compounds which was the motivation of this investigation.

SUMMARY

Our conceptual tools of the excised internal structure, strictly peri-condensed, and formula periodic table for benzenoids has led to our recognition of these constant-isomer series and their topological properties. Given that strictly peri-condensed benzenoids cannot have helicenic isomers or isomers with benzenoid holes (circulene isomers), these isomer numbers have no ambiguity. Even carbon nonradical strictly peri-condensed benzenoids have been speculated to be ultimate pyrolytic constituents, and thus these constant-isomer series represent a relatively more important group.[4,6,7] Strictly peri-condensed benzenoids on the left-hand staircase edge of Table PAH6 form two classes of constant-isomer series: a topologically unique singlet class and a topologically equivalent doublet class. Herein, we specifically claim that our algorithm has generated new isomer numbers, has led to the identification of new constant-isomer series, and has led to the identification of a new topological paradigm that may have universal implications since the polyhex system is a fundamental structure of nature.

REFERENCES

(1) Balaban, A. T.; Kennedy, J.; Quintas, L. *J. Chem. Educ.* **1988**, *65*, 304.
(2) Dias, J. R. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 124; *Can. J. Chem.* **1984**, *62*, 2914; *J. Mol. Struct.* **1986**, *137*, 9.
(3) Stojmenović, I.; Tosic, R.; Doroslovacki, R. in *Graph Theory Proceedings of the Sixth Yugoslav Seminar on Graph Theory*, Dubrovnik, April 18-19, 1985; Tosic, R.; Acketa, D.; Petrovic, V.; Eds.; University of Novi Sad: 1986.
(4) Dias, J. R. *Z. Naturforsch.* **1989**, *44A*, 765.
(5) Dias, J. R. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 61.
(6) Dias, J. R. *Handbook of Polycyclic Hydrocarbons, Parts A and B*; Elsevier: New York, 1987 and 1988.
(7) Dias, J. R. *Theor. Chim. Acta* **1990**, in press.
(8) Dias, J. R. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 15.
(9) Cyvin, S. J.; Brunvoll, *J. Chem. Phys. Lett.* **1989**, *164*, 635. Knop, J. V.; Müller, W. R.; Szymanski, K.; Trinajstić, N. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 159. Knop, J.; Szymanski, K.; Jericevic, Z.; Trinajstić, N. *MATCH* **1984**, *16*, 119-134. Hall, G. G. *Theor. Chim. Acta* **1988**, *73*, 425-435.
(10) Staab, H.; Diederich, F. *Chem. Ber.* **1983**, *116*, 3487.
(11) Funhoff, D.; Staab, H. *Angew. Chem. Int. Ed. Engl.* **1986**, *25*, 742.

The Knowledge-Based Organic Physical Property Data System (KB-OPDS)

ZHIHONG XU, QIAN DONG,* XINJIANG YAN, XIAOXIA LI, and LI GUO

Institute of Chemical Metallurgy, Academia Sinica, P.O. Box 353, Beijing 100080, People's Republic of China

Received January 25, 1990

The purpose of this paper is to report on a new and comprehensive computer program called the Knowledge-Based Organic Physical Property Data System (KB-OPDS), which was developed at the Institute of Chemical Metallurgy, Academia Sinica. This computer system is an effective integration of five functional software packages—Organic Physical Property Database, Prediction Package, Generalized Package of Automation of Group Additive Methods, Organic Structure Information Processing, and Systematic Qualification of Data and Models. The design principles, program functions, and software structure of KB-OPDS are discussed in detail.

Reliable physical property data are essential to many areas of chemistry and chemical engineering including process design, energy conversion, environmental engineering, organic synthesis, semiconductors, and ceramics processing. The development of computerized databases for chemicals and the computer manipulation of predictive methods have grown significantly in recent years. Chemical engineering databases in various forms have been designed by experienced scientists and engineers at many organizations in the world. These include the Physical Property Data Service (PPDS),[1] the DIPPR Pure Component Data Complication,[2] the Engineering Chemistry Data Bases (ECDB),[3] the DECHEMA[4] Thermophysical Property Databank,[5] and the TRC[6] Vapor Pressure Database.[5]

As a result of examining the chemical literature on databases and expert systems, we find that a need still exists for the development of mature data prediction systems. Such systems should evolve from a systematic and continuous effort. To meet the requirements of industry and academia, additional efforts must be made in accurate property prediction, data evaluation, and the linkage of databases to artificial intelligence (AI). The following sections of this paper will discuss two examples of the computerized database systems, the current evolution of combining AI technology with databases, and database developments in China.

DATABASE SYSTEMS AND DATABASE SYSTEM DEVELOPMENTS

Two Successful Database Systems. The hallmark of a database system is the combination of its reliability and practicality. One system of this type is PPDS. A widely used software system, it is a joint effort of the National Engineering Laboratory and the Institution of Chemical Engineers (U.K.). It is designed to provide reliable data for engineers involved in the chemical, petrochemical, and process industries. The system contains (a) a pure component database of thermal physical properties for 870 compounds and a program that calculates 18 constant properties and 20 variable properties for mixtures of up to 20 components, (b) a package for certain special purposes of calculating thermal physical data on liquid-phase salts and acid solutions, and (c) a user interface dealing with the direct entry of the user's data, regression of phase equilibrium data, and property prediction; this can help users obtain useful data even if their compound is not in the PPDS database. A feature of the PPDS system is that it offers engineers more than one predictive model for each property. It also permits manual or automatic selection of an appropriate model.

Another important database is the Pure Component Data Compilation under the sponsorship of the Design Institute for Physical Property Data (DIPPR) of the American Institute of Chemical Engineers (AIChE). The DIPPR database now provides 26 constant properties and 13 temperature-dependent properties of the 1023 compounds. One major feature of this project is quality control of data in the database. In the DIPPR Data Compilation, two quality codes are used to designate the quality level of data. One is for each property, while the other is for the correlation coefficient. This is of considerable assistance to users because it provides them with an effective method for selecting data from the database and

0095-2338/90/1630-0256$02.50/0

offers a professional opinion of its quality.

Artificial Intelligence Application in Database System. The First International Workshop on Expert Databases was held in the U.S. in 1984. It marked a new stage of the study in database development with the introduction of and partial integration with artificial intelligence.[7] The significance of this stage was to strengthen efforts in knowledge processing. Current data handling techniques and expert systems are not satisfactory for many major applications. On the one hand, expert systems in existence may employ hundreds and thousands of rules of scientific knowledge but are deficient in data storage and retrieval, particularly in handling large quantities of data; on the other hand, some data management systems, although capable of processing numerous data, lack in their knowledge processing ability. Therefore, an approach to merge the two technologies is proposed below to create a more complete knowledge-based data system.

In 1983, researchers at Carnegie-Mellon University developed an expert system, CONPHYDE, for the purpose of assisting engineers in making decisions on "best" models for gas–liquid equilibrium calculations.[8] This is usually regarded as the first step in a knowledge-based chemical thermodynamic data system. The framework of the existing expert system PROSPECTOR, which is designed for exploring hard rock minerals, was adopted in CONPHYDE to present thermodynamic knowledge. At the AIChE 1987 annual meeting, e b Kelly introduced another expert system that dealt with thermodynamic model selection and experimental data evaluation.[9] It is similar to the CONPHYDE system in its rules of knowledge representation and processing.

Recently, a paper titled "A Knowledge Based System for the Selection of Thermodynamic Models" was published in *Computers and Chemical Engineering*.[10] The goal of this work was to imitate human experts in this field in choosing the best calculation methods for a specific engineering problem.

Database Development in China. In China, there has been considerable activity in the development of chemistry databases. Examples of databases in physical property study are the following: (1) The Chemical Engineering Physical Property Data System was built at the Beijing Institute of Chemical Engineering and contained over 5000 organic and inorganic compounds. (2) The New Physical Property Data System developed at the Dalian Institute of Chemical Engineering on IBM PCs with the unique feature of functional group analysis along with the automatic selection of best methods. (3) The Tianjing Chemical Engineering Database, another data system on IBM PCs, was released at Tianjing University in 1989. Some of the most current predictive methods were adopted in that system.

ECDB has been under development at the Institute of Chemical Metallurgy (ICM), Academia Sinica, for 10 years and has been financed by the Chinese National Science Foundation and the Chinese National Five-Year Plan for 1986–1990. It is a cluster of highly reliable databases and software created for the chemical and petrochemical industry and contains five databases: Inorganic Thermochemistry Data Base, Non-Electrolyte Vapor–Liquid Equilibrium Data Base, Aqueous Thermodynamic Data Base, Program Package for Calculation in Engineering Chemistry, and a Chemistry Library.[11] Recently, ICM has released two new software packages—the Inorganic Property Estimation system[12] and the Knowledge-Based Organic Physical Property Data System (KB-OPDS). They represent a new effort in the integration of AI technology with database developments at ICM.

This paper focuses on the presentation of the design principle, system architecture, and software functions as well as the artificial intelligence and graph theory application in KB-OPDS.

SYSTEM ARCHITECTURE

(1) Design Principle. KB-OPDS is mainly designed to provide high-quality data for the study and development of processes in chemical and petrochemical engineering and to assist engineers in the selection of appropriate predictive methods. In addition, KB-OPDS offers detailed structural information for organic compounds to develop new predictive methods. The primary goals of this project are as follows:

(a) A database should be built consisting of a large number of chemical compounds and physical properties to meet the basic needs of the Chinese chemical industry. The database should contain multiple retrieval schemes that facilitate data searching.

(b) A data system should be provided to users with applicable and precise predictive methods when experimental data for chemical thermodynamic properties are not available. In general, due to their simplicity and practicality, group additive methods play an important role in the estimation of physical and chemical thermodynamic properties. However, the methods demand that the users be able to analyze chemical structure and understand the relationship between property contribution and substructure elements. Therefore, generalized software for automatic analysis of functional groups suitable to diverse group additive methods is one of the essential components of KB-OPDS.

(c) A software package should be created for organic structure information processing with a view toward improving group additive methods and other predictive methods. Although some group additive methods have wide applicability, their accuracy is sometimes limited because they only consider the contribution to a property from individual groups within a molecule, without regard to the effect of nearest- or next-nearest-neighbor interactions and stereoisomerism.

(d) Data should be tested and evaluated either from the database or from other data sources to maintain good quality control of the database. This approach has difficulties not only because of the huge quantities of data but also because of limitations in computerized methods for data evaluation. A long-range goal at ICM is to develop practical schemes for assessing the quality of models and data by computer to increase the reliability of KB-OPDS.

(2) Overall Structure of KB-OPDS. As shown in Figure 1, KB-OPDS is an effective integration of eight components: (1) a data bank, (2) a knowledge base, (3) a data management program, (4) a property prediction program, (5) a software program for automation of group additive methods, (6) a program dealing with organic structure processing, (7) a package for systematic qualification of data and models, and (8) a menu-driven user interface.

METHODOLOGIES USED FOR CREATING KB-OPDS

Some advanced concepts and techniques, such as artificial intelligence, knowledge engineering, graph theory, and modular design of software, were adopted in the course of establishing KB-OPDS, so it has the elements of computation, inference, analysis, and judgement. It contains a large number of properties and predictive methods representing the major achievements of the study on chemical physical property techniques in the 1980s. Much of the effort devoted to the creation of KB-OPDS can be regarded as a preliminary attempt to combine the database with the knowledge base and to set up a comprehensive computer system with the overall capability of chemistry knowledge processing.

(1) Knowledge Engineering Application. To begin with, we explored a large number of literature sources reporting chemical and physical properties as well as the molecular structure of compounds. Then, we examined various predictive techniques and experimental data and consulted with many

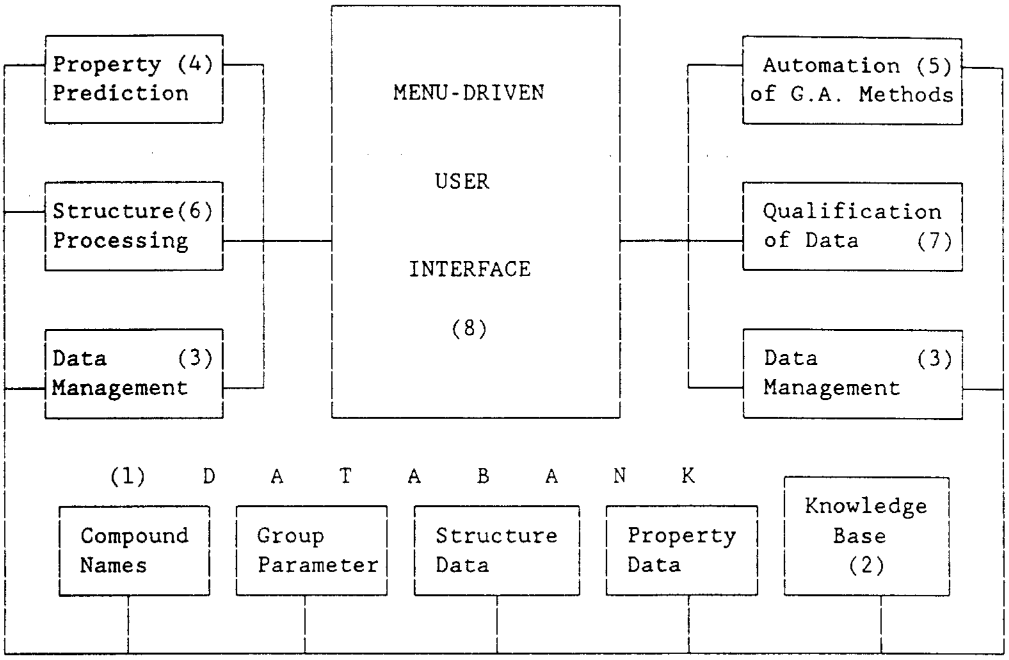

Figure 1. Outline of KB-OPDS.

experts. Consequently, a great deal of specialized knowledge and experience was captured so that a knowledge base and an inference mechanism could be constructed.

A primary focus of the knowledge engineering application in KB-OPDS was the knowledge representation of molecular structures. All domain-specific knowledge (e.g., compound structures, compound classification, principles of each group additive methods) was encoded in a knowledge base. A series of sets of alphanumeric symbols called "coding sets" were defined for the description of the domain-specific knowledge. Some of the coding sets devised for use in KB-OPDS for the manipulation of organic structural information are (a) The Organic Compound Coding Specially Designed for Chemical Engineering (OCSCE),[13] a chemical line rotation for chemical engineers to communicate with the KB-OPDS system; (b) a main group coding set that describes groups, fragments, and general molecular structure information; (c) a standardized substructure coding set that specifies groups and fragments for each group additive method; (d) a structural information coding set that represents the stereo structure effect of molecules; and (e) another coding set that is used for compound classification.

The inference mechanism is composed of hundreds of "IF THEN" rules and substructure matching strategies with alternative searching algorithms for cyclic and acyclic structures. The main function of the inference mechanism is to determine how the stored knowledge is to be used in parsing manipulation, group analysis, and substructure matching.

The knowledge base and inference mechanism have permitted the full automation of the entire process of organic structure analisis. The analysis starts with the parsing of the OCSCE coded formula. The first step produces the main group codes. This is followed by a step that determines adjacency matrices for the molecule. The final step matches and generates the specified fragment codes for each group additive method.

(2) Graph Theory Application. The nature of graph theory is such that it permits the study of the relationship between vertices and edges which constitute an object. As organic molecular structures are graphical objects in the above context, they can be studied and analyzed by using graph-theoretical tools. Most chemical engineering data systems contain two parts: a database and a prediction program. The disadvantage of such systems is that they are unable to take full advantage of compound structural information. As a result, the system capabilities are limited in structure and substructure match, automatic analysis of functional groups, online parameter processing for group additive methods, and the acquisition of stereo structural information. To enhance KB-OPDS, a program package for organic structure information processing has been constructed on the basis of graph theory. The main function of the program is to analyze structure information identified from compounds with the aid of connection tables, adjacency matrices, distance matrices, and numerical chromatic distance matrices.

(3) Computer-Aided Evaluation of Data and Models. The key to the establishment of database reliability lies with data evaluation. Data evaluation is still at a stage of development where it can be done only by experts in the field. We implement three methods for computer-aided evaluation of physical property data: (a) data error checking by the principles of chemical thermodynamics; (b) outlier checking according to the magnitude of a property; and (c) cross evaluation of data and modules from various predictive models.

(4) Modular Design of Software Packages. The concepts of software engineering involve extensive applicability, user friendliness, modular design, functional integrity, and shared data management in the software development. Being a large, comprehensive software system with many functions and many thousands of lines of source code, a great deal of consideration at the design stage for KB-OPDS was given to effective programming by introducing a modular design concept. One approach of the developing KB-OPDS is that, first, each subsystem could be constructed as a stand-alone package dealing with a specific subject, and then an integrated system would be created from those independent packages. Another approach is to implement standardized module design in program structure depending on similar procedures in the property calculation.

SUBSYSTEMS IN KB-OPDS

The eight components of the KB-OPDS system form five functional subsystems, which will be briefly described in the following sections. On the one hand, the subsystems of KB-OPDS are stand-alone systems, each fulfilling a particular task such as data retrieval, property prediction, and automation

Table I. Retrievable Data

	data items		data items
1	boiling point	21	flash point
2	melting point	22	flammability limit
3	freezing point	23	autoignition temperature
4	critical temperature	24	refractive index
5	critical pressure	25	solubility parameter
6	critical volume	26	radius of gyration
7	critical compressibility	27	solid density
8	acentric factor	28	liquid density
9	dipole moment	29	saturated pressure
10	standard enthalpy of formation	30	enthalpy of vaporization
11	standard Gibbs energy	31	solid heat capacity
12	absolute entropy	32	liquid heat capacity
13	liquid density	33	ideal gas heat capacity
14	liquid molar volume	34	second virial coefficient
15	triple point temperature	35	liquid viscosity
16	triple point pressure	36	vapor viscosity
17	enthalpy of fusion-melt Pt	37	liquid thermal conductivity
18	ent enthalpy of combustion	38	vapor thermal conductivity
19	van der Waals volume	39	surface tension
20	van der Waals area		

Table II. Retrievable Characteristics

1	toxicity	11	narcotic
2	irritant	12	carcinogen
3	insolubility	13	fumigant
4	reactivity	14	lachrymator
5	oxidant	15	unpleasant odor
6	flammability	16	hygroscopicity
7	explosivity	17	polymer
8	solubility	18	intermediate
9	corrosivity	19	solvent
10	refrigerant or extinguishant	20	decomposability
		21	others

Table III. Fundamental Properties

symbol	name of property
T_b	normal boiling point
T_f	normal freezing point
T_c	critical temperature
P_c	critical pressure
V_c	critical volume
Z_c	critical compressibility
ω	eccentric factor
D_{ipm}	dipole moment
H_{vb}	enthalpy of vaporization at boiling point
V_{lb}	liquid volume at boiling point
$H_f{}^{\circ}{}_{298}$	standard enthalpy of formation
$G^{\circ}{}_{298}$	standard Gibbs energy

Table IV. Thermal Properties

symbol	name of property
P_{vp}	liquid saturated vapor pressure
P_{vp}(s)	solid saturated vapor pressure
C_{pL}	pure liquid heat capacity
H_v	pure liquid vaporization heat
L_{den}(Vs)	saturated liquid density
L_{den}	supercooled liquid density

Table V. Thermodynamic Properties

symbol	name of property
$C_p{}^{\circ}$	ideal gas heat capacity
$H_f{}^{\circ}$	ideal gas enthalpy difference
G°	ideal gas Gibbs energy difference
V	fluid volume
Z	compressibility factor
f/P	coefficient factor of fugacity of fluid
$H - H^{\circ}$	enthalpy difference of fluid
$S - S^{\circ}$	entropy difference of fluid
$C_p - C_p{}^{\circ}$	heat capacity difference of fluid
γ	activity coefficient of mixture

Table VI. Transport Properties

symbol	name of property
VISG-LP	low-pressure pure gas viscosity
VISG-P	high-pressure pure gas viscosity
VISGM-LP	low-pressure gas mixture viscosity
VISGM-P	high-pressure gas mixture viscosity
VISL-LT	low-temperature pure liquid viscosity
VISL-T	high-temperature liquid mixture viscosity
VISL	high-pressure pure liquid viscosity
TCDG-LP	low-pressure pure gas thermal conductivity
TCDG-P	high-pressure pure gas thermal conductivity
TCDGM-LP	low-pressure gas mixture thermal conductivity
TCDGM-P	high-pressure gas mixture thermal conductivity
TCDL	pure-liquid thermal conductivity
ST	pure-liquid surface pressure

of group additive methods, while, on the other hand, the subsystems constitute an integrated software system by connecting all programs through the master process and subprocess[14] generated at run time. These enable users to access KB-OPDS in two ways, either by directly accessing any one of the subsystems or by entering the user interface through interactive mode and then accessing each subsystem.

The source code of the KB-OPDS system is written in FORTRAN 77. It contains nearly 600 subroutines with a total of approximately 60000 FORTRAN statements. The system operates on VAX 11/780 computers, and the data management software adopted in the database is Rdb/VMS,[15] which is commercially available on VAX 11/780 computers. The system is currently available from ICM.

(1) Organic Physical Property Data Base (OPDB).[16] OPDB is a collection of the following data sources: (1) *The Properties of Gases and Liquids* (Reid);[17] (2) *Handbook of the Thermodynamics of Organic Compounds* (HTOC);[18] (3) *Critical Data of Pure Substances* (DECHEMA);[19] (4) *Data Compilation Tables of Properties of Pure Compounds* (DIPPR);[2] and (5) *Thermodynamic Data for Pure Compounds* (TDPC).[20]

A reorganization of the above data sources was conducted before data were loaded into the database. A master index table was generated, composed of 5739 chemical compounds. Each compound appears only once in the table. In addition, the master index table contains the following 11 fields describing each compound: field 1, KB-OPDS sequence number; field 2, Chemical Abstracts Service Registry Number; field 3, molecular formula; field 4, compound name; field 5, OCSCE structural code; fields 6–11, points for each data source. On the basis of the master index table, a complete data set, which includes 12 property data items for 5739 compounds, is produced as recommended data from OPDB.

Rdb/VMS is used in OPDB for data management, such as data storage, data retrieval, and data uptake. Its high level language calling interface facilitates the user's retrieval operations in conjunction with the FORTRAN program. Six searching paths provided by the database enable users to locate certain compounds and their relevant property information from CAS Registry Number, compound name, formula, and OCSCE structural code as well as characteristic code indicating compound characteristics, i.e., toxicity, solubility, explosivity, flammability, etc. In addition, users may find other properties for some compounds by seeking those compounds for which the value of their properties ought to agree with a given range. In searching information of interest in the database, users can work more effectively by entering up to three retrieval requirements for the computer at a time.

There are 39 kinds of physical property data (Table I) and 21 items relating chemical characteristics for organic compounds (Table II) available for the users of OPDB.

(2) Physical Property Estimation Program (PPEP).[21] Approximately 25 properties in 4 categories may be estimated in the PPEP program. The categories are fundamental properties, thermal properties, thermodynamic properties, and transport properties. These are listed in Tables III–VI. When selecting predictive techniques for PPEP, we centered our concern on the precision and applicability of the methods in the light of recommendations by the literature.[10]

It is well-known that a single predictive method is not applicable to all kinds of substances because of difference between polar and nonpolar substances. So, a method is first taken here to classify all compounds into four groups in accordance to their dipole moment value (DMV), i.e., nonpolar and weak polar compounds (DMV = 0.0–0.5), regular polar compounds (DMV = 0.5–1.5), extremely polar compounds (DMV > 1.5), and compounds with unknown polarity. A further step is to divide the predictive methods in PPEP into four classes on te basis of their applicability toward the compounds with the different DMV.

To help users to select an appropriate method and to reduce input requirements for users, most substances with unknown dipole moment value in the OPDB are assigned a polarity in the light of the relation between the polarity and the compound structure along with the concept that the compounds with similar structure could be in accord with those having the same polarity. So, when calculating a property, PPEP offers users a function of automatically displaying a menu of predictive methods suitable to a group of compounds of the same polarity.

Two approaches are used in PPEP for helping engineers gain a better understanding of predictive methods involved in the system so as to use methods appropriate to their specific problems. One of the approaches is to supply users with online help information concerning the application range, input data, and average calculated deviation as well as references for each technique. In the second approach, PPEP sets a flag code that tells the program to monitor the entire property calculation process and report error messages in detail. When the property prediction is successful, a flag code appears in the results file showing that the calculation was completed. When the prediction is not successful, the results file indicates error messages, such as insufficient input of data and no group contribution.

(3) Generalized Program of Automation of Group Additive Methods (GPAG).[22] GPAG is a software package specifically designed to compute physical properties through various group additive methods. The system acts as follows: when it starts, users are only required to enter OCSCE coded formulas; GPAG is capable of automatically analyzing functional groups of molecules and extracting structure information from OCSCE coded formulas and finally carrying out the property calculation. In addition, it has the ability to offer analyzed group codes, the number of groups and group parameters for the user's own program. The package currently involves 21 group additive methods that calculate 16 physical properties (Table VII).

One of GPAG's unique features is the suitability of its methodologies for different kinds of group additive methods. GPAG covers the majority of group additive methods in the field of thermodynamics. Since the principles of the group additive methods and organic structural information can be encoded by generalized representation methods, the system can be easily updated, with a minimum of programming, by the addition of a new group additive method, the elimination of an obsolete method, or the modification of an existing method.

The GPAG program can be executed in a flexible manner, either acting as an independent property prediction package or integrating with PPEP without modification of the source codes.

Table VII. Methods and Properties in GPAG

pointer	methods[a]	properties
1	Fedors	critical temperature
2	Lebas	liquid molar volume at NBT
3	Missenard	liquid heat capacity
4	Reichenberg	low-pressure gas viscosity
5	Macloed–Suden	pure liquid surface tension
6	Chueh–Swanson	liquid heat capacity
7	Schroeder	liquid molar volume at NBT
8	Bondi	liquid volume
9	Lyderson	critical constants
10	Morris	pure liquid viscosity
11	Souder	pure liquid viscosity
12	Orrick–Erbar	low-temperature liquid viscosity
13	Thomas	pure liquid viscosity
14	UNIFAC	activity coefficient
15	Veter	critical volume
16	Tyn–Calus	binary liquid diffusion coefficient at infinite dilution
17	Joback	critical constants, boiling point, melting point, gas heat capacity, standard enthalpy of formation at normal pressure
18	Ambrose	critical constants
19	Benson	standard enthalpy of formation and heat capacity
20	Pedley	standard enthalpy of formation
21	Rihani–Doraiswamy	gas heat capacity

[a] Reid et al. *The Properties of Gases and Liquids*, 4th ed.; McGraw-Hill: New York, 1987.

(4) Organic Structure Information Processing (OSIP).[23] Briefly, the function of the OSIP program is, in the light of graph theory application in chemistry,[21] to capture useful structural information from the compound structure for the KB-OPDS system. The program is equipped with the following four functions.

(a) The program supplies GPAG with structural and stereochemical information on molecules such as cyclic or alicyclic, cis or trans, and a fragment connected to a chain or to a ring and so on.

(b) OSIP suggests a new scheme of structure matching for organic molecules. On the basis of the chromatic graph and the distance matrix, a chromatic distance matrix matching algorithm is proposed for quickly and accurately conducting a compound match in data retrieval.

(c) The program advances a judging algorithm of the nearest similarity in adjacent connection for organic molecular similarity, i.e., the program uses this algorithm to judge if two molecules coincide in their structures and what percentages of complete coincidence can be found among all the possible fragments within two molecules. This function tends to be a powerful tool for property study; for example, when the properties of an unknown substance are predicted by the group additive method, the contribution of unknown groups in the compound should be derived from those in the compounds with the most similar structure.

(d) OSIP contains several topological indices,[24] such as the Wiener distance index, the Randič connectivity index, the Balaban average distance sum connectivity index, and the Platt index. When users input OCSCE codes or a connectivity table for a molecule, the program will output topological indices for given compounds. Additionally, there is a property predictive method, combining topological indices with group additive methods, created in OSIP. In this method, the critical temperature, T_c, can be estimated without the need of the normal boiling point, and the results derived match well with other accurate methods for critical properties.

(5) Systematic Qualification of Data and Models (SQDM).[25] SQDM is an attempt to develop the useful evaluation techniques in various ways under the support of KB-OPDS. With the aid of SQDM, data in OPDB and models in or out of

PPEP have been tested and evaluated systematically and have provided the ability to improve the reliability of data and models in the KB-OPDS system. The three methods employed in SQDM are as follows.

(a) Outlier Checking According to the Magnitude of a Property. The concept behind this method is that a property value of a compound is generally within a definite range, and if the value deviates a great deal from its normal range, it is suspect. By use of the VAX-11/780 computer, the data file that contains property data to be tested is sorted by using the SORT command in Digital Command Language,[26] and then the result file from SORT can be used for checking the magnitude of data through rewriting it with the "F" format in the FORTRAN program. In doing so, errors will occur and can be located when outliers appear in the rewriting.

(b) Qualification of Data and Methods on the Basis of Basic Theories. An important way of checking data in SQDM is to use the basic theories in chemistry as a criterion; for example, "the vapor pressure of a saturated liquid at normal boiling point is 1 atm" may be taken as a criterion of data check. By comparing the calculated value of vapor pressure at the normal boiling point (T_b) with the values of T_b from different sources, T_b of C_4H_8O (CAs Registry Number 109-99-9) in HTOC is found to be too large [P_{vp}(bar) = 5.102 at 399.15 K)], while the value in the literature[10] is P_{vp}(bar) = 0.9771 at 338.0 K and the value in DECHEMA is P_{vp}(bar) = 0.9987 at 338.67 K.

(c) Cross-Checking Data by Comparing Predicted Data with Experimental Data. The predictive models of temperature-dependent properties were systematically tested against experimental data at a number of temperature points, and SQDM imitates what an expert may consider and propose for the expression of test results. This facilitates a vast amount of data testing for critical properties.

SUMMARY

For the purpose of making this paper more interesting and informative, an example of the use of the prediction system of KB-OPDS is provided (Appendix 1) as is a description of the format in the User Input File (Appendix 2).

For KB-OPDS to accomplish the anticipated goals of this project, a four-year effort by the members of the Organic Physical Property Data Center at ICM has been made on the above aspects. The project has been sponsored by the Ministry of Chemical Industry, China, and financed by the Chinese National Five-Year Plan for 1986–1990 as well.

The hope is that KB-OPDS will be accepted as a practical tool that reduces substantially the time and effort engineers and chemists currently require to obtain reliable physical property data. It may also serve as a basis to study more accurate predictive methods of the group contribution type. From a long-term point of view, KB-OPDS is a phased software product that needs to be continuously improved and up-dated in the following ways: (a) create a knowledge-based package to aid users in selecting the best suitable predictive methods; (b) develop an analytical scheme to accommodate physical property estimation for complex problems; (c) advance new techniques for physical property estimation; (d) establish a package for online parameter processing of group additive methods; and (e) develop and enhance practical computer techniques for data evaluation.

ACKNOWLEDGMENT

Contributions to data input by Liguang Chen and Meishan Peng at ICM are gratefully acknowledged. We also appreciate the cooperation provided by the Beijing Institute of Chemical Engineering during this development effort and the discussion with Dr. Eugene S. Domalski and Dr. David B. Neumann at the National Institute of Standards and Technology during the preparation of this paper.

APPENDIX 1. EXAMPLE OF THE USE OF THE PREDICTION SYSTEM OF KB-OPDS

The prediction of the critical temperature for five compounds, e.g., 1-propanol, 1-butanol, 1-pentanol, 1-hexanol, and 1-heptanol, from KB-OPDS is illustrated here. The corresponding experimental values are 516.25, 536.75, 562.95, 586.15, and 610.15, respectively. User input is indicated in boldface. Once KB-OPDS is started, the first screen will display as follows:

```
                  Knowledge Based
        Organic Physical Property Data System
               KB-OPDS VERSION 1.0
                   PREPARED BY
        The Institute of Chemical Metallurgy
                 Academia Sinica
                  Beijing, China
```

Then the system prompts the user for the mode of input. Input mode 1 allows the user to interactively input specific information, i.e., compound system, number of compounds, given temperature and pressure points, phase state of compound system, etc., under the prompting of the KB-OPDS system. In input mode 2, the user is permitted to edit the above data to a specific input file, User Input File. See Appendix 2 about the details of the file.

```
INPUT MODE
  1 = Interactive Mode
  2 = User Input File Mode
INPUT MODE? 2
```

In the following screen the user is prompted for files in which the experimental data and the constants required by calculation may be provided either by KB-OPDS or by the user. File 2 is the same sort of file as the UIF file.

```
DATA SOURCE FILE
  File 1 = Data from KB-OPDS
  File 2 = Data provided by users
DATA SOURCE FILE? (1,2) 1
```

The next screen presents options for the prediction. Option 1 is selected as critical temperature belongs to the category of fundamental property.

```
All properties Classification:
  1. Fundamental Properties
  2. Thermal Properties
  3. Thermodynamic Properties
  4. Transport Properties
  0. Property List of Each Class
PROPERTY SELECTION: 1
```

Under options 1–4 the fourth screen shows detailed properties in fundamental property category.

```
 Fundamental Property Classification:
 1. Normal Boiling Point
 2. Normal Freezing Point
 3. Critical Temperature
 4. Critical Pressure
 5. Critical Volume
 6. Critical Compressibility
 7. Acentric Factor
 8. Dipole Moment
 9. Enthalpy of Vaporization at Boiling Point
10. Liquid Volume at Boiling Point
11. Standard Enthalpy of Formation
```

```
    12. Standard Gibbs Energy
    SELECTION: 3
```

According to the input from the User Input File and the above selections, KB-OPDS automatically displays a menu of relative predictive methods. The user is given a chance to select any one of those methods; Option 0 in this screen enables the user to understand more about those methods.

```
    Property: Critical Temperature
    Estimation Methods Available:
      1. Ambrose (Group)
      2. Joback (Group)
      3. Fedors (Group)
      4. Klincewitz
      0. More information on each method
    FOR COMPONENT (1) [CODE=2146]
    FOR COMPONENT (2) [CODE=2563]
    FOR COMPONENT (3) [CODE=2733]
    FOR COMPONENT (4) [CODE=3326]
    FOR COMPONENT (5) [CODE=3828]
    INPUT YOUR CHOICE: 1
```

The prediction is performed at this point, and the outcome is shown on the screen and stored in the output file, OPDS.OUT.

```
          PROPERTY: Critical Temperature Tc
              METHOD: Ambrose (Group)
Comp. Code     Tc (K)     Calc. Status
2146           512.30     Normal
2563           536.50     Normal
2733           562.80     Normal
3326           588.00     Normal
3828           610.80     Normal
    CONTINUE WITH Tc? (Y/N[Y])
```

KB-OPDS provides a default option for continuing with Tc by other alternate methods; otherwise, a new property prediction can be done if the user goes back to the Property Classification Menu and has another choice of properties. Here are the results of Tc from the JOBACK method.

```
          PROPERTY: Critical Temperature Tc
    METHOD: JOBACK (Group)
Comp. Code     Tc (K)     Calc. Status
2146           520.0      Normal
2563           536.90     Normal
2733           556.10     Normal
3326           574.60     Normal
3828           591.90     Normal
```

The process can be continued for other predictions or terminated.

```
    CONTINUE WITH Tc? (Y/N[Y]) N
    CONTINUE WITH OTHER PROPERTIES? (Y/
    N[Y]) N
    Do you want to KEEP your input file? (Y/N[Y])
    Welcome to KB-OPDS again !
    Output file is OPDS.OUT
```

APPENDIX 2. DESCRIPTION OF THE FORMAT IN THE USER INPUT FILE

The UIF file is composed of a series of explanations and prompts. A number of designated places in the file are left for users to enter their responses to the prompts. The file is automatically generated by the KB-OPDS system in response to an option of input mode 2 when an UIF file does not exist. In the next step, KB-OPDS switches to VMS editor and enables users to give their answers following the sign → in the file. The program system will return from VMS editor and continue after the file is done. The input for the above example is shown here in boldface.

```
       Items of Input                               Description

 SYSTEM CODE:                                system type of the compound to
 1 = Pure System  2 = Mixture System         be predicted
       SYSTEM CODE --> 1


 NO. COMPS (<=10):                           number of compounds for each
 NO. COMPS = Number of Compounds             prediction is limited to 10.
      NO. COMPS --> 5


--- FOR PURE SYSTEM                         code of each pure substance, CODE
COMPS:                                      defined by the KB-OPDS system,
CODE COMPS = Code of each compound          may be obtained from the
      CODE COMPS 1 --> 2146                 database of KB-OPDS or from the
      CODE COMPS 2 --> 2563                 KB-OPDS User's Menu.
      CODE COMPS 3 --> 2733
      CODE COMPS 4 --> 3326
      CODE COMPS 5 --> 3828
      ......


--- FOR MIXTURE SYSTEM                      code and mole fraction for a CODE.
FRAC. COMP:                                 component in the mixture
Use comma to separate CODE and FRAC
      CODE, FRAC. COMP 1 -->
      CODE, FRAC. COMP 2 -->
      CODE, FRAC. COMP 3 -->
      CODE, FRAC. COMP 4 -->
      CODE, FRAC. COMP 5 -->
      ......


NO. T POINTS:                               number of temperature points to
If independent of T, enter 0.               be calculated is limited to 10
NO. T POINTS --> 0                          for each process.


T EACH POINT:                               value of each temperature point
UNIT: K
      T POINT 1 -->
      T POINT 2 -->
      T POINT 3 -->
      T POINT 4 -->
      T POINT 5 -->
      ......


NO. P POINTS:                               number of pressure points to be
If independent of P, enter 0.               calculated is limited to 10 for
each  NO. P POINTS --> 0                    process.


P EACH POINT:                               value of each pressure point
UNIT: bar
      P POINT 1 -->
      P POINT 2 -->
      P POINT 3 -->
      P POINT 4 -->
      P POINT 5 -->
      ......


PHASE STATE:                                a given phase state for each
1 = GAS  2 = LIQUID  3 = SATURATED          process
4 = SOLID  0 = UNKNOWN
      PHASE STATE --> 1


COMP. POL.                                  the indication of polarity of a
1 = Nonpolar or Weak Polar Compound         compound in a pure system
2 = Polar Compound                          either by the user or by
3 = Extremely Polar Compound                KB-OPDS
4 = Unknown
5 = Default Value of polarity by KB-OPDS
      For Compound 1, COMP. POL. -->
      For Compound 2, COMP. POL. -->
      For Compound 3, COMP. POL. -->
      For Compound 4, COMP. POL. -->
      For Compound 5, COMP. POL. -->
      ......


SYS. POL.                                   the indication of polarity of
1 = Nonpolar or Weak Polar compound         the mixture system only by the
2 = Polar Compound                          user
3 = Extremely Polar Compound
4 = Unknown
5 = Default Value of polarity by KB-OPDS
      For the Mixture, SYS. POL. -->
```

REFERENCES AND NOTES

(1) Edmonds, B. PPDS—Physical Property Data Service, Thermodynamic Databases. *CODATA Bull.* **1985**, *No. 58*, 6.

(2) Danner, R. P.; Daubert, T. E. The DIPPR Data Compilation Project, Thermodynamic Databases. *CODATA Bull.* **1985**, *No. 58*, 14.

(3) Xu, Z.; Wang, L. *Anal. Chim. Acta* **1988**, *210*, 115-121.

(4) The Deutsche Gesellschaft fur Chemisches Apparatewesen, Chemische Technik and Biotechnologie e.V. (DECHEMA), Frankfurt, Germany.

(5) Available from Technical Database Services, Inc., 10 Columbus Circle, New York, NY 10019.

(6) The Thermodynamic Research Center, part of the Texas Engineering Experiment Station, The Texas A&M University System.

(7) Brodie, M. L.; Jarke, M. On Integrating Logic Programming and Databases. *Proceedings of the First International Workshop on Expert Databases*; Benjamin/Cummings Publishing Co., Inc.: Menlo Park, CA, 1986.

(8) Alcantara, R. B.; Westerberg, A. W.; Rychener, M. D. *Comput. Chem. Eng.* **1985**, *9*, 127-142.

(9) Kelly, E. B.; Holste, J. C.; Hall, K. R. AIChE 1987 Annual Meeting, Session 138.

(10) Gani, R.; O'Connell, J. P. *Comput. Chem. Eng.* **1989**, *13*, 397-404.

(11) Dong, Q.; Xu, Z.; Wang, P. *Anal. Chim. Acta* **1988**, *210*, 181-187.

(12) Huang, G. Doctoral Thesis, the Institute of Chemical Metallurgy, Academia Sinica, Beijing 100080, China, 1989.

(13) Xu, Z.; Sun, R.; Dong, Q.; Yan, X. *Comput. Appl. Chem. (China)* **1989**, *6* (2), 1-5.

(14) *VAX/VMS Run Time Library Reference*; Documentation Publishing Center of 2000 Series: Beijing, 1987.

(15) *VAX Rdb/VMS database design and definition*; Documentation Publishing Center of 2000 Series: Beijing, 1987.

(16) Guo, L., et al. A database subsystem of KB-OPDS. In *KB-OPDS*, Research Report to the Chinese Academy of Sciences and the Ministry of Chemical Industry, China (Aug 1989) pp 21-40. Available from the Institute of Chemical Metallurgy, Academia Sinica, P.O. Box 353, Beijing, China.

(17) Reid, R. C., et al. *The Properties of Gases and Liquids*, 4th ed.; McGraw-Hill: New York, 1987.

(18) Stephenson, R. M.; Malanowski, S. *Handbook of the Thermodynamics of Organic Compounds*; Elsevier Science Publishers B.V.: Amsterdam, The Netherlands, 1987.

(19) Simmrock, K. H.; Janowsky, R.; Ohnsorge, A. *Critical Data of Pure Substances (Ag - C7)*; Chemistry Data Series; DECHEMA: Frankfurt, 1986; Vol. II, Part 1.

(20) Smith, B. D.; Srivastava, R. *Thermodynamic Data for Pure Compounds*; Elsevier Science Publishers B.V.: Amsterdam, The Netherlands, 1986, Parts A and B.

(21) Li, X., et al. A Prediction Subsystem of KB-OPDS. In KB-OPDS, Research Report to the Chinese Academy of Sciences and the Ministry of Chemical Industry, China (Aug 1989) pp 41-68. Available from the Institute of Chemical Metallurgy, Academia Sinica, P.O. Box 353, Beijing, China.

(22) Dong, Q., et al. A Generalized Subsystem of Automation of Group Additive Methods. In *KB-OPDS*, Research Report to the Chinese Academy of Sciences and the Ministry of Chemical Industry, China (Aug 1989) pp 70-86. Available from the Institute of Chemical Metallurgy, Academia Sinica, P.O. Box 353, Beijing, China.

(23) Yan, X., et al. A Subsystem for Organic Structure Information Processing. In *KB-OPDS*, Research Report to the Chinese Academy of Sciences and the Ministry of Chemical Industry, China (Aug 1989) pp 88-125. Available from the Institute of Chemical Metallurgy, Academia Sinica, P.O. Box 353, Beijing, China.

(24) King, R. B.; Rouvray, D. H. *Graph Theory and Topology in Chemistry*; Elsevier Science Publishers B.V.: Amsterdam, The Netherlands, 1987.

(25) Li, X., et al. Systematic Qualification of data and Models. In *KB-O-PDS* Research Report to the Chinese Academy of Sciences and the Ministry of Chemical Industry, China (Aug 1989) pp 126-147. Available from the Institute of Chemical Metallurgy, Academia Sinica, P.O. Box 353, Beijing, China.

(26) *VAX Digital Command Language Reference*; Documentation Publishing Center of 2000 Series: Beijing, China, 1987.

Computational Techniques for Vertex Partitioning of Graphs

XIAOYU LIU, K. BALASUBRAMANIAN,[*,‡] and M. E. MUNK[*]

Department of Chemistry, Arizona State University, Tempe, Arizona 85287-1604

Received March 5, 1990

A powerful vertex-partitioning algorithm is developed and applied for vertex partitioning of graphs of chemical and spectroscopic interest. The codes developed on the basis of these algorithms are tested and compared for performance with other methods based on the Morgan algorithm and the principal eigenvector algorithm based on the Givens-Householder method. The newly developed algorithm and codes appear to be more powerful than the Morgan and the principal eigenvector algorithms for vertex partitioning of graphs.

INTRODUCTION

The graph-vertex-automorphism partitioning problem has received considerable attention in recent years.[1-12] The vertex partitioning is of fundamental interest since it has many practical applications. First and foremost of all, it provides a solution for computer perception of the hidden topological symmetry of a molecule. In our group we have been interested in building a comprehensive computer-assisted structure-elucidation system.[11,12] A critical problem encountered in this work is that given the neighborhood table (equivalently the adjacency matrix) of a molecule, can an automated algorithm and code be written to yield its topological symmetry.

The graph-vertex-partitioning problem also finds important application in the computer generation of ^{13}C and proton NMR signals of molecules and their intensity patterns. We have described in an earlier manuscript the application of the vertex partitioning to generate ^{13}C NMR spectra.[24]

Some early techniques for generating vertex partitioning of graphs have been based on the Morgan algorithm.[1-5] and the principal eigenvector algorithm. Randič and co-workers[10] have formulated the canonical vertex-labeling method to generate the vertex-automorphism partitionings of graphs. Although many of these algorithms and the codes based on these algorithms are simple to use, they often lead to convergence problems and oscillatory behaviors. Herndon and co-workers[6-8,25] have discussed these aspects in considerable depth. These algorithms, including the canonical labeling technique, often do not provide satisfactory solutions for highly transitive graphs, viz., those graphs which contain many vertices of the same degree. For graphs which contain several isospectral points (two nonequivalent vertices in the graph which yield isospectral graphs if identical fragments are attached to these vertices), direct application of both the Morgan algorithm and the principal eigenvector schemes do not provide correct solutions.

Shelley and Munk[11] have developed an extended Morgan algorithm approach for vertex partitioning of graphs. As discussed in the literature (see ref 25), this algorithm cannot discriminate between isospectral points. However, a revised procedure proposed by Shelley and Munk[12] subsequently circumvents this problem. In their second paper, Shelley and Munk have used procedures to construct the sequence number permutations in the early stage of their algorithm. A combination of this procedure and the earlier methods was shown to construct the automorphism partitioning of many graphs, including graphs which contain isospectral points. This second paper of Shelley and Munk[12] has apparently been overlooked in the literature.

In this investigation we develop and apply algorithms and codes for vertex partitioning of graphs. We also critically compare the performance of newly developed codes with the codes based on Morgan, principal eigenvector, and other algorithms. It is demonstrated that the codes developed here

[‡] Camille and Henry Dreyfus Teacher-scholar.

0095-2338/90/1630-0263$02.50/0