box on the right side of the Main Panel allows the user to enter a description of the molecules desired. For example, to select the set of molecules that have molecular weights greater than 300 and CBG-binding less than 50 nM, one would specify "molecular-weight $>$ 300 and CBG $<$ 50nM." Navigator uses a reimplementation of the Daylight fingerprint algorithm[4] for substructure search. We use Navigator as an interface for accessing the Available Chemicals Directory database[5] of more than 100 000 commercially available compounds.

Two operations that can be performed on sets are *subsetting* and *sorting*. Both of these are available as buttons on every set display window. A set is subsetted using the same kinds of criteria that can be used to create sets initially. Sets can be sorted by any numerical property (including arithmetic combinations of properties, such as assay selectivity ratios). Navigator labels each set display with a cumulative description of the selection and sorting operations that were used in creating it.

Navigator allows data to be organized into *projects*, collections of data (molecules, assays, and so on) typically corresponding to a single drug discovery effort. The current project is displayed in the bright blue box on the left side of the Main Panel; one can click on this box to get a pull-down menu of alternatives. All operations are performed relative to the current project.

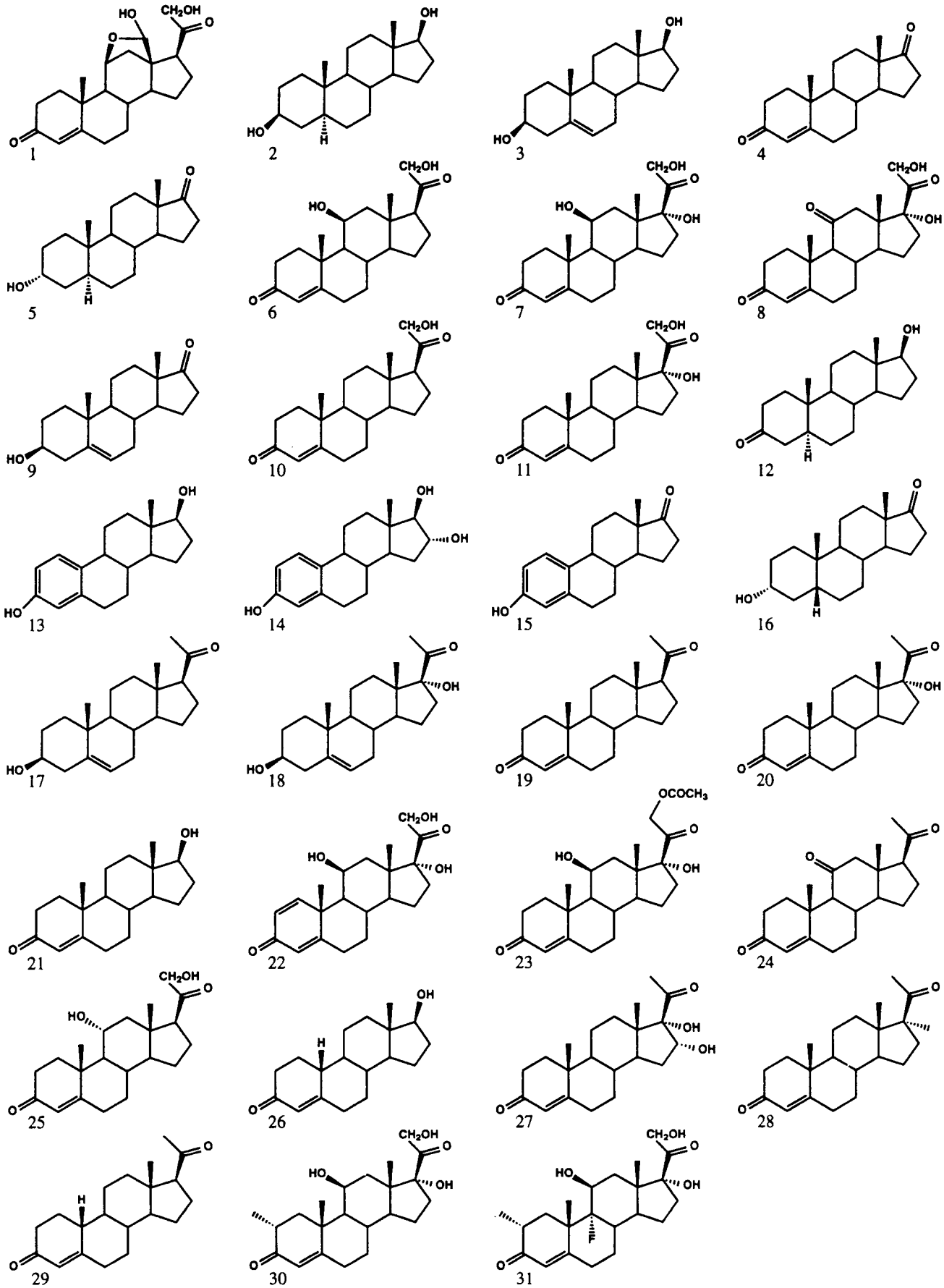

Figure 1. A set of 31 steroids, used as examples in this article.

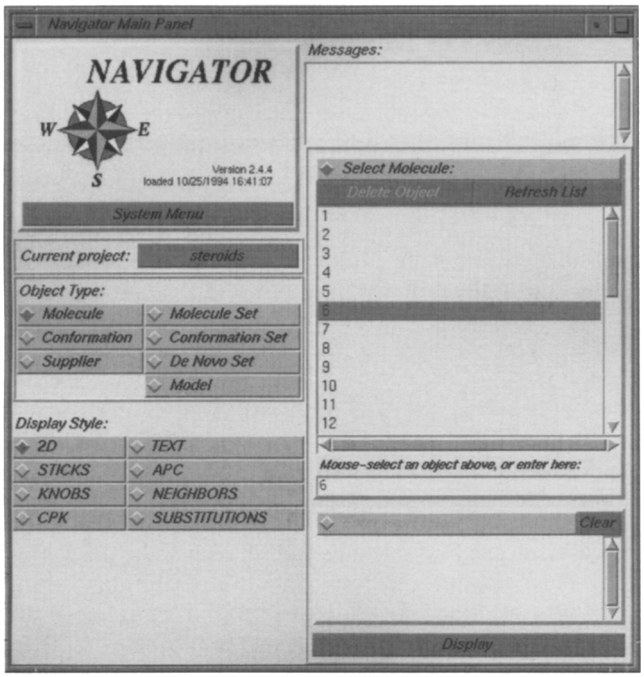

Figure 2. The Navigator Main Panel, with controls set to display molecule **6** *in the 2D style. The user clicked on Molecule to select molecule display, so the menu on the right provides a list of molecules (rather than of another sort of object); the user then clicked on* **6**, *which displayed it in a new window.*

NEIGHBORS: TOOLS FOR SUBSTITUENT ANALYSIS

We have found that informal discussions of structure–activity relationships often involve questions like "Don't we have the methoxy analog of that? Does anyone remember what its molecule number was?" At this point, either the question is dropped, or someone starts flipping through a thick stack of pages of chemical database printout, scanning each page for the relevant molecules. It is rare for the on-line version of the database to be used, because it is actually slower than paper when it's necessary to look at every molecule, and because previous chemical database systems have not provided tools that answer questions like these.

Navigator provides two related tools for accessing molecules according to similarity. Both tools define similarity in terms of substitutions, which is the way medicinal chemists most commonly think of the relationship between related compounds. We call two molecules *neighbors*, with a single substitution, if more than half their structure is identical and if one can be transformed into the other by breaking a single bond and replacing the substituent. We can define neighbors with two or more substitutions similarly. This notion of "neighbors" is quite different from heuristic similarity search methods,[6] such as Tanimoto coefficient,[4] found in some other database systems. Such similarity metrics are generally correlated with intuitive similarity, but substituted analogs are not guaranteed to have a high Tanimoto similarity, and some pairs of molecules with high Tanimoto similarity may look dissimilar by eye. The Navigator metric (number of substitutions) has a simple, algorithmic definition that makes chemical sense.

Displaying a molecule in Neighbors style (Figure 4) shows the molecule in three dimensions, with various positions labeled with letters. These are positions at which neighboring molecules differ from the displayed one. At the top of the display is a table of these neighbors with their substitutions. Navigator has a table of known substituent names and structures, which it uses to generate substituent names for Neighbors displays; when it does not recognize a substituent, it describes it as "R." The table of neighbors starts with the molecule (**10**) itself, with the substituents it has at the variable positions. Then come neighbors with a single substitution. For example, steroid molecule **6** is identical to **10** except that **6** has a hydroxyl group at the e position, where **10** has only a hydrogen (Figure 1). Neighbors with two substitutions appear further down the table, starting with **2.** Some of these have a hydrogen substitution at the c position, where the substituent on molecule **10** is described as "missing"; this is the addition of a hydrogen to the carbonyl, yielding a hydroxyl.

The Neighbors display style allows the user to answer rapidly questions such as "What was the name and assay of the allyl analog?," and "The new assay results show that 966 is significantly selective, but it's only micromolar. We made some other molecules that looked like 966 a few years ago; maybe some of them were more active?"

The Substitutions style presents neighbor information organized by the substituted positions rather than by the neighboring molecule (Figure 5). Although the Substitutions information could be derived from the Neighbors display, the Substitutions style makes it easy to isolate the effect on activity of substitutions at a particular position. Whereas each molecule appears in the Neighbors display

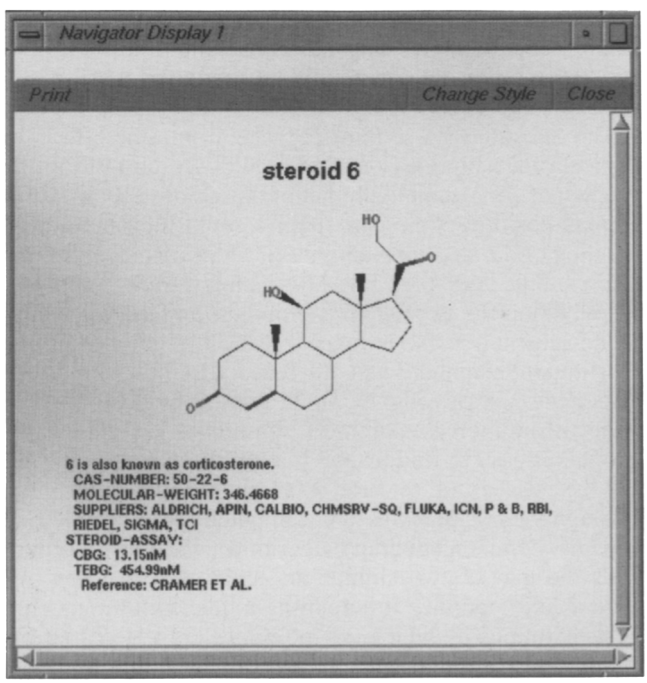

Figure 3. Display window for molecule **6**, *shown in 2D style.*

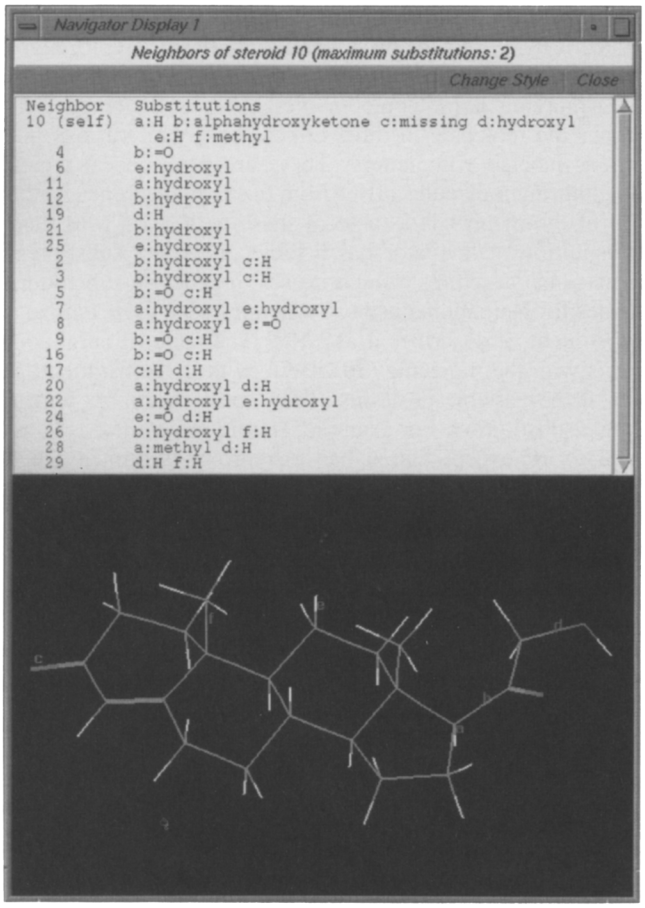

Figure 4. The Neighbors display for molecule ***10****. Substitution positions are labeled with letters; neighbors of* ***10****, with their substitutions, are listed in the table at the top.*

once, and each substitution position potentially many times, in the Substitutions listing all the neighboring molecules with substitutions at a particular position are collected together in a row.

The Substitutions style offers insight into the effects of particular substitutions on a molecule. By comparing the potency of a molecule with that of its analogs substituted at various positions, one can attain a rapid informal understanding of the structure–activity relationships involved. To take a simple case, molecule **31** is identical to **30** except that **31** has a fluorine in place of a hydrogen. In fact, displaying **31** in Substitutions style reveals that **30** is its only single-substitution neighbor; and **30** binds CBG two orders of magnitude more strongly than **31** does. Obviously the fluorine substitution is extremely significant. A relationship like this is easy to find by eye in a set of 31 molecules, but may be missed in a set of 1 000 analyzed manually.

Neighbor relationships are computed using a modified maximal common subgraph algorithm.[7] Such an algorithm finds the part of two similar molecules that matches; we were able to optimize it by putting a bound on the number of substitutions by which the molecules are allowed to differ. Because this computation is still very expensive, it is applied to each molecule once only. When a molecule is added to the database, it is compared with all other molecules, and the result is saved on disk.

MOUSE-SENSITIVE CHARTS: MAKING PATTERNS VISIBLE

The Chart style for molecule set displays the structure of each molecule together with whichever properties the user wishes to display. A sample chart is shown in Figure 6. The name of each molecule appears beneath it, together with any synonyms. (Molecules can be referred to by any number of names; one name is taken as primary. We usually use numbers as primary names, for brevity.) The properties of each molecule appear just below its name. This provides a compact way of showing the most relevant information about a group of molecules. Charts are paginated for printing.

Charts are most useful when used in conjunction with the subsetting and sorting operations. One will rarely want a chart of a complete drug discovery project database, which frequently includes hundreds or thousands of molecules. Instead, one can first select the subset relevant to answering a particular structure–activity relationship question. Sorting the subset by the assay of interest and charting the result will often lead directly to insight into the structure–activity relationship.[8] For example, in Figure 6, observe that all of the high-activity molecules have a carbonyl at the 3-position (left end of the molecule), whereas the low-activity ones have a hydroxyl. Examining those with high activity, note that the α-hydroxy-ketone substitution at the 17β-position (right end) provides the best activity, with smaller substitutions less favored. Other patterns like this emerge rapidly, especially when the chart is used in conjunction with Neighbors displays.

The two rules we derived in just a few seconds of Navigator analysis account for the steriod structure–activity data substantially better than do 2D QSAR methods.[3] (Three-dimensional methods can provide a more detailed and quantitative explanation of these molecules than such rules, however.)

Navigator charts are not simply passive displays; they are mouse sensitive. Right-clicking on a molecule in a chart produces a new display window showing that molecule and all its associated textual information. Left-clicking on a molecule causes it and all its neighbors to be boxed. The pattern of boxes in the chart allows the user to correlate structure with activity rapidly. In Figure 6, the user has clicked on molecule **6,** thereby causing all its neighbors to be boxed. Not surprisingly, the structural relatives of mol-

Bond	Mol 10	Substitutions on neighbors
a	H	hydroxyl: 11 7 8 20 22 methyl: 28
b	alphahydroxyketone	hydroxyl: 12 21 2 3 26 =O: 4 5 9 16
c	missing	H: 2 3 5 9 16 17
d	hydroxyl	H: 19 17 20 24 28 29
e	H	hydroxyl: 6 25 7 22 =O: 8 24
f	methyl	H: 26 29

Figure 5. Table of substitutions for molecule ***10****. The graphical portion of the display is identical to that in Figure 4.*

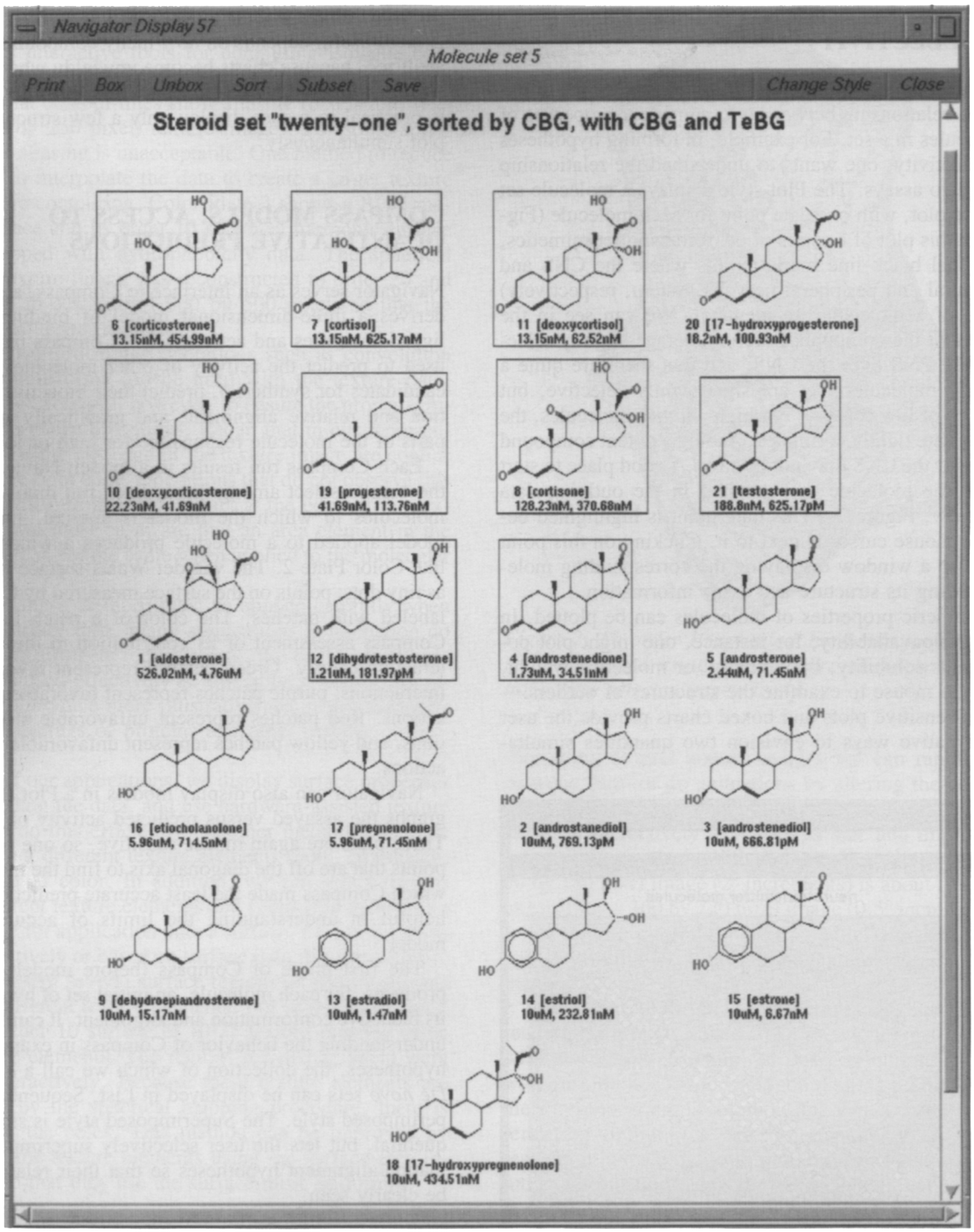

Figure 6. Chart of steriod molecules, with CBG- and TeBG-binding assays, sorted by CBG assay. Neighbors of molecule **6** are boxed.

ecule **6** include most of the most active molecules and none of the least active ones.

More generally, clicking on the Box button on a chart allows the user to specify an arbitrary subset of the molecules to box. For example, one might want to box all those that are highly active in a different assay from that by which the chart is sorted, to obtain intuition into the relationship between the two assays; or box all those that are highly flexible, when examining entropic considerations. The user can also box or unbox a single molecule by middle-clicking on it. This can be used to select by hand task-relevant molecules from a chart; the subset operation then collects and displays only these molecules. The subset can be saved in a disk file.

Navigator charts are conceptually similar to the molecular spreadsheet feature of other databases, but make more effective use of screen space, so that many more molecules can be viewed at once, and provide new functionality, in the boxing and mouse-sensitivity mechanisms.

MOUSE-SENSITIVE PLOTS: INSIGHT INTO SELECTIVITY

In structure–activity studies, it is frequently useful to examine the relationship between two numerical properties of the molecules in a set. For example, in forming hypotheses about selectivity, one wants to understand the relationship between two assays. The Plot style displays a molecule set as a scatterplot, with one data point for each molecule (Figure 7). In this plot of a series of neurotransmitter mimetics, the diagonal black line marks points where the CNS and PNS (central and peripheral nervous system, respectively) activities of a molecule are identical. We can see in the example that the compounds are on average slightly selective for the PNS over the CNS, and that there are quite a number of molecules that are significantly selective, but they are all of low activity. For high-activity molecules, the two assays are tightly correlated; finding a potent compound selective for the CNS may not be easy. A good place to start might be the molecule that resulted in the outlying data point (arrow, Figure 7). This data point is highlighted because the mouse cursor is next to it. Clicking on this point will pop up a window displaying the corresponding molecule, showing its structure and assay information.

Any numeric properties of molecules can be plotted. In exploring bioavailability, for instance, one might plot potency versus solubility, lipophilicity, or molecular weight, and use the mouse to examine the structures of outliers.

Mouse-sensitive plots and boxed charts provide the user with alternative ways to envision two quantities simultaneously, and the associated structures spatially. Which style is more useful depends on how many compounds are being analyzed, because charts become unwieldy when large; and on how many of the structures are truly relevant, because it is practical to mouse-display only a few structures from a plot simultaneously.

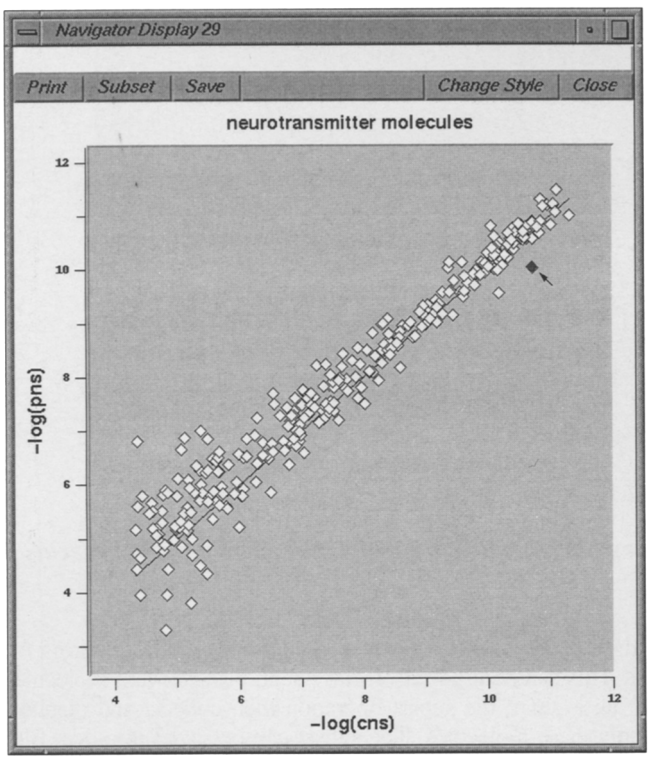

Figure 7. Plot of CNS versus PNS activities of a set of neurotransmitter ligands. One point (by arrow) has been selected with the mouse, and the corresponding molecule can be displayed by mouse clicking.

COMPASS MODELS: ACCESS TO QUANTITATIVE PREDICTIONS

Navigator serves as an interface to Compass, a system that derives a three-dimensional model of binding based on ligand structures and activities.[1] A Compass model can be used to predict the activity of other molecules (typically, candidates for synthesis), predict their bioactive conformation and relative alignment, and graphically indicate the parts of the molecule responsible for high or low activity.

Each Compass run results in a model; Navigator allows the user to select among these. Each run involves a set of molecules to which the model is applied. Displaying a model applied to a molecule produces a window looking like Color Plate 2. The van der Waals surface is displayed as tiny dots; points on the surface measured by Compass are labeled with patches. The color of a patch indicates the Compass assessment of its contribution to the binding interaction energy. Green patches represent favorable steric interactions; purple patches represent favorable polar interactions. Red patches represent unfavorable steric interactions, and yellow patches represent unfavorable polar interactions.

Navigator can also display models in a Plot style, which graphs the assayed versus predicted activity of molecules. These plots are again mouse sensitive, so one can click on points that are off the diagonal axis to find the molecules for which Compass made the least accurate predictions. This is helpful in understanding the limits of accuracy of the model.

The first phase of Compass (before model generation) produces, for each molecule, an initial set of hypotheses for its bioactive conformation and alignment. It can be useful in understanding the behavior of Compass in examining these hypotheses, the collection of which we call a *de novo set*. *De novo* sets can be displayed in List, Sequential, and Superimposed style. The Superimposed style is similar to Sequential, but lets the user selectively superimpose conformation/alignment hypotheses so that their relationship can be clearly seen.

IMPLEMENTATION

The Navigator database and all operations that manipulate molecular structure are implemented in Lucid Common Lisp (as are large parts of Compass, which shares the database). The database has facilities for importing and exporting chemical information in most of the common formats, including those of Biosym, ChemOffice, ISIS/Base, MacroModel, and Sybyl, facilitating the application of Navigator to existing databases. Structures and associated text and numeric information can also be entered by hand. Structures are stored in the format in which they are im-

ported or entered, and converted to an in-memory internal representation when referenced. The number of molecules stored, and the amount of information associated with each, are limited only by available disk space.

Navigator runs on Silicon Graphics workstations under Motif. Most of the user interface and graphics code is written in Tcl/Tk.[9] Tk supplies many graphical user interface tools that simplified the implementation, and Tcl provides an interpreted, dynamic environment for rapid prototyping. We used the Tk "canvas" abstraction for all two-dimensional displays; this enabled mouse sensitivity and provided Postscript generation for hardcopy. Three-dimensional displays were implemented in C with calls to the SGI Graphics Library (GL). The integration of Lisp, Tcl/Tk, and C/GL, using the Lucid foreign function interface, was straightforward. It allowed us to combine their different strengths synergistically, and saved considerable effort over what would have been required with any one of the systems alone.

CONCLUSION

Navigator seamlessly integrates a variety of new tools that can help medicinal chemists discover structure–activity relationships. These include ways of displaying the relationships between substituted molecules, charts displaying sets of molecules together with their assays, mouse-sensitive plots of assay relationships, and access to Compass, a system for three-dimensional quantitative structure–activity relationship discovery.

ACKNOWLEDGMENTS

We thank Mike Ross and anonymous Arris internal reviewers for helpful criticism of the manuscript, and Barr Bauer, Ajay Jain, Tom Jenkins, Brad Katz, Teri Klein, Elaine Kuo, Mike Ross, Laura Whitman, Jan Vågberg, and Peng Zhou, our test users, for enduring prerelease bugs and making many useful suggestions for Navigator functionality.

REFERENCES

1 One such tool is Compass (Jain, A.N., Koile, K., and Chapman, D. Compass: Predicting biological activities from molecular surface properties. *J. Med. Chem.* 1994, **37,** 2315–2327), described further in this article. For a review of others, see Kubinyi, H. (ed.). *3D QSAR in Drug Design*. ESCOM Science Publishers, Leiden, 1993

2 ChemFinder is a product of Cambridge Scientific Computing, Inc.; Daylight is a product of Daylight Chemical Information Systems, Inc., Irvine, California; ISIS/Base is a product of MDL Information Systems, Inc.; UNITY is a product of Tripos, Inc.

3 Jain, A.N., Koile, K., and Chapman, D. Compass: Predicting biological activities from molecular surface properties. *J. Med. Chem.* 1994, **37,** 2315–2327; Cramer, R.D., Patterson, D.E., and Bunce, J.D. Comparative molecular field analysis (CoMFA). *J. Am. Chem. Soc.* 1988, **110,** 5959–5967; Good, A.C., So, S., and Richards, W.G. Structure–activity relationships from molecular similarity matrices. *J. Med. Chem.* 1993, **36,** 433–438

4 *Daylight Theory Manual,* version 4.40b. Daylight Chemical Information Systems, Inc., Irvine, California, 1995

5 The ACD database is a product of MDL Information Systems, Inc.

6 Fisanick, W., Lipkus, A.H., and Rusinko, A. Similarity searching on CAS Registry substances. 2. 2D structural similarity. *J. Chem. Inf. Comput. Sci.* 1994, **34,** 130–140; Hagadone, T.R. Molecular substructure similarity searching: Efficient retrieval in two-dimensional structure databases. *J. Chem. Inf. Comput. Sci.* 1992, **32,** 515–521; see also the special issue of *J. Chem. Inf. Comput. Sci.*, 1992 **32**(6), collecting papers presented at the May 1992 Workshop on Similarity in Organic Chemistry

7 Chen, L., and Robien, W. MCSS: A new algorithm for perception of maximal common substructures and its application to NMR spectral studies. *J. Chem. Inf. Comput. Sci.* 1992, **32,** 501–506; Bayada, D.M., Simpson, R.W., Johnson, A.P., and Laurenço, C. An algorithm for the multiple common subgraph problem. *J. Chem. Inf. Comput. Sci.* 1992, **32,** 680–685

8 We are indebted to Rick Lathrop, who showed us this technique

9 Ousterhout, J. *Tcl and the Tk Toolkit*. Addison-Wesley, Reading, Massachusetts, 1994