processing. It does appear, however, that the phrase coding and input procedure lends itself quite advantageously to on-line interactive processing. The procedure is presently being adapted for on-line testing.

ACKNOWLEDGMENT

CAS is pleased to acknowledge the partial support of this work from the National Science Foundation (Contract C656).

APPENDIX. ALGORITHM-ARTICULATED INDEX ENTRY SAMPLE

Listed in Table VI is a sampling of the index entries generated from coded phrases via the articulation algorithm during the course of the testing. The index entry ratings assigned by the document analysts are also shown. The purpose of this listing is only to illustrate the quality of text modifications generated by the algorithm. The headings in this listing are the *as-dictated* headings which have *not* been subjected to the CAS heading terminology control systems. Therefore, in some cases, the headings shown are not the headings which would appear in the CAS volume indexes.

See Table IV for meaning of ratings.

LITERATURE CITED

(1) J. E. Armitage and M. F. Lynch, "Articulation in the Generation of Subject Indexes by Computer", *J. Chem. Doc.*, **7**, 170–8 (1967).
(2) J. E. Armitage and M. F. Lynch, "Some Structural Characteristics of Articulated Subject Indexes", *Inf. Storage Retr.*, **4**, 101–11 (1968).
(3) R. D. Nelson, W. E. Hensel, D. N. Baron, and A. J. Beach, "Computer Editing of General Subject Heading Data for *Chemical Abstracts* Volume Indexes", *J. Chem. Inf. Comput. Sci.*, **15**, 85–94 (1975).
(4) R. J. Rowlett and F. A. Tate, "A Computer-Based System for Handling Chemical Nomenclature and Structural Representations", *J. Chem. Doc.*, **12**, 125–28 (1972).
(5) R. Salvador, "Automatic Abstracting and Indexing", Masters Thesis, The Ohio State University, Columbus, Ohio, 1969.

ADAPT: A Computer System for Automated Data Analysis Using Pattern Recognition Techniques

A. J. STUPER and P. C. JURS*

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802

Received December 11, 1975

An interactive computer system has been developed for the convenient application of pattern recognition techniques to chemical problems. The system includes subsystems to perform the following functions: graphical input of molecular structures for the generation of files of structures stored as connection tables; file handling and revision; development of descriptors from molecular structure connection tables for pattern recognition analysis; prior feature selection and preprocessing; discriminant development; feedback feature selection. The system is modular, allowing its execution on a relatively small computer and allowing the user to add or delete routines easily.

The introduction of the digital computer into the chemical laboratory offers the chemist a new and exciting tool. A multitude of tasks including literally hundreds of numerical integrations can be handled by the device with a speed and accuracy unapproachable a mere decade ago. Not all problems faced by the chemist, however, lend themselves to such exacting solution: frequently, equations describing processes of interest are difficult or impossible to obtain, and a host of problems have not yielded to a satisfactory or usable theoretical explanation. In the absence of theoretically based solutions, empirically derived methods will often suffice to yield useful and practical solutions to complex problems.

Standard approaches to the extraction of information from complex data forms have included linear optimization, information theory, and a plethora of statistical analysis techniques. Since the early 1950's pattern recognition methods have also been applied to a variety of data interpretation problems and have paralleled the computer's growth in speed and sophistication with a corresponding expansion in scope and capacity. Pattern recognition techniques have found application in such varied fields as computer and information science, engineering, statistics, biology, physics, medicine, and physiology. Each of these disciplines has adapted the basic methods of pattern recognition to its own specific requirements.

Analyses using pattern recognition have also encompassed a number of chemical problems in areas including mass spectrometry,[1-5] infrared spectrometry,[6-8] stationary electrode polarography,[9] material production problems,[10] NMR spectroscopy,[11,12] and gas chromatography.[13] Recently, several papers have appeared which indicate the utility of these techniques in the search for correlations between molecular structure and biological activity.[14-22] Two pattern recognition program packages have been described and offered to potential users.[23,24]

Pattern recognition methods are uniquely suited to a variety of studies because of several novel attributes. No mathematical model is used, but rather relationships are sought which provide definitions of similarity between diverse groups of data. Pattern recognition techniques are able to deal with high-dimensional data (data for which more than three measurements are used to represent each object). Such high-dimensional data cannot be directly visualized or displayed. In addition, pattern recognition techniques can deal with multisource data or data in which the relationships are discontinuous. In multisource data each measurement can be the result of an independent generating algorithm or experiment, and each can have a different scale, origin, distribution, etc., from all the other measurements. Therefore, there need be no direct functional relationship between the measurements in multisource data as there must be, for example, in an absorbance vs. concentration plot. For many chemical problems, and especially for those providing multisource data, it is difficult to know in advance whether an appropriate set of measurements has been generated to effect a satisfactory solution. The generation of sufficiently informative multisource measurements can become in itself a major part of the

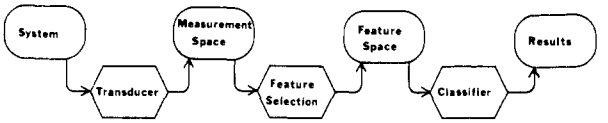

Figure 1. Schematic representation of a pattern recognition system.

overall pattern recognition experiment. When a number of measurements are available, pattern recognition can be used to judge their relative quality or utility with regard to specific questions. It is this ability to define relations through use of a diverse set of measurements which affords pattern recognition techniques their utility in such a wide variety of fields.

When properly used, pattern recognition techniques allow the chemist to develop criteria which relate the presence of properties to a particular subset of the total number of measurements. Once the important measurements are identified, they can be used to guide the development of subsequent experiments. For example, if a chemist were to find that ten structural parameters were important indicators of a particular biological effect, then he might hypothesize several as yet unstudied structures and use the results from the pattern recognition analysis to make an educated guess as to their effects. Alternatively, the fact that the particular ten parameters were shown to be important may lead to added insights into the problem. This ability to pick a subset of the original measurements which contains the bulk of the total information content is extremely desirable. As relations between several variables are not easily deduced through observation, this is an extremely useful capability of pattern recognition.

Before a chemist can routinely use pattern recognition techniques, the procedures he employs must be available in a convenient form. However, the architecture suggested in many of the standard pattern recognition texts is not generally usable. In the next sections we quickly review this architecture and expand upon it, while presenting a set of interactive computer programs which are currently operational and have been designed to serve as an aid in data intepretation through use of pattern recognition techniques.

BASIC PATTERN RECOGNITION SYSTEM ARCHITECTURE

Figure 1 shows a schematic representation of a basic pattern recognition system. This diagram, or a similar one, appears in many of the standard pattern recognition texts.[25-30] It consists of three interrelated subunits: a transducer, a feature extractor–preprocessor, and a classifier. Note that this simplified diagram avoids the question of the practicality of implementing any such techniques. Many operations fall between each of these schematic steps, and each operation may consist of a series of individual procedures, all of which must interface to each other. Additionally, a set of operations that has been developed for a specific problem may be entirely unsuitable for another. This would necessitate the development of a specific system for each new problem and could result in the chemist ignoring pattern recognition techniques in favor of others which are easier to apply. In order to apply pattern recognition techniques to studies of molecular structure–biological activity correlations, the data must be taken through a number of individual steps. These are listed in order to show how interrelated the steps become.

(a) Identify data set.

(b) Enter molecular structures. A complete description of the structure of each molecule must be entered into a file.

(c) Generate usable file. A subset of compounds must be selected from the master structure file. This may involve searching of keys for the structures, and will require carrying along an identifying label for each structure.

(d) Develop descriptors. The molecular structures stored in a general purpose form (e.g., connection tables) must be decomposed into sets of descriptors. The three general classes are topological, geometrical, and externally generated descriptors.

(e) Form data matrix. The subset of the available descriptors to be used is identified, and a matrix of data is generated. It may be partitioned into a training set and a prediction set.

(f) Prior feature selection. Techniques can be applied to determine which descriptors are expected to be most important.

(g) Develop discriminant. The data set is used to develop a discriminant function. After development, the discriminant function can be tested on unknowns to assess predictive ability.

(h) Feedback feature selection. The results of classification can be used to identify the most useful descriptors.

A general purpose system can be designed to implement all these steps. If the data set to be studied consists of a set of molecular structures, then all the steps are necessary. If the data set to be studied consists of numerical measurements (e.g., mass spectra), then steps b, c, and d are unnecessary and the analysis begins with step e. Whatever the source of the data, it will eventually be formed into a data matrix, where each row of the matrix contains all the measurements for one of the objects in the data set, and each column of the matrix consists of one particular measurement for all the objects. Preprocessing and prior feature selection of the data can be expressed as transformations of this data matrix. A potential pitfall in the implementation of pattern recognition techniques lies in the development of conventions concerning how these data matrices are constructed or input, labeled, stored, and accessed. If these operations are designed for a specific problem, then procedures suitable for one set of data may not suffice for others. Thus, one of the primary prerequisites for a useful general-purpose pattern recognition system is a general, data-independent, *file management* system. In the next section we describe a set of interactive computer routines known collectively as ADAPT (Automated Data Analysis using Pattern recognition Techniques). This system provides a generalized framework that takes into account the practical considerations inherent in the implementation of the pattern recognition framework shown in Figure 1.

ADAPT ARCHITECTURE

Figure 1 does not make clear the inherent diversity of the data handling problem. Not only must measurements from the transducer(s) be input, but they must be stored and labeled. Each data point must be given a class designation and identification number. Class designations must be easily assigned or modified. This ease of definition and redefinition is of utmost importance in the overall data analysis. The source of the data is also important. Sources such as digitized spectra or complex molecular structures would have widely different storage requirements. Since the operations performed on one type of data may bear little similarity to the operations performed on other types of data, a system designed with a high degree of modularity is required. To accomodate these requirements, the ADAPT system is designed to use computer routines known as overlays. Each overlay can execute independently, obtaining all necessary information either from a set of disc storage files or by interaction with the user. This mode of operation offers several advantages, the most obvious of which is a savings in core storage.

The modularity provided through use of overlays greatly decreases the complexity of the system and provides a means to incorporate additional algorithms into the system at any time. Thus the entire system is adapted to any user's individual

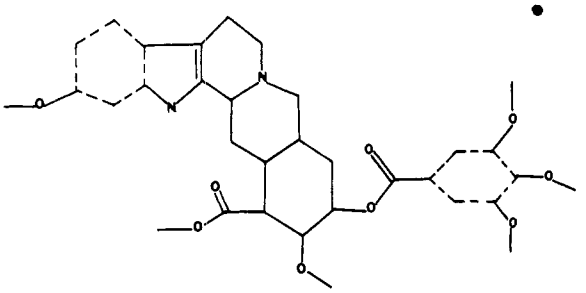

Figure 2. Reserpine as entered through UDRAW.

```
                              RESERPINE
        ID NO.   318.00                          DIRECT ACCESS NO.   277
ACTIVITIES   TMA   SED   HYPO
              1     1     1
CURRENT ACTIVITY CLASSIFICATION IS  NONE

                          BOND INFORMATION
ATOMS   C   N   O   P   S  CL   F  BR     BONDS     SNGL     DBL     TRPL     FENL
 44    33   2   9   0   0   0   0   0       49       34       3        0        12
                          RING INFORMATION
  ATOM     RING     ATOM     RING     ATOM     RING     ATOM     RING
    1    6 0 0      2    6 5 0      3    6 5 0      4    6 0 0
    5    6 0 0      6    6 0 0      7    5 6 0      8    5 6 0
    9    5 0 0     10    6 0 0     11    6 0 0     12    6 6 0
   13    6 6 0     14    6 0 0     15    6 6 0     16    6 6 0
   17    6 0 0     18    6 0 0     19    6 0 0     20    6 0 0
   21    6 0 0     30    6 0 0     31    6 0 0     32    6 0 0
   33    6 0 0     34    6 0 0     35    6 0 0
```

Figure 3. Report from AUTOGN for reserpine.

requirements since only those overlays which are relevant to the particular problem at hand need be executed. In addition, these routines are relatively inexpensive to use because they do not require large-scale facilities for execution. Finally, overlays are interactive in the sense that the user directs which manipulations are to be performed upon the data.

ADAPT thus consists of a framework within which an unlimited number of overlays can be supported. Each overlay performs a specific, independent operation ranging from initial input of data to final output of results. The general utility of the system arises from the fact that the user has a large number of options to choose from, and he can conveniently interact with his data set.

Interaction with ADAPT is provided via a Tektronix 4010 CRT terminal. Data are stored in a series of defined files on cartridge discs. This allows fast access and ease of manipulation. Currently, ADAPT consists of approximately 70 defined files which use 2.4 million bytes of storage (one cartridge disc). The ADAPT routine uses approximately 90,000 bytes of core storage for its largest overlay and is currently implemented using a 16-bit MODCOMP II/25 computer located in the Department of Chemistry at The Pennsylvania State University.

The next few sections will describe the individual overlays which are used in ADAPT to perform operations such as data handling, class development, descriptor generation, data input, and classification/feature selection processes.

DATA HANDLING

The ADAPT system is designed for input of two types of data—chemical structures and generalized data vectors. In the structural input mode an overlay call AUTOGN accepts chemical structures drawn in the familiar two-dimensional form on the screen of the CRT terminal. A subroutine, UDRAW, converts this structure to a compressed, augmented connection table. A more detailed description of UDRAW's capabilities can be found in the literature.[31] AUTOGN is capable of maintaining a library of 1000 structures and associated auxiliary information.

Once a structure is input, AUTOGN assigns it a direct access identification number (DAN number), and all relevant information keyed to the structure is stored in the appropriate defined files. This information can be altered by use of AUTOGN's alter mode and can similarly be reviewed in total or in part by AUTOGN's review mode. Unwanted structures are easily deleted from the structure files and new structures are easily added. Input times depend only upon the user's proficiency in drawing structures upon the CRT screen.

Figure 2 shows the molecule reserpine as it would be entered through the routine UDRAW, and Figure 3 shows the report output by AUTOGN. If it is desired to save this report, facilities for making a hardcopy (photocopy) of the CRT image are available. This routine can also display any structure from the file without its ancillary information, label the positions on the structure which are found to be contained in a ring, include carbon atoms in the display, as well as perform a variety of other display operations.

To speed structural input, AUTOGN stores up to ten different structural backbones, that is, structural forms which frequently occur in a series of molecules to be input. These can then be made to appear upon the initial UDRAW sketch pad, and a complete molecule can be built up starting from this backbone. This allows the user to input a series of structurally similar compounds without redrawing the base structure each time. In many studies it is necessary to generate three-dimensional models of molecules using a molecular mechanics routine named MOLMEC, to be described below. The backbone fragments can be modelled just as if they were complete molecules. Modelling of the backbones prior to their use in the generation of larger molecules decreases the amount of time MOLMEC must spend upon the larger molecule in order to place it in its lowest energy configuration.

As auxiliary information the user can assign a name and identification number to a structure for his own reference (e.g., index to his notebook files), as well as store a 16-word message concerning the structure.

Since the stored molecular structures will eventually be used in a classification scheme, the user may specify up to 65 different "class type" (activity) abbreviations with which to reference the structure. Any one structure may be present in as many as five of these classes simultaneously. To ensure error-free input, only activities specified by the user will be accepted. In addition to this, the user may specify a current activity to be any four-character code. This current activity can then be used as a special search indicator. A further discussion of the significance of the activity designations is given in the classmaker section of this paper.

In addition to maintaining files of molecular structures, AUTOGN also maintains a library of up to 100 substructures and a file of 30 environment fragments. These files can be reviewed and entries can be added or deleted with a minimum of effort. The entries in the substructure file are input through UDRAW in the same way structures are input. The environmental fragments are input through the CRT or the card reader using a simple code. The entries in these two files are used as data by several of the descriptor development routines.

Once a structure file is created it can be queried by a series of overlays which give various pieces of printed output regarding its contents. Information such as lists of the members and their respective DAN numbers can be generated, and lists of compounds which are indicated as having a given general or specific activity reference can be selectively printed. The user can employ these routines to quickly obtain specific information concerning the data set saved on the disc. These queries and all other commands are entered by the user via the CRT terminal.

CLASS DEVELOPMENT

Once the structural data are stored, they must be organized into classes. ADAPT treats all problems as two-class (yes, no) problems. However, overlays are available which can

quickly change class designations of the members.

ADAPT allows the user to create a list of 300 members which are to be used in the subsequent analysis. This list is called the working list. The working list contains a list of identifying numbers for the set of data that has been chosen for inclusion in the active data set. Its generation involves searching the structure files using search criteria entered by the user; no actual data are moved during the construction of the working list, however. An overlay, CLSMKR, creates and maintains this working list in any of three ways: (1) the user can specifically ask for any member by DAN number from the structure file, (2) the structure file can be searched using an activity from the list of general activities, or (3) the structure file can be searched using the specific activity originally specified in the AUTOGN section of ADAPT.

In the case of the general activity list, any member of the structure file containing the desired activity will be included in the users choice of yes or no classes, even though it may be labeled as having many different activities. Once a compound is entered in a yes or no list, it is prevented from being entered in the other list. Also, if a compound is entered as being one of the several general activities, it cannot be accessed as being any other general activity until its old activity designation has been discarded.

The ability to specify several different activities for a single molecule was designed with drug studies in mind. (This is the origin of the term activity as opposed to the term class designation.) Since many drugs have multiple activities, the use of a list which holds the major actions is a great aid in rapidly gathering structures which are considered to have this action "in general". Similarly, those structures cited to have a specific activity can be noted. This specific activity can then be used to quickly obtain all those drugs in the file which "specifically" have the desired activity. Obviously this procedure can be used to indicate the class of a structure in studies in which pharmacological activity is not the factor determining the class designation. The purpose of this type of designation is to identify a series of structures so that they can jointly be retrieved from the library on the basis of a one to four character user specified symbol, rather than individually retrieved by their DAN number. This allows rapid assembly of multiclass data into a two-class form.

The activity search is done by comparing bit strings and is extremely rapid even when using a relatively long access time moving head disc. A typical search requires two seconds of wall clock time. The generation of a complete working list takes no more than five minutes. Up to 600 members can be included in the yes and no classes. Of these, a maximum of 300 can then be selected to form the working list of compounds with which to perform a pattern recognition study.

The class designations of the members of the working list can be changed after the working list has been constructed. Thus, one could perform a series of studies using the same set of data, but with class memberships differently allocated, all without having to execute CLSMKR more than once. This is especially convenient for use in studies where the property being trained for is quantitatively known and one wishes to train a series of discriminants with different threshold cutoffs between the yes and no classes.

Up to this point in the information flow through the ADAPT system no data manipulations have occurred, and therefore the procedure is an interactive, real time process. The user can easily direct the data set construction and monitor his progress using output from the CRT display or the line printer. Changes are extremely easy. The time required for a change is limited by the speed with which the user can specify the new parameters. Lists are constructed from commands input via the CRT or a card reader. Card reader input is especially convenient since data sets can literally be constructed within seconds.

DESCRIPTOR GENERATION

Once the working list is constructed, the user can begin to generate descriptors for the molecules, whose identifying numbers are contained within it. ADAPT currently contains six separate descriptor generation routines which allow for the generation of descriptors derived from the connection tables formed in AUTOGN. Each routine has its own disc storage area and can be executed independently. These descriptor development routines are especially useful in that they allow the computer to act as an internal transducer for the overall pattern recognition system. This is highly desirable since the computer is much more consistent than a human for descriptive processes. Errors in the input of the parameters could conceivably cause erroneous rules to be developed. Use of the computer as a transducer minimizes this possibility.

The methods of generation of these descriptors are described in greater detail in the accompanying paper. A list of the names of the routines and a very brief description of the function of each is given below:

DODABN—Generates atom and bond fragment descriptors as well as a length parameter. Atom descriptors giving the number of C, N, O, S, P, and halogen atoms can be generated. Bond descriptors that can be generated are single, double, triple, phenyl, delocalized, and ionic.

DNVIRN—This routine gives a measure of the molecular environment of atom-centered fragments within a structure. Several different environments can be considered (connectivity, bond type, and bond-connectivities). AUTOGN maintains a user-generated list of possible atom fragments to be used in the environment calculation. The user chooses which of the fragment environments are to be calculated. The routine provides a report of results upon command.

DSSS—This routine finds the number of times a given substructural fragment appears in the molecules. The method of calculation is that proposed by Sussenguth[32] with changes by Zander.[33] Substructures to be searched are input by the user through AUTOGN. The user specifies which of the substructures are to be used in the DSSS search procedure. A report of results is produced on command.

DGEO—This routine uses the modelled structure to generate the principal axes for each of the structures. The eigenvalues for the principal axes X, Y, Z, and the ratios of eigenvalues for X/Z, Y/Z, and X/Y are stored as descriptors. The axes are derived from the eigenvectors of the three-dimensional matrix of atom positions which is generated by MOLMEC.

MVOL—This routine calculates the molecular volume of a molecule. The method of calculation is that proposed by Bondi.[34] It currently can be implemented for structures which have previously undergone the molecular modeling process. In order to calculate the molecular volume for structures which have not been modelled, the routine uses a table of standard bond lengths.

The overlays MVOL, DGEO, and AUTOGN use structures which have been modelled into a three-dimensional configuration which is considered a minimum strain energy configuration for the molecule. This energy-minimized structure has great potential for descriptor development. The MVOL routine is only one of many possible descriptive routines that could result from the application of this procedure.

The modelled three-dimensional structures are produced by a molecular mechanics routine called MOLMEC. It is similar to the several types of molecular modeling routines which have been reported.[35,36] Basically MOLMEC develops a structure which minimizes the energy of a molecule that is considered to be described by a summation of strain energy contributions due to the bonded, nonbonded, torsional, hybrid, and angular interactions within the molecule. A complete description of

MOLMEC is given in the following paper.

Through the application of the descriptor development routines, files of descriptors are built up for each molecule contained in the working list. These lists of descriptors, or subset of them, can then be used for pattern recognition analysis by the remainder of ADAPT.

GENERALIZED DATA INPUT

It was previously noted that in addition to structural input, ADAPT also can accept information in a generalized vector format. This mode of input is implemented to mimic the descriptor development routines. Since each descriptor generation routine has its own storage area, and its file structure is similar, data external to the ADAPT descriptor development routines may be treated as data from another descriptor package.

The routine enables the use of input sets of up to 300 objects or events, each of which can have up to 150 descriptors. This information could be a set of digitized spectra, physical measurements made on some system, data from on-line experiments, or descriptions of molecules in the working list that were generated external to the ADAPT descriptor development routines. For example, if a physiochemical measurement were determined experimentally for each of the molecules in the working list, it could be entered into ADAPT through XIN for analysis along with the internally generated descriptors. In this manner a basic data set can be constructed which contains information from a wide variety of sources.

At this point a data set consists of a working list and a number of files of descriptors saved independently on disc files. To proceed with the analysis of the data, a data matrix must be formed.

CLASSIFICATION AND FEATURE SELECTION

Up to this point all operations were those necessary to create or manipulate descriptors. From this point on we are concerned with the actual manipulations to be performed using these descriptors. The first step in this process is to form an active data set from those measurements that the user has determined to be significant. In general, not all of the measurements made upon a system are of interest. The user may wish to use only certain pieces of this information in the study. The overlay COLATE allows the user to pick any subset of the available descriptors and use them to form an active data set. Thus, the user can choose to include in an active data set only those descriptors that he suspects to be most important, or he can test several different subsets of descriptors in subsequent trials. This active data set can then be subjected to a whole spectrum of pattern recognition techniques.

COLATE stores the active data by descriptor. Each component of the descriptor contains the value of the descriptor for a given element in the order that it appears in the working list (generated by CLSMKR or by the XIN input). As an example of this type of storage, consider the case where one wishes to store 300 spectra each having 100 peak positions. COLATE would store these as a 100 by 300 matrix where the rows correspond to the intensities at peak positions labeled 1 to 100 and the columns correspond to the spectra, numbered 1 to 300, from which the peaks were taken.

This storage format is easily accessed by a classifier or any other program through the use of two arrays containing the indirect addressing pointers to the active rows and columns. The active row list is a list of which rows (molecules or spectra) of the data matrix are to be used, and the active column list contains a list of which descriptors are to be used. Initially, these arrays are set such that all the rows and columns defined by the working list and by COLATE are active. The user may

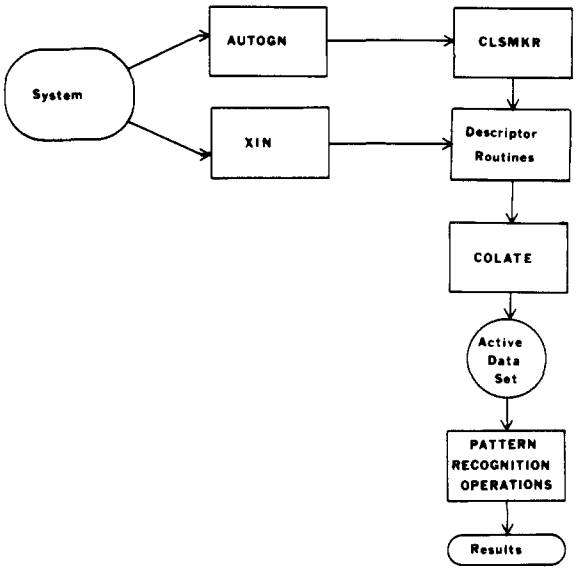

Figure 4. Diagram of information flow through ADAPT.

then deactivate any of these based upon results of feature selection or classification. Service routines exist which quickly perform such operations as generation of training and prediction sets, changing of class designations, and removal of unwanted descriptors. Many of these operations simply involve redefining which members in the rows and columns are to remain active. In this manner, manipulation of the active data set is quickly and easily accomplished. Figure 4 shows a summary of the operations one performs in order to place the data in a form suitable for pattern recognition analysis. Note that Figure 1 shows all the operations prior to this point as being a two-step procedure.

After it has been collated, the data matrix still consists of raw descriptors. ADAPT supports a wide variety of preprocessing or a priori feature selection methods which can be used to transform the data. COLATE can be used to implement the results from an a priori feature selection method in that it allows the user to specify which descriptors are to be used. There are a number of routines available to the user which calculate correlation coefficients, Fisher ratios, separation abilities of individual descriptors, separation abilities of all descriptors taken pairwise, a statistically based U statistic quantity, and others.

After preprocessing, the data set is presented to a discriminant development or a classification routine. ADAPT can support any of the common methods: linear learning machine, K nearest neighbor classifier, least-squares procedures, etc.

The weight vectors from the discriminant development routines can be used in a feedback loop to perform feature selection using variance feature selection.[37] This allows identification of a minimum set of descriptors which support linear separability for the data set. Alternatively, the user can perform feature selection himself based upon his own review of the results from any of the above discriminant functions or classifiers. The important point to be made is that the user has a wide number of choices and can conveniently exercise his ingenuity as he tries to extract meaning from large sets of high-dimensional data.

SUMMARY OF ADAPT ARCHITECTURE

A visual summary of the ADAPT system is presented in Figure 5. This figure explicitly shows the relationship between the major segments of the framework. Note how the working list is used to direct the operations performed by MOLMEC and all of the descriptor development routines. The working list is used by COLATE to create an active data set on which

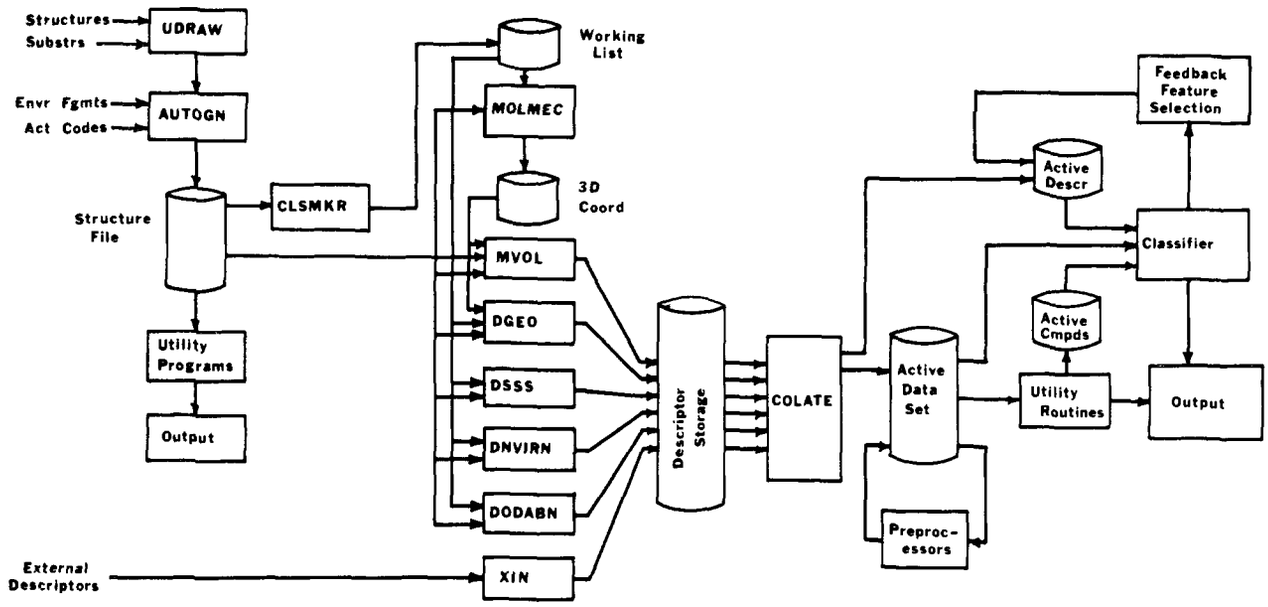

Figure 5. The ADAPT system.

the pattern recognition techniques will operate. Also shown is the fact that the preprocessing routines replace the data formed by COLATE with the preprocessed data. Not shown, however, are the many auxilliary routines which interact with the files to provide information to the user, and which enable him to change the contents of the active lists.

DISCUSSION

This report has described the system called ADAPT, which was designed to overcome the two major problems which arise in applications of pattern recognition to chemistry. The first is file management. One cannot interact with the data unless the tools are available to efficiently mold the data into useful forms. The use of the defined file system is one solution to this problem. The old files are easily expanded and new files are easily added as the system grows.

The second problem is that of having a routine which restricts the user to a fixed set of procedures. This is also solved by the use of overlays which interact with the file management system. Expansion of the capabilities of the system is relatively easy because of the modular construction of ADAPT. Newly developed procedures can access the data easily since it is already transformed into a useful form by the file maintenance routines. Use of overlays also allows the system to be executed on a small laboratory computer. This makes the system cheaper to implement and usually improves the access of the computer to the user, thus enhancing interactive computing.

It has not been our purpose to design a set of machine-independent routines which allow any user to create interactive pattern recognition routines which will run on any computer system. ADAPT is somewhat machine dependent. This allows us to take full advantage of the capabilities of the MODCOMP. However, ADAPT has been written almost entirely in Fortran IV, and it is conceivable that if one were willing to mimic some of the software functions provided by the MODCOMP, as well as have the appropriate graphics capabilities, ADAPT could be relocated. Most of the system dependence is in the graphics input and output since the Fortran graphics subroutines are designed to be used with the Tektronix CRT terminal.

The ADAPT routines have significantly increased the speed of discriminant development. They have also allowed a diverse collection of data types to be studied. A previous study[22] which had taken some 11 hours to complete using batch-type processing was rerun in approximately 2 hours of interactions with the ADAPT routines. Other studies in the area of structure activity relations have been completed as well, one of which contained 65 separate classes. Studies involving olfaction have been run, and various improvements in the area of feature selection have come about as a direct result of the use of the ADAPT routines. Results of these studies will be reported at a later date.

We feel that with development of this type of computer system, a major barrier to studies of chemical applications of pattern recognition techniques has been removed. The major questions of the ultimate utility of these techniques can now be addressed. ADAPT is currently being used in our laboratory on a wide spectrum of relevant chemical problems. Through the use of structural input routines we have developed a series of data sets with which we hope to investigate the application of pattern recognition techniques to drug structure activity studies. Also planned are experiments dealing with biological fluid analyses, spectral interpretation, and mixture identification. In short ADAPT has shown itself to be a useful tool which significantly decreased the time required to maintain data files, and allows the chemist to easily interact with various forms of data in an attempt to gain a better insight and understanding of the relations contained therein.

ACKNOWLEDGMENT

The authors would like to acknowledge W. E. Brugger for his assistance and support of this project. The National Science Foundation provided partial financial support for the MODCOMP computer used.

REFERENCES AND NOTES

(1) P. C. Jurs, B. R. Kowalski, and T. L. Isenhour, *Anal. Chem.*, **41**, 21 (1969).
(2) P. C. Jurs, B. R. Kowalski, T. L. Isenhour, and C. N. Reilley, *Anal. Chem.*, **42**, 1387 (1970).
(3) D. D. Tunnicliff and P. A. Wadsworth, *Anal. Chem.*, **45**, 12 (1973).
(4) K. Varmuza and P. Krenmayr, *Z. Anal. Chem.*, **266**, 274 (1973).
(5) P. Kent and T. Gaumann, *Helv. Chim. Acta*, **58**, 787 (1975).
(6) R. W. Liddell and P. C. Jurs, *Anal. Chem.*, **46**, 2126 (1974).
(7) S. R. Lowry, H. B. Woodruff, G. L. Ritter, and T. L. Isenhour, *Anal. Chem.*, **47**, 1126 (1975).
(8) G. S. Zander and P. C. Jurs, *Anal. Chem.*, **47**, 1562 (1975).
(9) L. B. Sybrandt and S. P. Perone, *Anal. Chem.*, **43**, 382 (1971).
(10) B. R. Kowalski, *Chem. Technol.*, 300 (May 1974).
(11) B. R. Kowalski and C. A. Reilly, *J. Phys. Chem.*, **75**, 1402 (1971).
(12) C. L. Wilkins et al., *J. Am. Chem. Soc.*, **96**, 4182 (1974).
(13) J. J. Leary et al., *J. Chromatogr. Sci.*, **11**, 201 (1973).

(14) K. L. H. Ting et al., *Science*, **180**, 417 (1973).
(15) S. A. Hiller et al., *Comput. Biomed. Res.*, **6**, 411 (1973).
(16) J. T. Clerc, P. Naegeli, and J. Seibl, *Chimia*, **27**, 639 (1973).
(17) B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.*, **96**, 916 (1974).
(18) C. L. Perrin, *Science*, **183**, 531 (1974).
(19) K. C. Chu, *Anal. Chem.*, **46**, 1181 (1974).
(20) R. O. Mathews, *J. Am. Chem. Soc.*, **97**, 935 (1975).
(21) K. C. Chu et al., *J. Med. Chem.*, **18**, 539 (1975).
(22) A. J. Stuper and P. C. Jurs, *J. Am. Chem. Soc.*, **97**, 182 (1975).
(23) L. A. Cox, Jr., R. H. Pritchard, and C. F. Bender, "RECOG: A Polyalgorithm for the Analysis of Generalized Data Sets. An Operator's Manual," UCID-16443, Rev. 1, Lawrence Livermore Laboratory, 1975.
(24) J. R. Koskinen and B. R. Kowalski, *J. Chem. Inf. Comput. Sci.*, **15**, 119 (1975).
(25) J. T. Tou and R. C. Gonzalez, "Pattern Recognition Principles", Addison-Wesley, Inc., Reading, Mass., 1974.
(26) W. S. Meisel, "Computer-Oriented Approaches to Pattern Recognition", Academic Press, New York, N.Y., 1972.
(27) N. J. Nilsson, "Learning Machines", McGraw-Hill, New York, N.Y., 1965.
(28) H. C. Andrews, "Introduction to Mathematical Techniques in Pattern Recognition", Wiley-Interscience, New York, N.Y., 1972.
(29) R. O. Duda and P. E. Hart, "Pattern Classification and Scene Analysis", Wiley-Interscience, New York, N.Y., 1973.
(30) P. C. Jurs and T. L. Isenhour, "Chemical Applications of Pattern Recognition", Wiley-Interscience, New York, N.Y., 1975.
(31) W. E. Brugger and P. C. Jurs, *Anal. Chem.*, **47**, 781 (1975).
(32) E. A. Sussenguth, *J. Chem. Doc.*, **5**, 36 (1965).
(33) G. S. Zander, Thesis, Department of Chemistry, The Pennsylvania State University, 1974.
(34) A. Bondi, *J. Phys. Chem.*, **68**, 442 (1964).
(35) J. E. Williams, P. J. Stang, and P. v. R. Schleyer, *Annu. Rev. Phys. Chem.*, **19**, 531 (1968).
(36) E. M. Engler, J. D. Andose, and P. v. R. Schleyer, *J. Am. Chem. Soc.*, **95**, 8005 (1973).
(37) G. S. Zander, A. J. Stuper, and P. C. Jurs, *Anal. Chem.*, **47**, 1085 (1975).

Generation of Descriptors from Molecular Structures

W. E. BRUGGER, A. J. STUPER, and P. C. JURS*

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802

Received December 11, 1975

To apply pattern recognition methods to studies of structure–activity relations, the molecular structures must be decomposed into numerical descriptors. The descriptors generated are of two types: topological and geometrical. Topological descriptors include: fragments, which code the atom and bond types; substructure descriptors, which code the presence or absence of particular, explicitly defined, substructures; and environmental descriptors, which code the immediate surroundings of an atom center of interest. The geometrical descriptors are derived from a strain-energy minimized three-dimensional structure, and they code the shape and size of the molecule. These descriptors can be used in conjunction with pattern recognition programs to investigate relationships between molecular structure and biological activity.

INTRODUCTION

A major problem in any pattern recognition system is the generation of informative pattern vectors which permit the correct classification of the objects or system under investigation.

In chemical applications of pattern recognition, several types of input data can be used to generate pattern vectors. Early investigations used single source data such as mass spectra,[1-3] infrared spectra,[4,5] NMR spectra,[6,7] and stationary electrode polarograms[8] as input to the pattern recognition system. The object of these investigations was to determine chemical structure information or to elucidate information about a chemical system. In these studies, the transformation of the spectra into pattern vectors was accomplished by first digitizing the spectra and then selecting all or parts of each spectrum to be the pattern vector. The major difficulty arose in obtaining consistent and representative spectra of a large number of compounds. However, with the abundant number of library reference spectra now available, this problem is lessened.

A second type of input consists of a compound's molecular structure alone. The purpose of these studies, where the molecular structure serves as the input, is to search for correlations between structure and the physical, biological, or pharmacological properties of the compound. Examples of investigations of this type have been published including the generation of the low-resolution mass spectrum of a given molecule,[9] structure–activity studies of drugs,[10-14] and studies of cancer chemotherapy agents.[15,16] For studies of this type, the generation of descriptors from the molecule's structures to be used in the pattern vectors constitutes a major problem. Unfortunately, there is no single method available to express all of the chemical information embedded within a chemical structure. Therefore, the pattern vector must be constructed of bits and pieces of information extracted from the molecular structure. It is the intention of this paper to describe the molecular development routines employed in the ADAPT pattern recognition system detailed in the companion paper.

Table I. Connection Table for 2-Methyl-6-bromobenzothiazole

Atom no.	1	2	3	4	5	6	7	8	9	10	11
1	1	4				4					
2	4	1	4								
3		4	1	4							1
4			4	1	4						
5				4	1	4			1		
6	4				4	1	1				
7						1	3	2			
8							2	1	1	1	
9					1			1	4		
10								1		1	
11			1								7

Atom type	Numeric codes	Bond type	Numeric codes
C	1	Single	1
O	2	Double	2
N	3	Triple	3
S	4	Aromatic	4
F	5	Delocalized	5
Cl	6	Ionic	6
Br	7		
I	8		
P	9		