MAMBAs: A real-time graphics environment for QSAR

Frank E. Blaney,* Dorica Naylor,† and John Woods†

SmithKline Beecham Pharmaceuticals, Medicinal Research Centre, Essex, UK; †Oxford Molecular Ltd., The Magdalen Centre, Oxford, UK

MAMBAs (Multivariate Analysis Methods in Biomechanistic Activity Studies) is an integrated workstation-based graphics program designed for the investigation of quantitative structure activity relationships (QSAR). It combines many of the commonly used statistical techniques with an extensive database of substituent constants, a variety of molecular and substituent property calculations and detailed graphics-based table and graph editors. Graphical representations of standard substituent generation and optimization techniques are also included. These are all utilized within a state-of-the-art real-time graphics environment.

Keywords: QSAR, molecular graphics, multivariate statistical analysis

INTRODUCTION

It has often been said[1] that the father of medicinal chemistry was Paul Ehrlich. At the beginning of this century, it was he who stated that the biological action of active medicinal compounds was due to some complementary properties of the compound and its site of action (receptor)[2] and that the two fitted together as "parts of a mosaic." The importance of one such property, lipophilicity, was recognized as early as the nineteenth century, when Overton[3] and Meyer[4] independently showed a relationship between narcotic potency and the oil–water partition coefficient. Albert[5] over many years showed that there was a quantitative relationship between a variety of measured chemical properties and biological activity. Modern QSAR, however, owes most to the pioneering work of Corwin Hansch and his coworkers.[6] As early as 1963, he was quantifying the proposals of Ehrlich, by attempting to relate the biological activity of a series of similar compounds to their physicochemical properties. He did this by expanding on the linear free-energy (LFE) concepts first developed by Hammett.[7] Hansch was also the first to study the prediction of hydrophobicity[8] and the nonlinear relationship between this property and drug penetration and delivery. Hansch used the LFE substituent constants for electronic, steric and hydrophobic effects within a multivariate statistical framework to describe the biological activities of compounds. It was the applications of this technique which subsequently became known as quantitative structure activity relationships (QSARs).

Traditionally, the classical techniques of QSAR have been considered separately from molecular modeling. In drug or pesticide design, however, the ideal endpoint of many modeling studies would be a predictive QSAR equation. One of the reasons for this division has been that developments in the modeling field, over the last ten years in particular, have made full use of the advances in real-time graphics workstation technology. Cramer[9] has already demonstrated with the COMFA program[10] how one multivariate statistical technique, PLS,[11] could be used for drug design within a graphical environment. It was felt that "classical" QSAR could gain much from the use of computer graphics and that this would help to break down some of the barriers between the practitioners of QSAR and molecular modeling. It would also help to introduce the methods of QSAR to the medicinal chemist in a more user-friendly and familiar framework.

A new suite of computer programs known as MAMBAs (Multivariate Analysis Methods in Biomechanistic Activity Studies) was therefore developed which combined many of the standard statistical techniques used in QSAR, within a real-time graphical environment. The program includes an extensive substituent database and semiautomated calculation of substituent and molecular properties. A spreadsheet style table is used which can include both two-dimensional (2D) and three-dimensional (3D) graphical structures, text and numerical data. Once the data tables are set up, the techniques of multiple regression, cluster analysis, factor analysis, principal components analysis, nonlinear mapping and PLS, etc., can be readily applied. The tables and the results from the statistical analyses can be manipulated in real time. There is a powerful graph plotting function which enables the user to study plots in two, three, four or even five dimensions. Mathematical functions of table columns can be defined. It is also possible to input the results from some external calculations, such as output from the standard quantum mechanical programs, CNDO,[12] MOPAC[13] and Gaussian 90.[14] For those who wish, it is also possible to create data for the statistical suite of programs GENSTAT.[15] Although it is not currently possible, hardcopy is recognized

Color Plates for this article are on pages 185–186.

Address reprint requests to Dr. Blaney at SmithKline Beecham Pharmaceuticals, Medicinal Research Centre, Coldharbour Rd., The Pinnacles, Harlow, Essex, CM19 5AD, UK.

Received 28 December 1992; accepted 16 February 1993

as an important issue, and a postscript output is currently being developed.

PROGRAM LAYOUT

MAMBAs is written in FORTRAN 77, under the UNIX operating system and currently contains over 50,000 lines of code. It utilizes the GL graphics programming language[16] running under 4SIGHT[17] (a port to X-Windows/OSF Motif[18] is planned) and will run on most Silicon Graphics 4-D series workstations.

As with most modern workstation programs, MAMBAs is written in such a way that different major parts of the suite utilize different windows. As shown in Color Plate 1, upon start-up, three main areas appear on the screen. The window on the right contains a menu and dials which are intended primarily for the manipulation of molecules and their associated properties. The menu items which appear here are the main control features of the program. These are shown in Figure 1. The bottom area of the screen contains a text port. The remainder of the screen is taken up with the main working window, which may contain molecules, tables, graphs or the associated 2D drawing package. As already mentioned, the dials are meant primarily for the manipulation of molecules and their associated properties. While the program is not in any way to be considered as another modeling package, such simple manipulations are thought to be essential in the calculation and display of physicochemical factors. For 3D molecular displays, rotation and scaling with the dials is either continuous or by fixed increments. In other modes, e.g., 2D graphs or tables, certain rotations (continuous and fixed increment) or translations are not applicable. In these cases the appropriate dials will not be visible.

Further control and setup for the various options is carried out either with the right dials and menu window, which changes for many of the options, with dialogue boxes or with pull-down menus. (Color Plate 1 shows many of these features.)

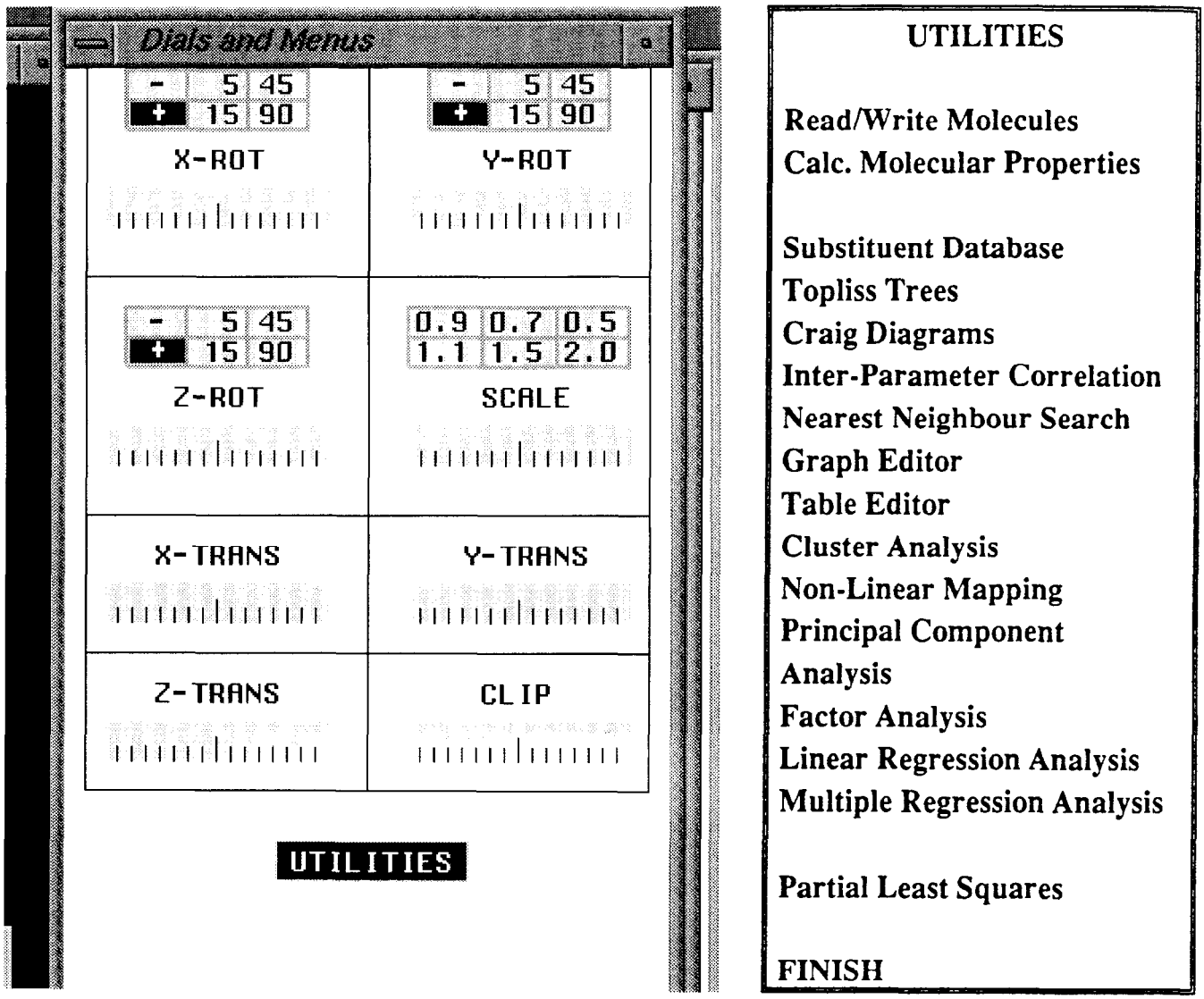

Figure 1. Dials and main start-up menu for MAMBAs.

UTILITIES

The utilities functions in MAMBAs are intended primarily to alter certain settings which under normal conditions are fixed. The most important such features are the sensitivity of the dials and the definition of the colors. These settings are under interactive control. One can also change the level of output to the text port, although this function was originally intended mainly for debugging purposes. The other settings which can be changed are mainly involved with text and data attributes. These include features such as font size, text color, the number of decimal places in the data columns, the maximum number of characters and lines allowed in text cells, etc. This list is constantly being added to as needs arise or as new features are added.

READING AND WRITING MOLECULES

The use of this function is obvious. The program, however, does not have its own 3D builder so it has been provided with file formats for a number of different external programs. These include the standard CSSR, PDB and Z-Matrix (MOPAC and Gaussian) formats together with the file structures of the commercial packages licensed within the company. Another useful feature is the ability to read in structures from the output of some commonly used quantum mechanical packages, again notably MOPAC and Gaussian 90. Simple M.O. properties such as potential-derived charges and dipole moments are then available for use in table generation.

QSAR has traditionally been associated with small-molecule studies. However, the dimensions of the program are such that proteins can be readily read in and displayed. The main purpose here is to display some commonly associated QSAR properties such as lipophilicity, electrostatic potential and molar refractivity on the protein surface and to highlight areas of potential interaction. No property or statistical calculations are currently carried out on protein structures.

PROPERTY CALCULATIONS

The main purpose of the MAMBAs Read/Write Molecules facility is to enable the user to perform a variety of molecular and substituent property calculations. Under the normal QSAR framework, properties can be divided into bulk properties of the whole molecule (eg., LogP, pKa, ionization potential, shape, molecular weight, volume and surface area, etc.) and substituent properties which are normally either electronic, hydrophobic or shape/steric in nature. The properties which can currently be calculated are summarized in Figure 2.

Molar refractivity calculations are based on the additivity principles described by Hansch et al.[19] Similar methods have been used in the literature for the estimation of LogP, most notably those of Nys and Rekker[20] and Hansch et al. The latter program[21] has been commercially available for several years and is widely used throughout the world. For this reason no attempt has been made within MAMBAs to include hydrophobic calculations. Similarly, although many published electronic substituent parameters are included in the program's internal database, it was felt that molecular electronic properties were best derived from the application of commonly available quantum mechanical programs. Simple interfaces for CNDO, MOPAC and Gaussian 90 have therefore been included and properties from these calculations can be displayed or calculated from the output. The properties currently available are shown in Figure 3.

Another important feature is the ability to align molecules in a meaningful way, so that property calculations and comparisons can be made in the same reference frame. This is particularly true of quantum mechanical properties such as dipole moment vector components. These alignments can either be of the least-squares fitting type or can be based on the steric-electrostatic alignment (SEA) method of Smith.[22]

The final group of properties which can be calculated may be broadly thought of as "geometric" in nature. Obvious ones here include molecular or substituent surface area and volume, and the usual intramolecular measurements of distances, angles and dihedrals. Various "dummy atoms," such as ring centroids and perpendiculars, may be defined

Molecular and Substituent Properties	
1:	Molecular Weight
2:	Surface Area and Volume
3:	Molar Refractivity
4:	Verloop Sterimol Parameters
5:	Moments of Inertia
6:	Principal Ellipsoids
7:	Internal Coordinates, distances etc. incl. dummy variables.
8:	Quantum Mechanical Properties

Figure 2. The current properties calculated in MAMBAs.

Molecule Based Electronic Properties	Atom Based Electronic Properties
Dipole Moment & Components	Mulliken Atomic Charges
Total Energy	Potential-derived Atomic Charges
Orbital Energies (HOMO & LUMO)	Atomic Eigenvectors
Ionization Potential	Superdelocalizability (where applicable)
Electrostatic Potential	

Figure 3. Electronic properties displayed or calculated from M.O. output.

and used in these calculations. A second important group here are the substituent steric parameters. Classically, these have presented the greatest difficulty for QSAR descriptions. The earliest substituent constant used was the Taft steric parameter E_S, which was derived[23] from experimental studies on the rates of hydrolysis of acyl substituted methyl esters relative to that of methyl acetate. All the commonly used E_S values are included in the internal substituent database. Hancock et al.[24] suggested a correction to the Taft parameter to take account of hyperconjugation effects. This takes the form of

$$E_S^c = E_S + 0.306(n - 3)$$

where n is the number of alpha hydrogen atoms. These corrected E_S values are also included in the database. An evident problem with this type of parameter, however, is the lack of experimental data. For spherically symmetrical substituents, a correlation was found between E_S and the substituent radius, and this property has been used to derive further values of E_S. However the most useful approach to the development of a novel nonexperimental steric descriptor has been that of Verloop et al.,[25] who have used one length and four width parameters to describe the geometric dimensions of the substituent. They have reported values for some 243 substituents. The Verloop method has been programmed into MAMBAs so that the user can calculate additional parameters for substituents not reported in the literature. Another descriptor which we have found to be useful as a steric parameter is the substituent principal ellipsoid (SPE). This gets around the problem that most substituents are not spherical, by describing them as ellipsoids, defined by their three principal axes. These, in turn, come from calculations of the moments of inertia of the substituents. Both Verloop parameters and SPEs can be displayed with the program and manipulated in real time. Color Plate 2 shows a molecule with a Verloop sterimol calculation (2a) and a substituent principal ellipsoid calculation (2b) portrayed.

SUBSTITUENT DATABASE

The substituent database was initially based on the published work of Hansch and Leo[26] and contains some 300 commonly found aromatic and aliphatic substituent values. Thirteen properties are currently contained within the database. These are Pi, MR, Swain and Lupton's F and R, Sigma Para, Sigma Meta, Sigma*, Taft Es and the Verloop Sterimol parameters L and B1–B4. It is an easy task, however, to add new substituents and properties to the database or to customize it to the user's requirements. Subsets of this database, e.g., "All Aliphatic" or "Well-Characterized[27] Aromatic" can be searched individually.

The database can be queried either by entering a substituent name or by graphically scanning through the whole file, which appears as a table on the screen. The former method is faster but often fails because of unforeseeable factors such as spelling mistakes! For this reason, entry can either be by name or formula. Common pseudonyms for substituents are also included; hence searching for parameters for the phenyl group may be done by typing *phenyl, c6h5* or even *ph*. If the user wishes to inquire about substituents with a given set of values, this is easily carried out. Queries can also be made about the nearest neighbors (in Euclidean distance) to a particular substituent, based upon any given set of parameters. Alternatively, related families of substituents (again based on a predefined set of parameters) can be examined using hierarchical cluster analysis techniques. As with any other data set, standardization of the data is necessary, although the option is provided to use nonstandardized data. Two methods are provided to do this,[28] viz., standardization either by range or by calculating the distance from the mean divided by the standard deviation. The use of Euclidean distance in such searches is based on our normal bias of thinking in 3D space. However, as Martin has pointed out,[28] there is no reason in higher dimensional hyperspace not to use alternative measurements such as the cube root of cubes. Some of these are provided in MAMBAs, although the results do sometimes look strange!

When synthesizing a series of compounds for biological evaluation, the optimal choice of substituents depends upon fully exploring the hyperspace defined by the parameter set of interest. Standard *lead optimization* methods for this include the well known Topliss Tree[29] and Craig Diagram[30] methods. Graphical representations of both these methods can be obtained from the MAMBAs program. The latter has been extended so that the plots of any two parameters within the database can be examined. Furthermore, by using the "Interparameter Correlation" option on the main menu, plots of 3, 4 or even 5 parameters can be displayed as 3D graphs using variable color ramps and multiple symbols for the fourth and fifth values. (See Color Plate 3.) These graphs can, of course, be manipulated in real time. For example, one can pick a point on the graph and information will immediately be displayed about the name of the substituent, the values of associated parameters and those of additional parameters not included in the plot. All these graphs are under the control of the MAMBAs Graph Editor.

GRAPH EDITOR

In addition to the database plots described above, the graph plotting and manipulation routines are general for any data set and form a valuable integral part of the program. The purpose of the graph editor is to provide the user with complete control over the appearance of the display. This includes such features as notching and labeling of the axes, symbols, fonts, color, labeling of points and, of course, rotation, translation and scaling of the graphs. Axes can be instantaneously swapped, a useful feature which is often missing from other graph plotting programs. Single or multiple lines (or curves) can be added to the graphs. Information for these lines can come from such sources as linear regression, curve or contour-fitting routines, Cardinal or B-Spline fits, etc. The Graph Editor menu is shown in Figure 4.

TABLE EDITOR

A key feature of the program is the table editor. This enables the user to generate complete tables which can include 2D or 3D graphical structures, multiline text descriptions, and row and column editing. With the editor, rows and columns can be readily added, deleted or swapped around. In addition, data columns can be optionally defined as dependent, when

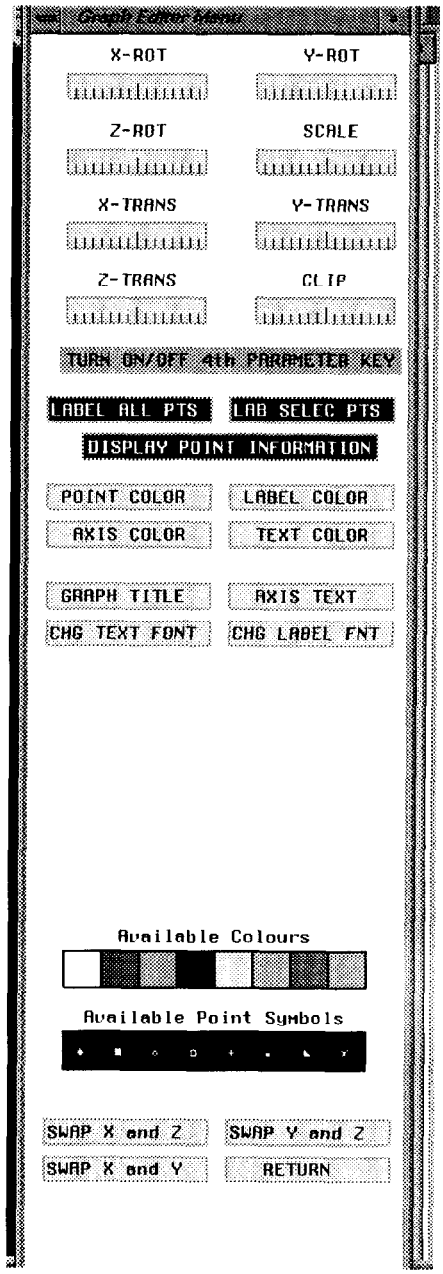

Figure 4. Graph editor menu.

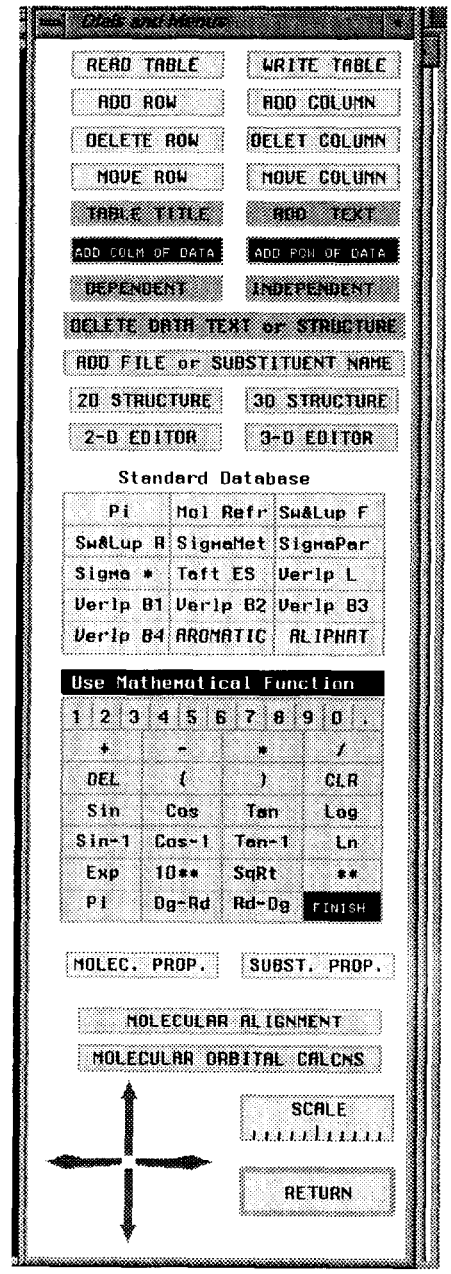

Figure 5. Table editor menu.

appropriate for the subsequent statistical methods of analysis to be employed. Rows and columns can also be temporarily excluded from the subsequent statistical analysis.

Figure 5 shows the main menu which controls the generation and manipulation of the table. The table itself is in the form of a molecular spreadsheet, which can contain data, text and structures. As described above, the graphical structures can be either 2D or 3D. At present, there is only one column of structures allowed in the table, although no problem is envisaged in expanding this to two or three. One can also have a generic structure which appears alongside the title at the top of the table. This could then, for example, contain an R- group, and the structures of the R group would appear in the table itself. Figure 6 shows a typical MAMBAs data table. If one is setting up a table of a series of related compounds where variation occurs only in the substituents, then if the parameters required are stored in the database, it can be automatically searched and the appropriate parameter values entered. To do this, the appropriate parameter is picked from the table editor menu and added to a blank column in the table. An appropriate column title will automatically appear. Entry of substituents is facilitated by means of a scrolling list of all those present within the database. As these are picked off, they are automatically entered into the table. (Color Plate 4)

The structures in the data table are saved in the data table file as a set of filenames which point to externally stored files. It was decided to construct the table in this way simply

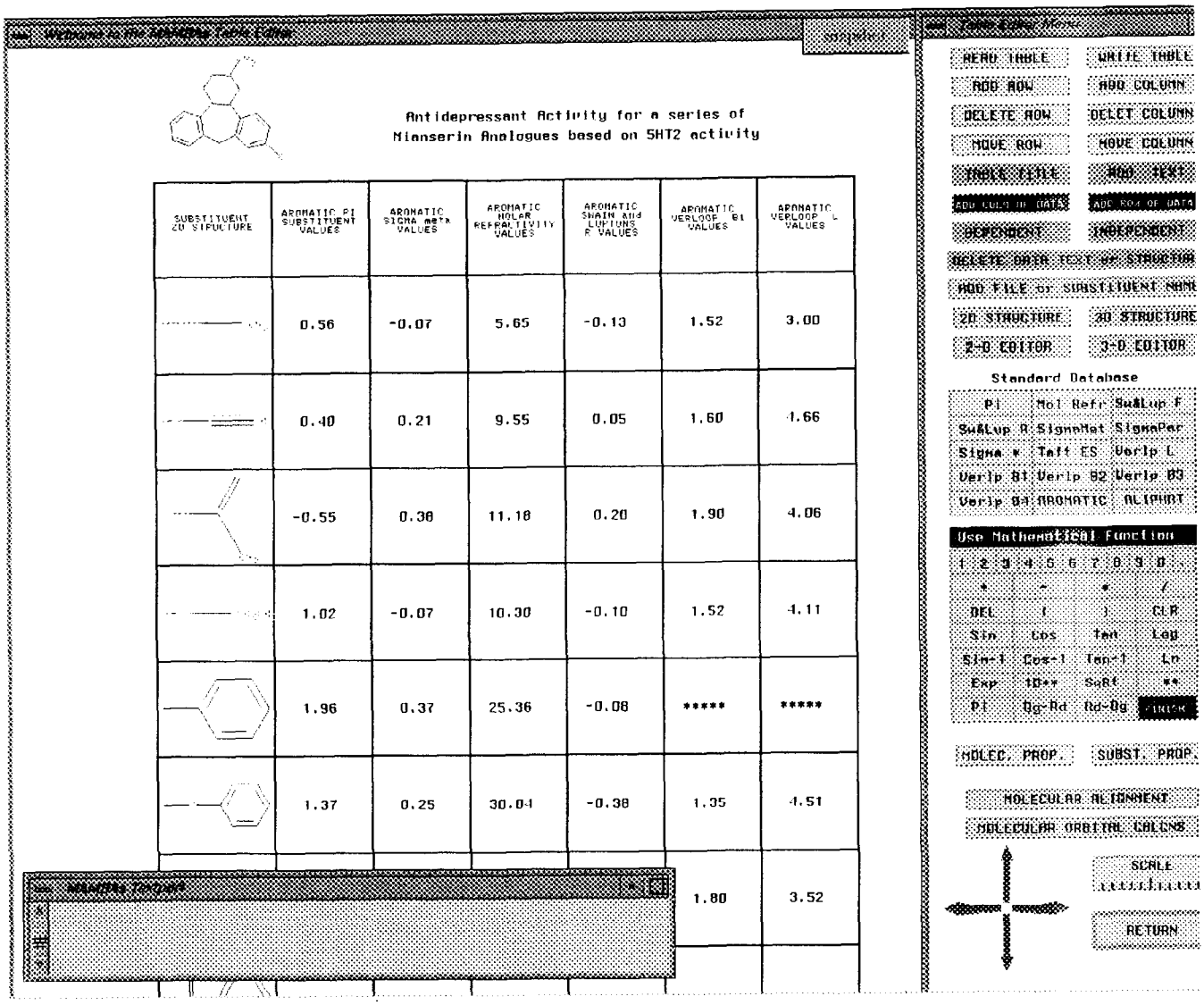

Figure 6. A typical MAMBAs data table with title and structures.

to save space. Also stored in the data table file is an associated 3 × 3 rotation matrix for each filename. This matrix contains information about the orientation of the structure in the data cell. An important feature of the table editor is the ability to interactively examine the structures in the table. Both 2D and 3D structures can be expanded into their own windows and manipulated in real time. Using this feature, it is possible to alter the view of the 2D or 3D structures within the table data cells, to show or hide the hydrogens and to calculate the simple properties 1–7 outlined in Figure 2. These are then automatically entered into the appropriate data column. To facilitate the handling of 2D information, MAMBAs comes complete with its own 2D structure drawing feature, which is shown in outline in Figure 7.

Evidently for all but the simplest data sets, the table will extend beyond the boundaries of the screen. Real-time manipulation of the table is therefore provided. This is achieved with the arrows, which allow X-Y translation, and the scale bar.

The final feature of the Table Editor worthy of mention is the mathematical function generator. This allows for the definition of the values of a column in terms of some mathematical function. The function can consist of any combination of variables, from other columns or from the standard database, with mathematical definitions from the table editor. A simple example might be

(COL 9) = (Aromatic − Pi)**2

This defines the value of each row in Column 9 as the square of the aromatic substituent Pi value of that row, as obtained from the database. All the equations start with a (blank) column value on the left of the equation, followed by an *equals* sign. The mathematical function parser first checks on the levels of parentheses and evaluates everything from the innermost set outwards. The other rule used in the parser is that the conventional hierarchy

** >> */ >> + −

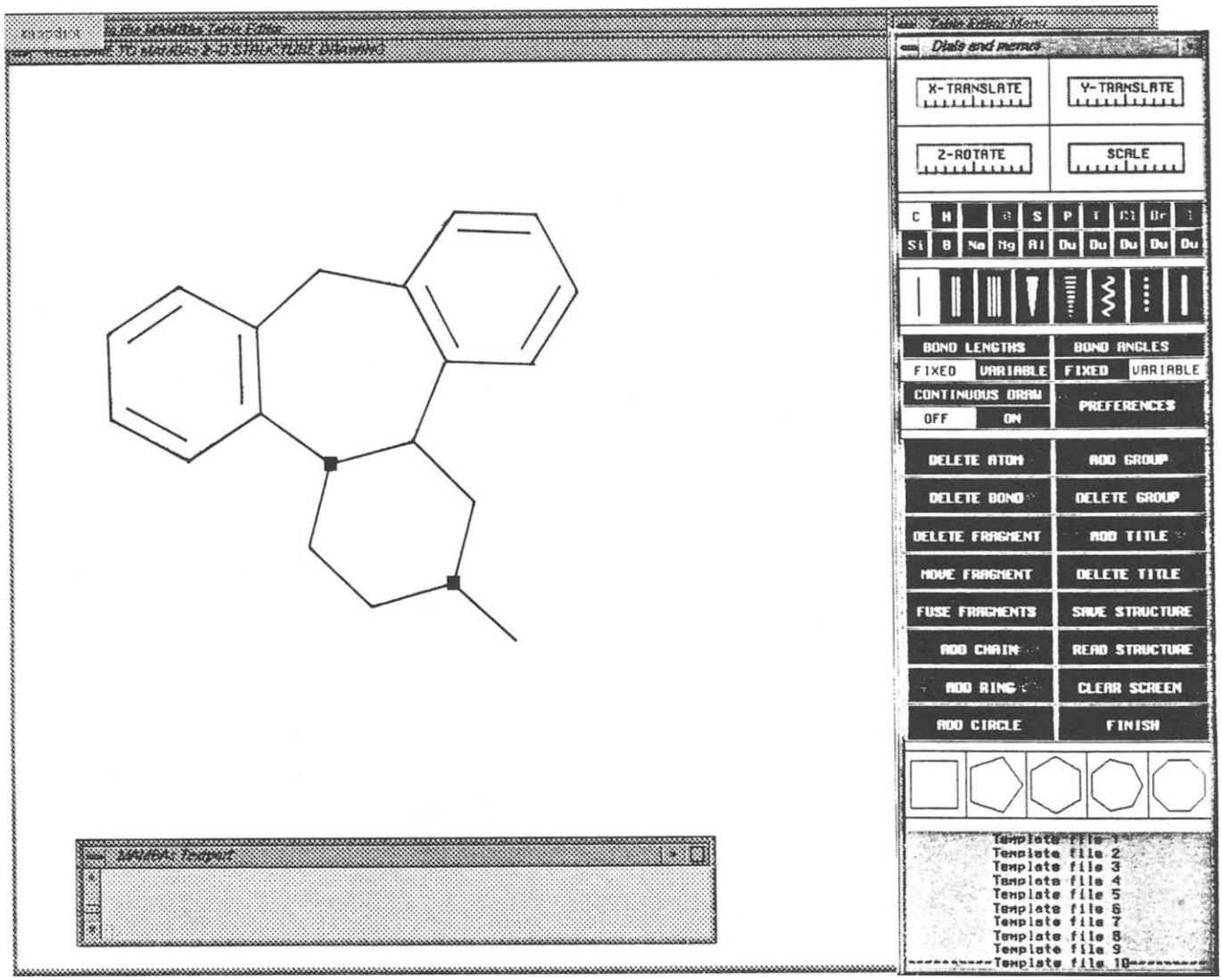

Figure 7. MAMBAs 2D structure drawing option.

is followed. It is very important to ensure that all the parentheses are included and that they balance, as the error checking is not always foolproof! It is still recommended that as many sets of parentheses as possible are used to make the expression completely unambiguous. Another example of an expression is shown below

(COL 9) = (3.5*(MR)) + (SQRT(COL 5)) +
((Aromatic − Pi)**2)

The interpretation of this expression is evident. Note, however, the liberal use of parentheses to ensure that the individual components of the expression are all evaluated before the final summation. All these expressions can either be entered at the keyboard or more simply defined using the mouse.

STATISTICAL ANALYSIS

Having set up the data table, the user can then go on and perform the statistical analysis on it. A variety of standard statistical analysis methods are therefore provided. These methods have been described in great detail in the literature,[31] and standard algorithms have been used. The statistical procedures provided in MAMBAs are single and multiple linear regression, factor analysis, principal component analysis, PLS[32], nonlinear mapping and a variety of cluster analysis[33,34] algorithms. Each of these methods has its own setup menu. For example, the menus for multiple regression, factor analysis and principal component analysis are shown in Figure 8. The top portion of each menu contains features which control the setup of the data table itself. These include the exclusion or inclusion of rows and columns, the choice of the data standardization method and the definition of the dependent data column where applicable, etc. Another feature present in all the statistical method menus, is the ability to plot any set of data columns, prior to analysis, using the standard graph editor.

The output from the statistical calculations is displayed as a series of graphical objects which can be fully manipulated in real time. Thus looking at the multiple regression menu in Figure 8, it can be seen that the correlation matrix, the analysis of variance and the standard regression information

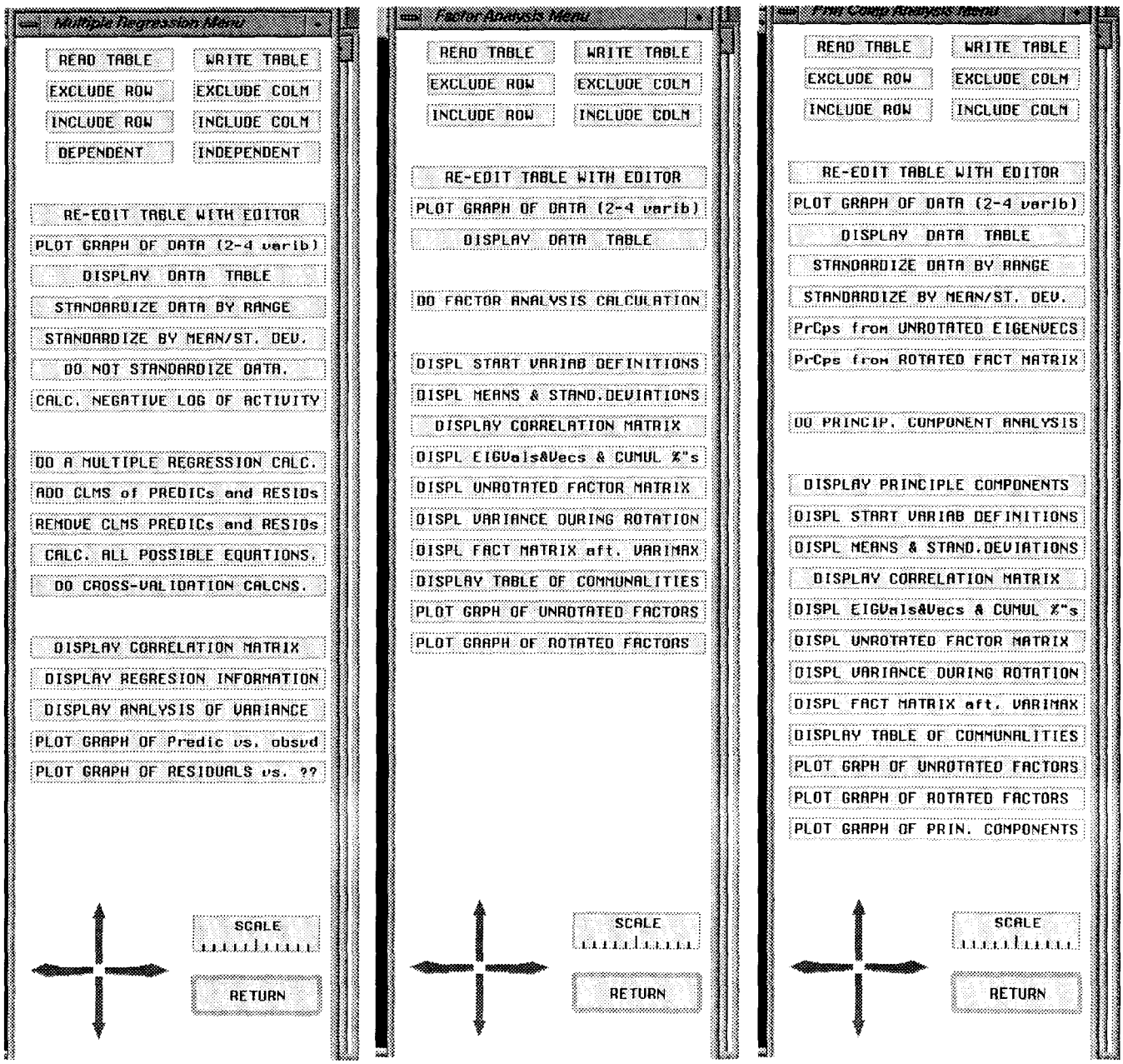

Figure 8. Menus for multiple regression, factor and principal component analysis.

(i.e., the regression equation, correlation coefficient, F-values, etc.) may all be viewed as seperate graphical objects and can be instantly interchanged on the screen. Many other utilities are also included. For example, in multiple regression analysis, an automatic graph of predicted versus observed data is obtained. The columns of predicted and residual values from the calculation can be added to the original data table and used in subsequent graph plots. The program will also calculate the set of regression equations for all possible combinations of up to six parameters and rank these according to their correlation coefficients. Finally, as with many statistical methods, cross-validation may be used to check the validity of the data. With methods such as factor analysis, PCA and nonlinear mapping, plots of the outputted factors can be obtained, or alternatively, one can perform cluster analysis on them.

The choice of clustering method to be used is dependent on many factors and has been the subject of discussion in many papers and books,[31,33,34] and Bawden[35] has given a comparison of some of these as applied to chemical problems, using his DISCLOSE program. As different methods are well known to produce quite different results on the same data set, a variety of clustering algorithms are provided. These are single-link hierarchical, minimum spanning tree, furthest neighbor, group average and binary splitting. The display of clustering is of the standard dendrogram type. Alternatively, individual clusters can be highlighted in the 2D or 3D graph displays.

During the statistical calculations in any one MAMBAs session, a continuous log of the input and outputted results is maintained. This can be printed out afterwards as hardcopy.

CONCLUSION

Many large statistical packages exist nowadays for the practice of QSAR calculations. To our knowledge, however, MAMBAs provides the first workstation-based package which fully integrates the techniques of real-time molecular modeling with the classical parameter calculations and statistical methods used in QSAR.

REFERENCES

1 Ariens, E. *Molecular Pharmacology*. Academic Press, London, 1964, vol. 1, p. *xiv*
2 *The Collected Papers of Paul Ehrlich*. Pergamon Press, London, 1956, vol. 1
3 Overton, E. *Zeitschrift fur Physikalisch Chemie*. 1897, **22,** 189
4 Meyer, H. *Archiv fur Experimentell Pathologie und Pharmakologie*. 1899, **42,** 109
5 Albert, A. *Selective Toxicity*. Chapman and Hall, London, 1973
6 Hansch, C., Muir, R.M., Fujita, T., Maloney, P.P., Geiger, F., and Streich, M. *J. Am. Chem. Soc.* 1963, **85,** 2817
7 Hammett, L.P. *Physical Organic Chemistry*. McGraw-Hill, New York, 1940
8 Fujita T., Iwasa, J., and Hansch, C. *J. Am. Chem. Soc.* 1964, **86,** 5175
9 Cramer III, R.D., Patterson, D.E., and Bunce, J.D. *J. Am. Chem. Soc.* 1988, **110,** 5959
10 COMFA is a registered trademark of TRIPOS Associates Inc.
11 Wold, S., Ruhe, A., Wold, H., and Dunn III, W.J. SIAM. *J. Sci. Stat. Comput.* 1984, **5,** 735
12 CNINDO. QCPE 141. Dept. of Chemistry, Indiana University, Bloomington, Indiana, 47405
13 Stewart, J.J.P. QCPE 455. Dept. of Chemistry, Indiana University, Bloomington, Indiana, 47405
14 Frisch, M.J., Head-Gordon, M., Trucks, G.W., Foresman, J.B., Schlegel, H.B., Raghavachari, K., Robb, M.A., Binkley, J.S., Gonzalez, C., Defrees, D.J., Fox, D.J., Whiteside, R.A., Seeger, R., Melius, C.F., Baker, J., Martin, R.L., Kahn, L.R., Stewart, J.J.P., Topiol, S., and Pople, J.A. Gaussian 90. Gaussian Inc., Pittsburgh, PA, 1990.
15 GENSTAT 5 is a widely used general purpose statistical package which has been developed by the Statistics Department of Rothampsted Experimental Station, and distributed by Numerical Algorithms Group Ltd., Mayfield House, 256 Banbury Rd., Oxford, OX2 7DE, UK
16 The Graphics Library (GL) programming language is a trademark of Silicon Graphics Inc., Mountain View, California
17 4Sight is an integrated windowing environment for IRIS workstations
18 The X-Window system is a trademark of the Massachusetts Institute of Technology. Motif is a trademark of the Open Software Foundation, Inc.
19 Hansch, C., Leo, A., Unger, S.H., Kim, K.H., Nikaitani, D., and Lien, E. *J. Med. Chem.* 1973, **16,** 1207
20 Nys, G.G., and Rekker, R.F. *Chimica Therapeutic.* 1973, **9,** 521
21 Medchem software is distributed by Daylight Chemical Information Systems Inc., Irvine, CA 92715, USA
22 Smith, G. QCPE 567. Dept. of Chemistry, Indiana University, Bloomington, Indiana, 47405, USA
23 Taft, R.W. in *Steric Effects in Organic Chemistry*. (M.S. Newman, Ed.) Wiley, New York, 1956, pp. 556–675
24 Hancock, C.K., Meyers, E.A., and Yager, B.J. *J. Am. Chem. Soc.*, 1961, **83,** 4211
25 Verloop, A., Hoogenstraaten, W., and Tipker, J. in *Drug Design*. (E.J. Ariens, Ed.) Academic Press, New York, 1976, vol. 7
26 Hansch, C., and Leo, A.J. *Substituent Constants for Correlation Analysis in Chemistry and Biology*. Wiley-Interscience, 1979
27 A "well-characterized data set" is defined in these circumstances as a data set in which all the information for each row and column is known. The data used here for well-characterized aromatic or aliphatic comes from the work of Hansch and Leo in Ref. 26 above
28 Martin, Y.C. *Quantitative Drug Design A Critical Introduction*. Marcel Dekker, Inc., New York, 1978, p. 235
29 Topliss, J.G., and Martin, Y.C. in *Drug Design*. (E.J. Ariens, Ed.) Academic Press, New York, 1975, vol. 5
30 Craig, P.N. *J. Med. Chem.* 1971, **14,** 680
31 Chatfield, C., and Collins, A.J. *Introduction to Multivariate Analysis*. Chapman and Hall, London, 1980
32 Helland, I.S. *Commun. Statist. Simula.* 1988, **17,** 581
33 Anderberg, M.R. *Cluster Analysis for Applications*. Academic Press Inc., New York, 1973
34 Hartigan, J.R. *Clustering Algorithms*. Wiley, New York, 1975
35 Bawden, D. *Anal. Chim. Acta.* 1984, **158,** 363