ch2(F)/ch3(G),
c(C)/ch2(H),
ch2(H)/ch3(I)]).

REFERENCES AND NOTES

(1) Silk, J. V. Realistic versus Systematic Nomenclature. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 146–148.
(2) Wiswesser, N. J. *A Line Formula Chemical Notation*; Crowell: New York, 1954.
(3) Balaban, A. T. Applications of Graph Theory to Chemistry. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 334–343.
(4) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
(5) Wirth, K. Coding of Relational Descriptions of Molecular Structures. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 242–249.
(6) Abe, H.; Kudo, Y.; Yamasaki, T.; Tanaka, K.; Sasaki, M.; Sasaki, S.-I. A Convenient Notation System for Organic Structure on the Basis of Connectivity Stack. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 212–216.
(7) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 2. GENSAL, a Formal Language for the Description of Generic Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 151–161.
(8) Sussenguth, E. H. A Graph-Theoretical Algorithm for Matching Chemical Structures. *J. Chem. Doc.* **1965**, *5*, 36–43.
(9) Prolog terms are data structures used in logic programs. They are defined inductively as follows: A constant (i.e., an integer or atom) is a term; a variable is a term; a compound term is a term. A compound term is written by using the notation f(t1,t2,t3...tN), where f is the name of the compound term and each of the N arguments t1,t2...tN is a term. f is often referred to as the *functor* and N as the *arity* of the term. It is conventional to use the notation f/N to refer to some functor f of arity N. For more details see references 10 and 11.
(10) Bratko, I. *Prolog Programming for Artificial Intelligence*; Addison-Wesley: New York, 1986.
(11) Clocksin, W. F.; Mellish, C. S. *Programming in Prolog*; Springer-Verlag: Bonn, 1981.
(12) The programs in this paper have been tested and run on Quintus Prolog Version 1.5 running under BSD 4.2 UNIX, ALS Prolog Version 1.2 running under MS-DOS Version 3.1, and Sussex Poplog Prolog Version 10.
(13) Read, R. C. A New System for the Design of Chemical Compounds. 1. Theoretical Preliminaries and the Coding of Acyclic Compounds. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 135–149.

PAD Programming and Its Application in Chemistry

ZHENG XIANMIN

Chemistry Department, Hua Chiao University, Quanzhou, Fujian, China

Received July 26, 1988

Problem analysis diagram (PAD) is the most vital representation of software design at present. A synopsis of its principles, graphic representation, writing mode, and structure is given. The application of this software engineering method to structured programming in chemistry is also discussed and exemplified with problems common to the teaching and research of chemistry. Users who are not computer professionals can program efficiently in solving chemical problems with the computer.

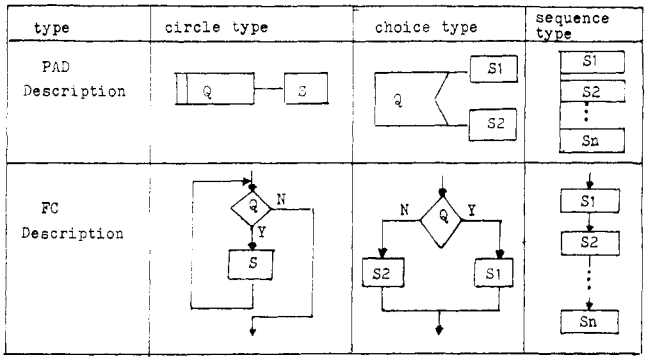
Software has developed into an epoch of software engineering in which the engineering method of expressing software design, manufacture, checking, and maintenance by graphics came into being. For applying software engineering methods to the field of chemistry so as to lessen the trouble that may be encountered in solving chemical problems, this paper presents the problem analysis diagram (PAD) method with special reference to its principle, graphic representation, writing mode, and structure. The application of PAD to structured programming in problems common to the teaching and research of chemistry is also discussed here.

SYNOPSIS OF PAD

PAD is the presentation of software design characterized as a two-dimensional tree structure.[1,2] By using PAD in structured programming, it is possible to represent program logic tersely so as to raise the efficiency of software design, manufacture, checking, and maintenance greatly and to make the program easy to read, remember, and understand.

(1) Basic Principle of PAD. As a two-dimensional tree structure representation for software design, PAD was generated on the basis of the improved Warnier diagram. Because of the use of the control structure of Pascal, it may be regarded as a two-dimensional expansion graph. Thus, PAD may also be regarded as an abbreviation of a Pascal diagram, known as the expansion graph of Pascal. Its essence lies in using the basic concept of top-down design and continual improvement so as to transfer the sketchy, vague idea for solving a problem into definite and thorough computerized processes.

Table I. Program Structure and Elementary Forms of PAD



Six symbols are used by PAD to describe treatment, repetition, selection, statement label, definition, and process, as shown in Figure 1.

Similar to other stipulated software methods, PAD provides procedures that should be followed by the software designer in designing a system or program.

(a) Petitioning of Sequence. Beginning with the design of a fuzzy concept of procedure and system, mark the sequence of each part of the existing process that is freely defined.

(b) Petitioning of Circulation. Mark the part that will be repeated and the condition of the beginning and end of repeating, that is, the condition of the beginning and end of the principal circulation process.

(c) Petitioning of Choice. Mark the condition of every process that will be implemented.

0095-2338/89/1629-0060$01.50/0

Name	Symbols	Explaination
Processing Frame		The process name or various statement are written in the frame.
Repetition Frame		Post-decision circle. Condition for circle are written in the frame.
		Pre-decision circle. Condition for circle are written in the frame.
Choice Frame		Condition for choice are written in the frame.
Statement Number		Statement number is written in the circle.
Definition or Definition Frame	def	For use in defining or resolving PAD.
		The definition of PAD name is written in the frame.
Subprograms Processing Frame		The subprograms are written in the frame.

Figure 1. Some symbols used in PAD.

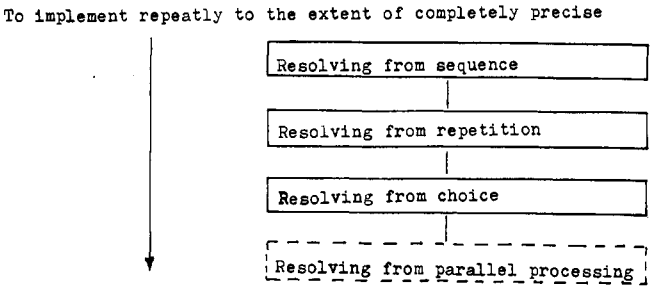
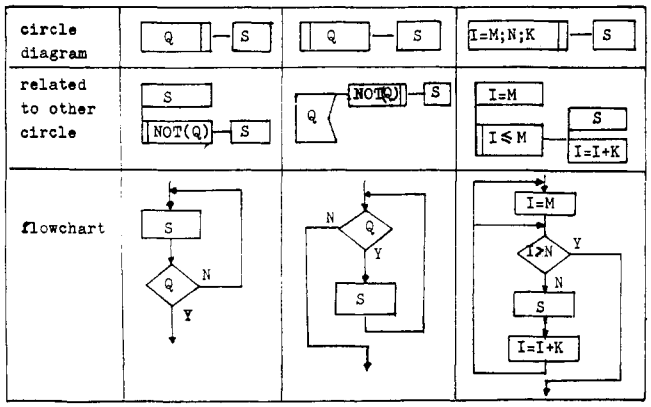


Figure 2. Processing flow of PAD method.

Table II. Relationship between Predecision and Postdecision



(d) Petitioning of Parallelism. Mark the parts that will be implemented in parallel.

The processes of the above procedures are shown in Figure 2.

(2) Program Structure and Basic Graphic of PAD. There are three elemental structure forms of PAD: (a) sequential structure, to deal with time series of more than two cases; (b) structure in circle, to implement over and over again as soon as the condition exists; and (c) structure of choice, to choose just one case meeting the demand when there are more than two cases that need to be treated.

Taking these three program structures of PAD and the corresponding flowchart (FC) as one side and the basic graphic of PAD as the other side, the relationship between them is shown in Table I. The structure in circle involves predecision, postdecision, and problem-oriented forms; their relationships are shown in Table II. The parallel structure for parallel processing is shown in Figure 3.

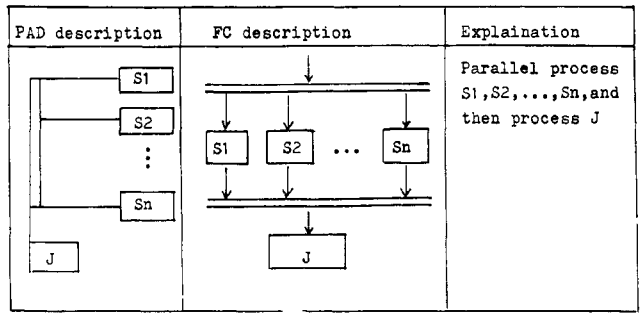


Figure 3. Diagram of parallel processing.

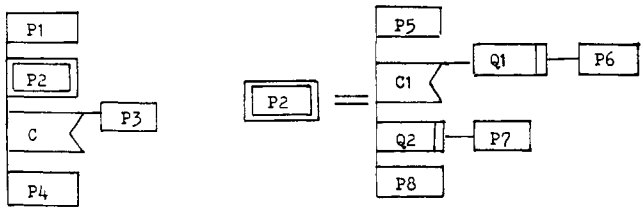
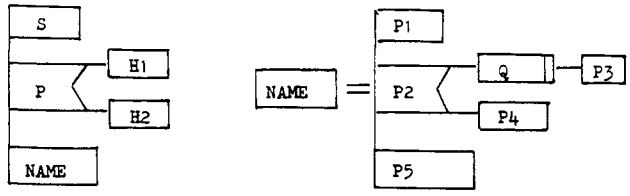


Figure 4. Definition of the usage of PAD details.



(a) PAD in preceding page (b) PAD in suceeding page

Figure 5. Definition of the linkage of successive pages.

Table III. Elementary Forms of Data Structure

elementary form	PAD	implication
sequence	A, B	Data A followed by data B
circle	N, A	To operate data repeatly to N times, N can be omitted if it is an unknown number, and the conditions for circulation can also be written as I = 1;N.
choice	Q, A, B	To choose data A or B according to condition Q.

(3) Writing Mode of PAD. As a two-dimensional tree structure program representation, PAD expresses a unique graphic/concept combination. It expresses sequential information vertically and describes branches and inlaid layers horizontally, which combine to form a PAD diagram.

In writing PAD, the user may start with defining the top chart of a program or program groups and then add the detailed chart by means of the symbols def or =. These symbols may be written as follow when they are used for defining PAD.

[S] def PAD or [S] = PAD

S is the name of PAD as defined. The following can be defined by means of the symbols def or =: (a) the usage of details of PAD (see Figure 4); (b) the usage of the linkage of successive pages (see Figure 5).

(4) Usage of PAD Describing Data Structure. PAD data structure can be described in a very simple audio-visual form, similar to the three elemental forms of program structures (see Table III).

(5) PAD and High-Level Languages. The writing mode of PAD bears no relation to language on the whole; it can be

Table IV. PAD Standard Diagram Used by BASIC

		PAD	BASIC
Linkage		H1 H2	H1 H2
Circle	Pre-Decision	L1 Q H L2	L1 REM IF NOT Q THEN L2 H GOTO L1 L2 REM
Circle	Post-Decision	L Q H	L REM H IF NOT Q THEN L
Circle	Problem-Oriented	I=M;N;K H	FOR I =M TO N STEP K H NEXT I
Choice	Simple Type	Q H L	IF NOT Q THEN L H L REM
Choice	Bifurcate Type	Q L1 H1 L2 H2	IF NOT Q THEN L1 H1 GOTO L2 L1 REM H2 L2 REM
Choice	Multiplexed Type	L1 H1 L2 H2 Ln Hn L I = 1 2 n	ON I GOTO L1,L2,...,Ln L1 REM H1 GOTO L Ln REM Hn L REM

applied to the programming of any high-level language. The expansion can be made by users according to the need, except for the standard graphic.

The standard graphic of PAD used by BASIC, FORTRAN, and COBOL is shown as Tables IV and V.

The source program can be formed easily from PAD, for it is a two-dimensional tree structure used in expressing the program. In PAD programming, the first step is to write out PAD graphic according to the mathematic model and then to program in light of this graphic. There are two programming methods: one is to write out the source program manually in light of PAD; the other is to key in PAD graphic to the computer directly with the support of the PAD system, and then the source program will be compiled automatically by the machine. The latter method is exemplified by all the programming in this paper.

APPLICATION OF PAD IN CHEMISTRY

Along with the rapid development of computer sciences, the computer has become an indispensable tool in chemistry. However, programming remains to be trouble for those who are not computer professionals. The application of PAD programming will be helpful for them. It will express program logic succinctly and will raise the efficiency of programming, manufacture, checking, and editing. This can be seen directly from the comparison of the program structure and elementary forms between PAD and FC (see Table I). Furthermore, with the support of the PAD system, the computer will work out programs corresponding to PAD that the users key in and will operate them, thus facilitating the work of program editing for those who are not computer professionals. The following illustrations on how to apply PAD in structured programming are exemplified by problems common to the teaching and research of chemistry. The processing of all the examples is the results of operation on IBM-PC/XT in light of keyed-in PAD graphic by the author.

(1) Calculation of Thermodynamic Data. In physicochemical studies, calculation of thermodynamic data in great number is very common and can be made simple and popular

Table V. PAD Standard Diagram Used by FORTRAN and COBOL

		PAD	FORTRAN	COBOL
Circle	Post-decision	L Q H	L CONTINUE H IF NOT Q GOTO L	PERFORM H. PERFORM H UNTIL Q.
Circle	Pre-decision	L Q H L1	L CONTINUE IF NOT Q GOTO L1 H GOTO L L1 CONTINUE	PERFORM H UNTIL (NOT Q).
Circle	Problem oriented	I=M;N;K H L	DO L I = M;N;K H L CONTINUE	PERFORM H VARYING I FROM M BY K UNTIL I K.
Choice	Bifurcated Type	Q L1 H1 L2 H2	IF NOT Q GOTO L1 H1 GOTO L2 L1 CONTINUE H2 L2 CONTINUE	IF Q H1. ELSE H2.
Choice	Simple Type	Q H L	IF NOT Q GOTO L H L CONTINUE	IF Q THEN H.
Choice	Computed Branched Type	L1 H1 L2 H2 L3 H3 L4 I = 0 0 0	IF I L1,L2,L3 L1 CONTINUE H1 GOTO L4 L2 CONTINUE H2 GOTO L4 L3 CONTINUE H3 L4 CONTINUE	
Choice	Multiplexed Type	L1 H1 L2 H2 Ln Hn L I =	GOTO (L1,L2,...,Ln),I L1 CONTINUE H1 GOTO L Ln CONTINUE Hn L CONTINUE	GOTO L1,L2,...,Ln DEPENDING ON I. L1. H1. GOTO L. Ln. Hn. L.

by application of the computer. Using PAD in structured programming not only makes the calculator program simple and convenient but also makes the thermodynamic data calculation succinct in logic so that it is understandable and easily modified. The preparation of PAD for thermodynamic data calculation of any ideal gas reaction is exemplified as follows:

(a) Setting up a Mathematical Model. For any ideal gas reaction gG + hH = nN + mM. According to the van't Hoff equation

$$(\partial \ln K_p/\partial T)_p = \Delta H^\circ(T)/RT^2 \quad (1)$$

where

$$\Delta H^\circ(T) = \Delta H^\circ_{298} + \int_{298}^{T} \Delta C_p \, dT \quad (2)$$

$$\Delta H^\circ_{298} = \sum \nu_i \Delta H_f^\circ{}_{298,i} \quad (3)$$

$$\Delta C_p = \Delta a + \Delta bT + \Delta cT^2 \quad (4)$$

When (3) and (4) are substituted into (2), we obtain

$$\Delta H^\circ(T) = \Delta H^\circ_{298} + \Delta a(T - 298) + \Delta b(T^2 - 298^2)/2 + \Delta c(T^3 - 298^3)/3 \quad (5)$$

To substitute (5) into (1) and integrate it

$$\ln K_p = -\Delta H^\circ_0/RT + \Delta a \ln T/R + \Delta bT/2R + \Delta cT^2/6R + I \quad (6)$$

where

$$\Delta H^\circ_0 = \Delta H^\circ_{298} - \Delta a 298 - \Delta b(298^2/2) - \Delta c(298^3/3) \quad (7)$$

$$I = \Delta G^\circ_{298}/298 + \Delta H^\circ_0/298R - \Delta a(\ln 298)/R - \Delta b(298/2R) - \Delta c(298^2/6R) \quad (8)$$

$$\Delta G^\circ(T) = -RT \ln K_p \quad (9)$$

$$\Delta S^\circ(T) = (\Delta H^\circ(T) - \Delta G^\circ(T))/T \quad (10)$$

Table VI. Thermodynamic Data Related to Reaction a

	C_6H_6	CH_4	$C_6H_5CH_3$	H_2
ν_i	−1	−1	1	1
$\Delta H_f°_{298,i}$ (J/mol)	82.93×10^3	-74.848×10^3	49.999×10^3	0
$\Delta G_f°_{298,i}$ (J/mol)	129.076×10^3	-50.794×10^3	122.298×10^3	0
a_i	−33.899	17.451	−33.882	29.079
b_i	471.872×10^{-3}	59.204×10^{-3}	557.045×10^{-3}	-0.837×10^{-3}
c_i	-298.344×10^{-6}	1.117×10^{-6}	-342.373×10^{-6}	2.013×10^{-6}

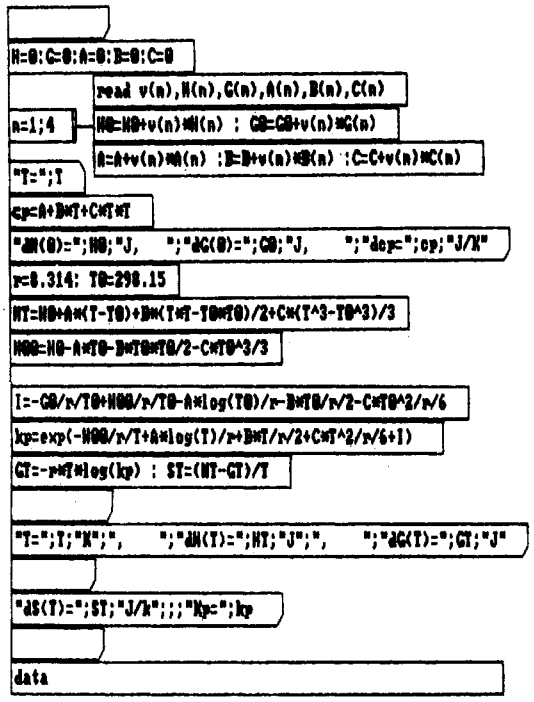


Figure 6. PAD graphic for thermodynamic data calculation of any ideal gas reaction.

(b) Writing of PAD. Let $H0 = \Delta H°_{298}$, $G0 = \Delta G°_{298}$, $H00 = \Delta H°_0$, $HT = \Delta H°(T)$, $GT = \Delta G°(T)$, $ST = \Delta S°(T)$, $A = \Delta a$, $B = \Delta b$, $C = \Delta c$, and $C_p = \Delta C_p$. Then, according to the preceding mathematic models, the PAD written in BASIC language is shown in Figure 6.

The PAD as shown in Figure 6 is suitable for the calculation of thermodynamic data in any ideal gas reaction. In calculation, the results can be obtained by merely supplying the given thermodynamic data of specific topic in a data statement, for example, to calculate $\Delta H°$, $\Delta G°$, $\Delta S°$, and K_p of the reaction

$$C_6H_6(g) + CH_4(g) = C_6H_5CH_3(g) + H_2(g) \qquad (a)$$

in case of $T = 773$ K. The thermodynamic data related to reaction a are shown in Table VI. When the data of Table VI are used in a data statement of PAD in Figure 6 and keyed into the computer, the results are obtained as follows:

run
T = ? 773
dH(0) = 41917 J, dG(0) = 44016 J, dC_p = 5.298788 J/K
T = 773 K, dH(T) = 47578.3 J, dG(T) = 43699.14 J
dS(T) = 5.018314 J/k, K_p = 1.114221E-03
Ok

(2) Balancing the Equation of Chemical Reaction. There are several methods of balancing the equation of chemical reaction. Balancing by the algebraic method has been taken seriously along with the application of computer in chemistry.[3,4] The algebraic method not only works well in universal programs for balancing various equations but it also is convenient for balancing some complex chemical equations. According to the mathematical model for solving first-order multivariate equations and taking advantage of the PAD programming technique, a universal PAD graphic for balancing chemical equations is worked and shown in Figure 7. The following two examples are given to show how to use Figure 7 to balance chemical equations.

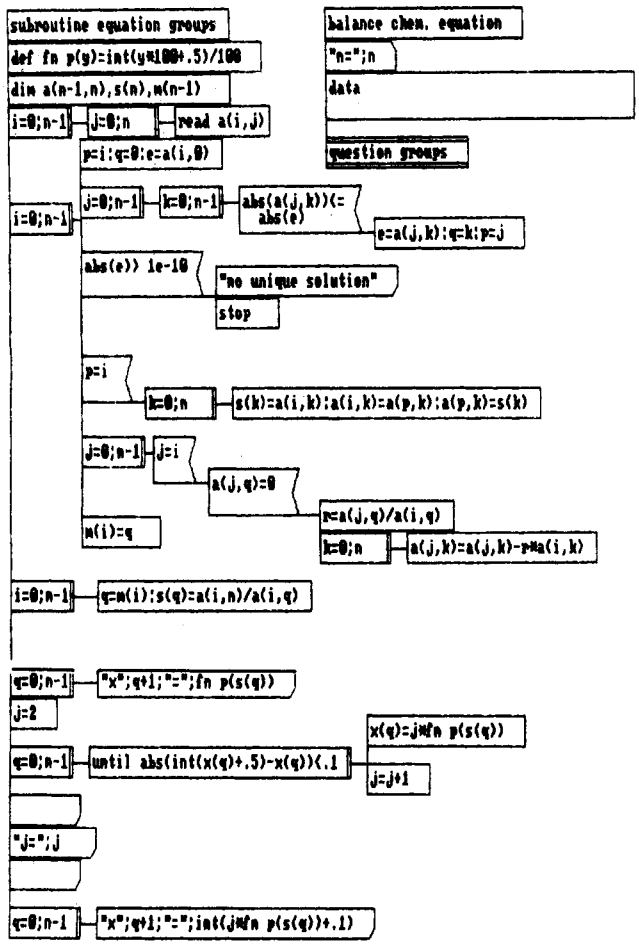


Figure 7. Universal PAD graphic for balancing chemical equations.

Example 1. To balance the equation

$$Pb(N_3)_2 + Cr(MnO_4)_2 \rightarrow Cr_2O_3 + MnO_2 + Pb_3O_4 + NO$$

suppose the equation after balancing is

$$X_1Pb(N_3)_2 + X_2Cr(MnO_4)_2 = X_3Cr_2O_3 + X_4MnO_2 + X_5Pb_3O_4 + X_6NO$$

Then the relational expression of atomicity of various elements on both sides of the equation will be

$$\begin{aligned} \text{Pb:}\ & X_1 - 3X_5 = 0 \\ \text{N:}\ & 6X_1 - X_6 = 0 \\ \text{Cr:}\ & X_2 - 2X_3 = 0 \\ \text{Mn:}\ & 2X_2 - X_4 = 0 \\ \text{O:}\ & 8X_2 - 3X_3 - 2X_4 - 4X_5 - X_6 = 0 \end{aligned} \qquad (11)$$

and let

$$X_1 = 1 \qquad (12)$$

Formulas 11 and 12 can be written in matrix symbol as

$$\begin{bmatrix} 1 & 0 & 0 & 0 & -3 & 0 \\ 6 & 0 & 0 & 0 & 0 & -1 \\ 0 & 1 & -2 & 0 & 0 & 0 \\ 0 & 2 & 0 & -1 & 0 & 0 \\ 0 & 8 & -3 & -2 & -4 & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

To fill in the values from the above matrix in the data statement of PAD in Figure 7 and to key them into the computer, the coefficients of every reaction component obtained after operation are shown as follows:

```
20 DATA 1,0,0,0,-3,0,0,6,0,0,0,0,-1,0,0,1,-2,0,0,0,0,0,2,
        0,-1,0,0,0,0,8,-3,-2,-4,-1,0,1,0,0,0,0,0,1
```

```
run
N = ? 6
× 1 = 1
× 2 = 2.93
× 3 = 1.47
× 4 = 5.87
× 5 = 0.33
× 6 = 6

J = 15

× 1 = 15
× 2 = 44
× 3 = 22
× 4 = 88
× 5 = 5
× 6 = 90
Ok
```

So the equation after balancing is

$$15Pb(N_3)_2 + 44Cr(MnO_4)_2 = 22Cr_2O_3 + 88MnO_2 + 5Pb_3O_4 + 90NO$$

Example 2. To balance the equation

$$P_2I_4 + P_4 + H_2O \rightarrow PH_4I + H_3PO_4$$

suppose the equation after balancing is

$$X_1P_2I_4 + X_2P_4 + X_3H_2O = X_4PH_4I + X_5H_3PO_4$$

By the above method, it may be written in matrix form as follows:

$$\begin{bmatrix} 2 & 4 & 0 & -1 & -1 \\ 4 & 0 & 0 & -1 & 0 \\ 0 & 0 & 2 & 4 & -3 \\ 0 & 0 & 1 & 0 & -4 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

To key in the values from the above matrix to the computer, the operation results are shown as follows:

```
20 DATA 2,4,0,-1,-1,0,4,0,0,-1,0,0,0,0,2,-4,-3,0,0,0,1,
        0,-4,0,1,0,0,0,0,1

run
N = ? 5
× 1 = 1
× 2 = 1.3
× 3 = 12.8
× 4 = 4
× 5 = 3.2

J = 10

× 1 = 10
```

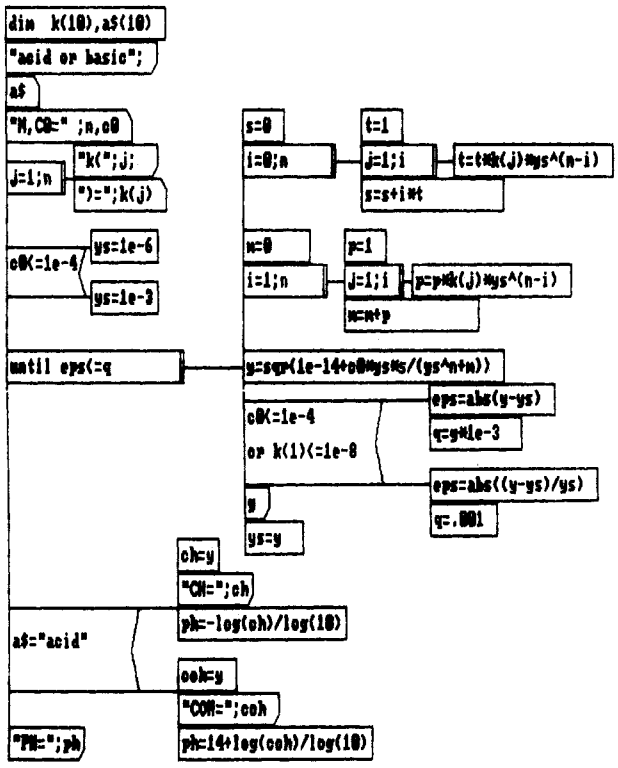


Figure 8. Universal PAD graphic for calculating the pH of various weak acid (base) balanced systems.

```
× 2 = 13
× 3 = 128
× 4 = 40
× 5 = 32
Ok
```

So the equation after balancing is

$$10P_2I_4 + 13P_4 + 128H_2O \rightleftharpoons 40PH_4I + 32H_3PO_4$$

(3) pH Value Calculation for Weak Acid (Base) Balance System. Weak acid and weak base balanced systems are complex systems commonly met with in analytical chemistry. The proton-transfer reaction of solvent water makes the calculation very complex; it deals with higher order equations in calculating precisely the pH value of a weak acid (base) balanced system. The application of the computer not only simplifies the calculation but also carries out the precise calculation on the pH value of various weak acid (base) balanced systems conveniently. The universal formula[5] for calculating precisely the pH value of a weak acid (base) balanced system is

$$y = \left(k_w + \frac{C_0 y \sum_{i=1}^{n} i \prod_{j=1}^{i} K_j y^{n-i}}{y^n + \sum_{i=1}^{n} \prod_{j=1}^{i} K_j y^{n-i}} \right)^{1/2} \tag{13}$$

in which y represents the concentration of H^+ (weak acid) or OH^- (weak base), n the dimension of weak acid (base), and K the dissociation constant. According to eq 13, the PAD of the universal calculator program prepared by using PAD in structured programming and the corresponding flowchart (FC) for calculating precisely the pH value of various weak acid (base) balanced systems are shown in Figures 8 and 9, respectively.

Two conclusions can be made from the comparison between Figures 8 and 9: (1) The programming structure of Figure 9 is complicated and tedious and the concept is fuzzy, while that of Figure 8 is simple and clear and the logical transpa-

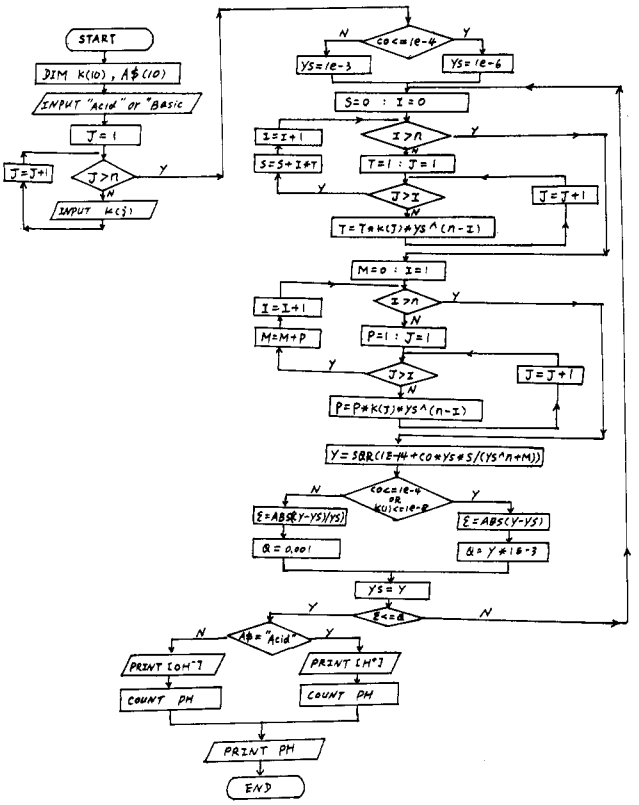


Figure 9. FC graphic for calculating the pH of various weak acid (base) balanced systems.

rence is high. (2) The computer can operate only after the tree graphic of Figure 9 is transferred to a source program manually, while for Figure 8, with the support of the PAD system, the computer can transfer the source program automatically and operate it according to the PAD of Figure 8 keyed in.

As examples of pH value calculation in balanced systems using Figure 8, the calculated results of pyrophosphoric acid of 10^{-8} mol/L and diamine of 10^{-8} mol/L are cited below:

```
run
acid or basic? acid
N,CO = ? 4,1e-8
k(1) = ? 3e-2
k(2) = ? 4.4e-3
k(3) = ? 2.5e-7
k(4) = ? 6.5e-10
  1.802966E-07
  1.310263E-07
  1.234047E-07
  1.22174E-07
  1.219739E-07
  1.219414E-07
CH = 1.219414E-07
PH = 6.91385
Ok
```

(a) pyrophosphoric acid

```
run
acid or basic? b
N,CO = ? 2,1e-8
k(1) = ? 3e-6
k(2) = ? 7.6e-15
  1.322876E-07
  1.061462E-07
  1.050009E-07
  1.0495E-07
COH = 1.0495E-07
PH = 7.020983
Ok
```

(b) diamine

The results of pH value calculation in some common and typical weak acid (base) balanced systems obtained by the above method are described in detail in another of the author's papers.[5]

(4) Processing of Experimental Data. Processing experimental data by computer is now very popular. A typical experiment in kinetics is the measurement of the velocity constant of sucrose hydrolysis by the polarimetrical method. A series of data measured in various reaction times (t) of hydrolysis is shown in Table VII. By use of PAD in structured

Table VII. Experimental Data from Sucrose Hydrolysis (a_∞ = −4.33, T_1 = 313.2 K, C_{H^+} = 1.8 mol/L)

	t (min)					
	5	10	15	20	25	30
a_t	8.85	4.10	1.05	−0.60	−1.75	−3.00

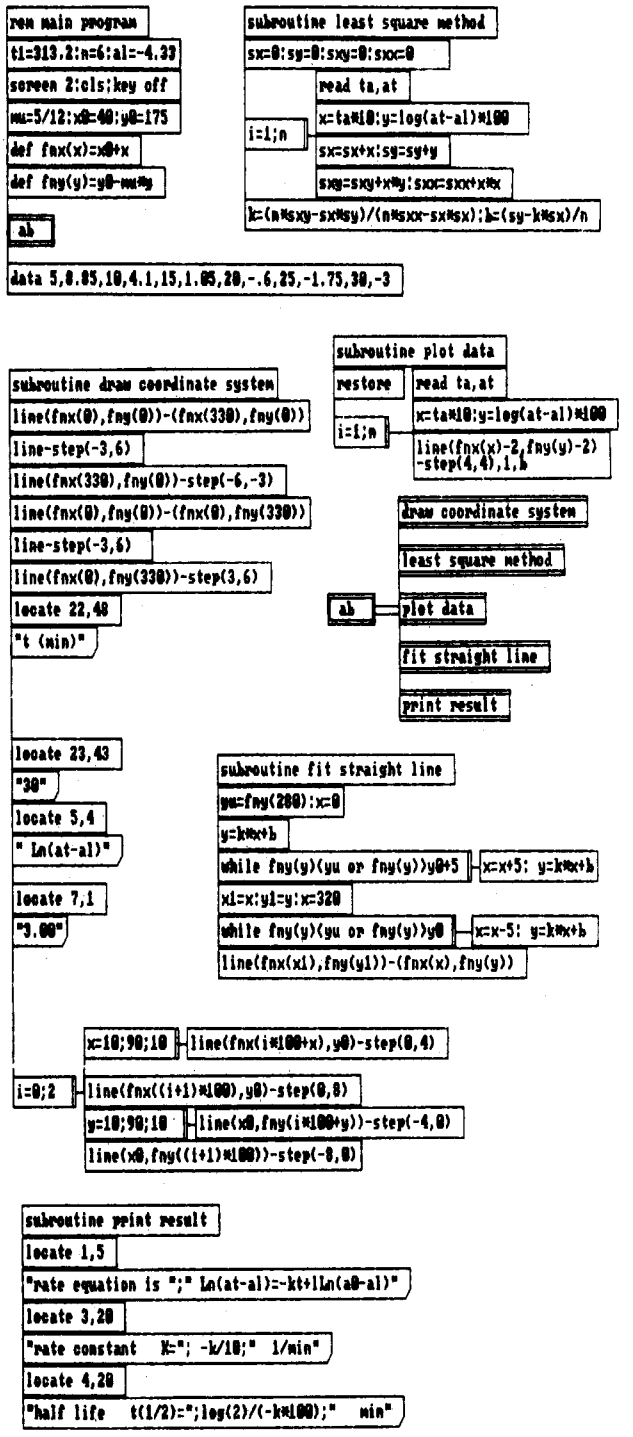


Figure 10. PAD graphic for determining the velocity constant and half-life of sucrose hydrolysis.

programming to process the experimental data in Table VII, the PAD for determining the reaction velocity constant and half-life is shown in Figure 10.

It can be seen from Figure 10 that PAD is made up of one main program structure and five subprograms that process drawing the coordinate system, performing the least-squares method, plotting the data, fitting the straight line, and printing the result, respectively. Program structures manifest them-

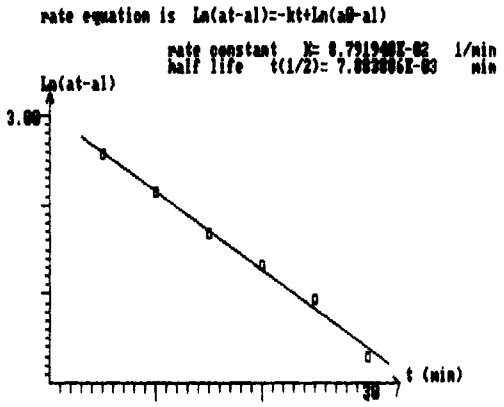


Figure 11. Operational results of Figure 10.

selves as succinct and transparent in logic, easy to read, remember, and understand. After the PAD structure is keyed into the computer, the operational results are shown in Figure 11.

CONCLUSION

From the above examples and the comparison of the program structure described between PAD and FC (see Table I) as well as the comparison between Figure 8 (the universal PAD graphic for calculating the pH of various weak acid balanced systems) and Figure 9 (the FC graphic for calculating the pH of various weak acid balanced systems), we obtain the following conclusions:

(1) Demonstrating the program by using PAD makes the program structure simplified, the logic succinct and transparent.

(2) It is easy and convenient to converse PAD into a source program (compiled manually or automatically by the computer).

(3) By use of PAD in structured programming, the program is easy to read, remember, and understand.

With the support of the PAD system, the computer can do many things such as compile programs, operate programs, and output operational results automatically according to PAD keyed in by the users. Thus, the efficiency of programming, editing, manufacturing, and checking can be raised by reducing many problems met with by workers who are not computer professionals. For this reason, to apply PAD in dealing with various problems in chemistry is a step forward in applying software engineering to the field of chemistry as well as a leap in the history of computing chemistry.

REFERENCES

(1) Nimura, Yoshihiko. In *Program Technique—PAD Structured Programming*; OMU Co.: Japan, 1985.
(2) Guilan, Yan; Jiayao, Liu. Lecture on Software Engineering. *Computer Studies in Fujian*; Fujian Publishing House: Fujian, China, 1986; pp 1-4.
(3) Kennedy, J. H. *J. Chem. Educ.* **1982**, *59*, 523.
(4) Blakely, G. R. *J. Chem. Educ.* **1982**, *59*, 728.
(5) Xianmin, Zheng. pH Value Calculation for Weak Acid (Base) Balanced System—PAD Programming Technique. *Proceedings of 2nd Computing Chemistry of the Chinese Chemical Society*; Jiangxi Publishing House: Jiangxi, China, 1988; Vol. 7, p 26.
(6) Peters, Lawrence J.; Belady, L. A. *Software Design: Methods & Techniques*; Yourdon: Englewood Cliffs, NJ, 1981.

Clustering a Large Number of Compounds. 1. Establishing the Method on an Initial Sample

LOUIS HODES
National Cancer Institute, Bethesda, Maryland 20892

Received March 25, 1988

The National Cancer Institute Division of Cancer Treatment has revised its drug-screening program. About 230 000 compounds in our repository are available for screening under the new protocol. This paper is the first on an attempt to extract a representative sample of these compounds by clustering. It reviews the establishment of the clustering method on a 4980-compound initial sample. The clustering algorithm is fairly simple. However, the molecular fragments employed to match the compounds are somewhat complex to distinguish a large number of compounds.

INTRODUCTION

The National Cancer Institute (NCI) Division of Cancer Treatment (DCT) Developmental Therapeutics Program (DTP) has been converting its primary screening program from in vivo mouse models to cell cultures derived from human cancers.[1] It should soon be possible to screen a large number of compounds, of which many are available from our store of several hundred thousand compounds that have been acquired over the years of our program for the earlier screens.

A search of our file revealed 232 000 compounds with inventory sufficient for this new test. The work reported here is an effort to find a representative sample of these compounds for large-scale testing on the new screens.

Such a sample may be obtained by clustering the compounds according to molecular structure. One or more compounds can be chosen from each cluster.

We assume that compounds with similar structure tend to have similar test results. However, this is only a first approximation, and there are many counterexamples. Therefore, we require the compounds in a cluster to be very much alike. Ideally, they should differ in only one functional group. Such a strict criterion will yield a relatively large number of clusters, which agrees with our need to test as many different substances as possible.

There is also the question as to whether this job stretches the limits of current computer capability. That is, the sheer amount of data may render the project infeasible.

In this paper we present work on an initial sample of 4980 compounds to show how the clustering was developed to yield a satisfactory separation of compounds. The sequel will describe the use of a new species of computer to accomplish the large clustering.

Much work on clustering of chemical structures has been done by Willett.[2] Willett et al.[3] have a discussion on the use of clustering to select compounds for biological testing. Since he has experimented with and reviewed a variety of methods,

0095-2338/89/1629-0066$01.50/0