CIAC Comprehensive Information System of Rare Earths

LU XU,* GUOQUAN LI, SHUYUN WANG, HONG LU, HUAYUN WANG, CHANGYU HU, XUHONG JIANG, YONGHUI XIAO, YUNDE XIAO, and XIANGYU LÜ

Changchun Institute of Applied Chemistry, Academia Sinica, Changchun 130022, Jilin, People's Republic of China

Received September 24, 1990

The CIAC (Changchun Institute of Applied Chemistry) Comprehensive Information System of Rare Earths is composed of three subsystems, namely, extraction data, physicochemical properties, and reference data. This paper describes the databases pertaining to the extraction of rare earths and their physicochemical properties and discusses the relationships between data retrieval and optimization and between the structures of the extractants and the efficiency with which they are extracted. Expert systems for rare earth extraction and calculation of thermodynamic parameters are described, and an application of pattern recognition to the problems of classification of compounds of the rare earths and prediction of their properties is reported.

INTRODUCTION

The rare earth elements are used widely in the metallurgy and ceramics industries and in the processing of electronic and luminescent materials because they possess special physicochemical properties. For these reasons, studies of the rare earth elements have become important in recent years.

The last several decades have seen considerable study of the properties and applications of these elements, and a great deal of data in these areas has been reported. Since 1970, a number of large databases of chemical information have been built, but there appears to have been no such activity concerning the chemistry or the physics of the rare earths. Consequently, we began in 1987 to construct the CIAC (Changchun Institute of Applied Chemistry) Comprehensive Information System of Rare Earths. This project was organized and supported financially by the Academia Sinica as part of a larger project, which called for the building of some 20 information systems relating to chemistry, physics, and biology. The rare earths system, which is the subject of this paper, is one of these 20 systems.

This rare earths system is an "information system" in that it not only contains several databases but is also capable of deriving secondary data based upon the original information and can support studies of relationships between data, classification of compounds, prediction of properties, optimization, and various applications of artificial intelligence, such as pattern recognition and expert systems. The system is written in Fortran 77 and runs on a Micro VAX-II.

I. THE RARE EARTHS INFORMATION SYSTEM

This paper is concerned with just two subsystems of the rare earths system: extraction of rare earths and their physicochemical properties.

1. Sources of Data. The data are gathered mainly from Handbooks, the chemical literature, and from direct measurements in our laboratory.

(*a*) *Handbooks.* Several handbooks concerning extraction and physicochemical properties of the rare earths have been published. These include the *Gmelin Handbook of Inorganic Chemistry*, the *Handbook on Physics and Chemistry of the Rare Earths*,[1] *Equilibrium Constants of Liquid–Liquid Distribution Reactions*,[2] and the *Handbook of Extraction*,[3] and so on. Handbooks are one of the major sources of data.

(*b*) *Chemical Literature.* Work pertaining to the study of the rare earths and the applications of their properties is fouund in every discipline of science and in papers in many different periodicals, such as journals dealing with general chemistry (50 titles), physical chemistry (53), and inorganic chemistry (30).

Table I. Major Items in Rare Earth Extraction

no.	item	no.	item
1	extractant	7	temperature
2	diluent	8	salt-out agent
3	auxiliary ligand	9	distribution ratio
4	metal ion	10	dissociation equiv constant
5	polar phase	11	reference
6	acidity	12	remarks

Table II. Major Categories of Rare Earth Compounds

no.	type	no.	type
1	metal	37	nitride
2	alloy	38	phosphide
3	oxometallate	39	hydride
4	intermetallic	50	oxyacetate
11	oxide	51	nitrate
12	hydrated oxide	52	sulfate
20	halide	53	oxalate
21	fluoride	54	carbonate
22	chloride	55	phosphate
23	bromide	56	chlorate
24	iodide	57	hydrohalogenate
25	oxyhalogenide	58	bromate
26	hydroxide	59	iodate
30	nonmetal	70	organic
31	boride	71	inorganic
32	carbide	72	carboxylic acid complex
33	sulfide	73	nucleic acid complex
34	selenide	74	organophosphorus complex
35	antimonide	80	complex halogenide
36	silicide		

(*c*) *Direct Measurement.* Data in this category are drawn from the authors' laboratory and from internal, unpublished reports of this Institute.

2. Types of Data.

(*a*) *Extraction of Rare Earths.* The major types of data pertaining to the extraction of rare earths are listed below:

1. Neutral phosphorus-containing extractants
2. Amine extractants
3. Carboxylic acid extractants
4. Extraction of inorganic acids
5. Extraction of diluents
6. Physicochemical properties and structural formulas of extractants and diluents

Currently, there are approximately 4000 records in the database. The main items in one set are illustrated in Table I.

(*b*) *Physicochemical Properties of the Rare Earths.* There are various different categories of rare earth compounds, and the major types are shown in Table II. There are about 4000 sets of physical properties data under consideration for the system. One set of physical properties is shown in Table III.

0095-2338/91/1631-0375$02.50/0

Table III. Partial List of Physicochemical Properties of Rare Earths

no.	item	no.	item
1	density	9	conductivity
2	melting point	10	optical property
3	color	11	refractive index
4	magnetic moment	12	refractivity
5	magnetic conductivity	13	evaporation velocity
6	Nell temperature	14	coefficient of expansion
7	Curie temperature	15	dielectric constant
8	resistance	16	relative viscosity

Table IV. Example of a Single Parameter Table[a]

sort:CHLORIDES					formula:MCl_2	
property:DENSITY					unit:G/CM3	
function:D						
condition: X-RAY METHOD; AT ROOM TEMPERATURE						
	Sc	Y	La	Ce	Pr	Nd
data:						4.542
	Pm	Sm	Eu	Gd	Tb	Dy
data:		4.788	4.856			5.025
	Ho	Er	Tm	Yb	Lu	
data:			5.166	5.272		

[a] D: density; condition: measure method and condition.

Table V. Example of a Multiparameter Table

sort:TRICHLORIDES					formula:MCl_3	
property:SOLUBILITY					unit %	
function:SOL & CEL						
condition: IN WATER;%:WEIGHT PERCENT;SOL:SOLUBILITY;CEL:CELSIUS						
	Sc	Y	La	Ce	Pr	Nd
data:		42.95	48.99		49.65	49.34
data:		25.00	15.40		26.20	15.40
	Pm	Sm	Eu	Gd	Tb	Dy
data:		49.20				
data:		40.00				
	Ho	Er	Tm	Yb	Lu	
data:						
data:						

(c) Data Organization. In this section, some of the physicochemical data will be used to illustrate the way in which the data are organized.

(i) Data Representation. Data pertaining to the rare earths, particularly the physicochemical data (see Table III), relate to a variety of disciplines and a number of different representations of the data are usual, including characters, numerics, curves, and so on. As a result, the range of the data can be very large and for these reasons, a number of rules have been developed for use in representation of data. First, the original data must not be distorted and second, the original information must not be lost. In seeking to attain these goals, different problems can arise. Some data cannot easily be represented in the usual numeric way, and this difficulty is exacerbated when the datum is a range of numbers, rather than a single number. For example, the melting point of Y_2Se_3 is reported[4] as ">1800 °C" which can be represented by a normal character string. The melting point of SmTe on the other hand is cited as "1910 to 1930 °C". This value must be decomposed into two integers and the linkage between them must be reestablished at search time in software.

(ii) Nomenclature of Data Fields. It is common for a specific type of data to be referred to by more than one name. Thus heat of formation is often termed enthalpy of formation; cell parameters are often called lattice constants and so on. However these data are stored, an interface must be provided that can accept any of the synonymous terms and locate the correct information.

(iii) Hierarchical Tree Structure of Data. As is well known, there are 17 elements in the rare earth family. Each of these elements and its compounds have hundreds of properties, but there is a great deal of interelatedness between the properties. Information concerning, for example, the solubility of La_2O_3 ties together data concerning the solvent, the temperature, and the purity of the solute. To organize such multidimensional data, the techniques of "folders" is used to define both logical and physical storage, as follows.

(a) The properties of the rare earth compounds are grouped into a number of tables according to their formulas and the property in question. The length of the tables is variable, and they contain an entry for each of the 17 elements. This entry is displayed explicitly, but is stored implicitly.

(b) The tables have a tree structure. They may be single parameter (see Table IV) or multiparameter, or tables which overlap and are interrelated. The table shown in Figure 4, which carries the densities of the chlorides, is an example of a single parameter table. In the multiparameter, table shown as Table V, the line headed "function" carries the names of the parameters, separated by an ampersand (&). Thus in Table V, "SOL & CEL" signifies solubility at a given temperature. Thus two lines of data are needed; the solubility of YCl_3 in water at 25 °C is 42.95 weight percent, i.e., 429.5 g/L.

The overlapping tables are multiparameter tables in which more complicated interrelationships between the data are present. As an example, the electrical conductivity data for R-RX_3 systems are shown in Table VI.[5] The data in Table VI all pertain to the materials at the head of the table (cf. Table VII), and Table VI can in principle be restated as a series of multiparameter tables, the first two of which are Tables VIII and IX.

Table VI. Conductivity of R-RX_3 Systems[a]

La + $LaCl_3$			Ce + $CeCl_3$			Pr + $PrCl_3$		
temp	rm	cond	temp	rm	cond	temp	rm	cond
910	0	1.40	855	0	1.20	830	0	1.09
910	0.46	1.61	855	0.13	1.22	830	1.1	1.20
910	1.57	1.84	855	0.24	1.23	830	3.2	1.39
910	2.33	2.46	855	0.31	1.28	830	6.0	1.68
910	5.1	3.21	855	0.58	1.31	830	10.6	2.22
910	7.8	5.02	855	1.00	1.38	830	15.3	2.90
910	10.3	7.26	855	1.73	1.56	830	17.8	2.90
910	10.3	7.23	855	2.50	1.78	830	17.8	3.37

[a] Temp: celsius temperature; rm: R (rare earth elements) moles %; cond: conductivity.

Table VII. Head of Relevant Overlapping Table

compound: fused-salt system
formula: R + RCl_3
property: conductivity
unit: 1/ohm·cm
number: 3
function: CNDM & mol % & CEL
condition: at cel; mol %: percent of R molar; cndm melting conductivity

Table VIII. Example 1 of Data in a Relevant Overlapped Table

	Sc	Y	La	Ce	Pr	Nd
data:			1.40	1.20	1.09	
data:			0	0	0	
data:			910	855	830	
	Pm	Sm	Eu	Gd	Tb	Dy
data:						
data:						
data:						
	Ho	Er	Tm	Yb	Lu	
data:						
data:						
data:						

Table IX. Example 2 of Data in a Relevant Overlapped Table

	Sc	Y	La	Ce	Pr	Nd
data:			1.61	1.22	1.20	
data:			0.46	0.13	1.1	
data:			910	855	830	
	Pm	Sm	Eu	Gd	Tb	Dy
data:						
data:						
data:						
	Ho	Er	Tm	Yb	Lu	
data:						
data:						
data:						

(iv) When these overlapping tables are searched, they first must be assembled. The search algorithm builds a "folder", which is a set of overlapping tables having the same header information (cf. Table VII). This is a device that can be used to describe the relationships between many multidimensional data points.

(v) To facilitate search and retrieval, all the compounds of the rare earth elements are divided into the 39 categories shown in Table II. Their properties are categorized into five classes: physical, chemical, thermodynamic, crystalline, and mechanical. Every piece of data thus has two codes and these two codes, together with the ID number of the compound, are combined into a complex seven-digit code, which provides a succinct summary of the data available for the compound. As an example, if compound no. 001 is an oxide (code 11) and has property 4, then its complex code will be 1140001. This not only simplifies searching but also facilitates the development of statistics across the whole file.

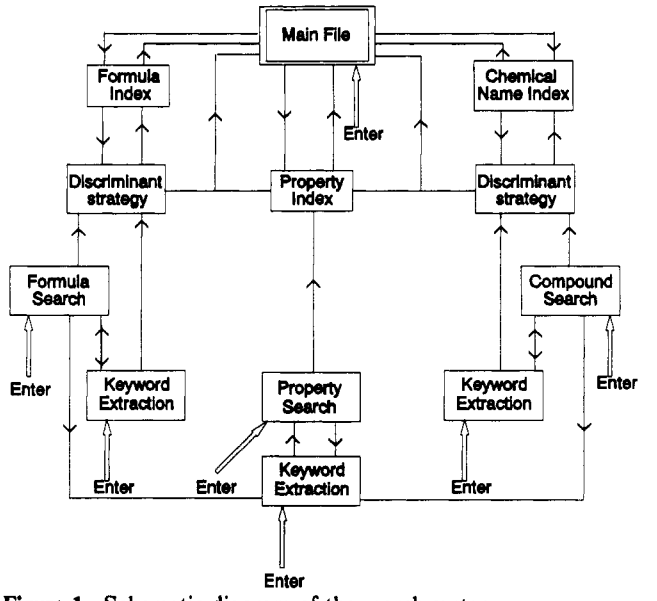

Figure 1. Schematic diagram of the search system.

3. Searching of Data. This system has several unique features which make it especially useful to searchers.

(a) Searches can be carried out not only for specific data types, such as extractant, metal element, polar phase and so on, but also for multiple data types. Automatic intersections are used to accomplish such searches.

(b) Data retrieved in searches can be plotted in two dimensions or in three-dimensional space by application of curve-fitting routines.

(c) A search may have multiple entry points, and this increases the speed and the ease of searching. As an example, in the physicochemistry database, the logical searching environment is shown in Figure 1. The program allows the user to begin retrieval from several different levels such as the name of the compound, its formula, or a combination of the two. This philosophy applies to both the physicochemical and the extraction databases.

II. OPTIMIZATION OF DATA TO SUPPORT EXTRAPOLATION

In order to elicit the relationships between the various parameters, a package of programs supporting curve fitting, multiple regression, and so on was assembled. These programs allow the derivation, for example, of the separation coefficient or extraction distribution ratio of two elements, the so-called β value, which is the ratio of the appropriate extraction distribution coefficients, D_1 and D_2.

Simplex and response surface analyses are used for the optimization of experimental design. The response surface method will be described in detail in this study. A response surface is the graph of a system response, such as an extraction distribution ratio plotted as a function of one or more of the system variables, which include the metal ion concentration, pH, temperature, diluent or salt-out agent, and so on. The response surface provides a means of visualizing the effect on the system of systematic variations in the different variables and assists in the development of an optimization scheme.

For example, the variation with concentration and acidity of the extraction distribution ratio (D) for different rare earth elements at constant temperature and extractant concentration has been studied. The acidity was adjusted by means of acidic

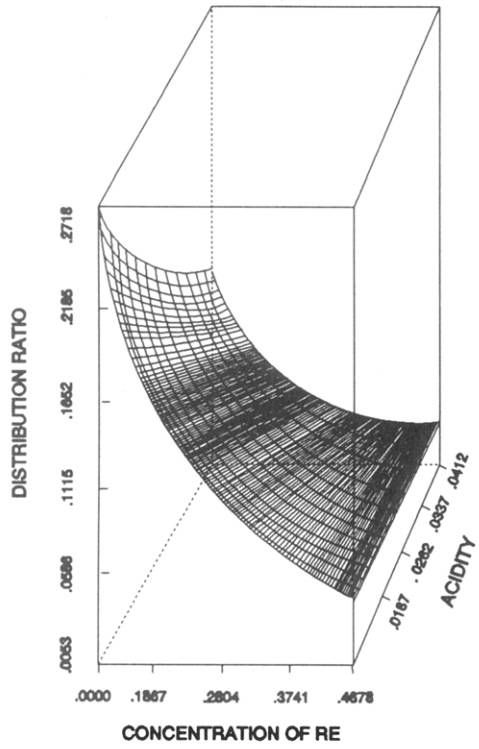

Figure 2. Distribution–concentration–acidity diagram for rare earth.

phosphate. The extractants used were P_{507} [bis(2-ethylhexyl)(2-ethylhexyl)phosphonate] and P_{204} [bis(2-ethylhexyl)-phosphonate]. Mathematical models of the efficiency of extraction for each rare earth were derived by stepwise regression. The function for the $LaCl_3$–HCl–P_{507} system is

$$\ln D = -4.01 - 9.73[H^+] - 0.99 \ln \{[La^{3+}] + [H^+]\}$$

where $[La^{3+}]$ and $[H^+]$ denote the initial concentrations of lanthanum and protons, respectively. The plot of this function in three-dimensional space is shown in Figure 2. This diagram shows that D (the extraction distribution) decreases as $[La^{3+}]$, the "RE", increases. If the value of $[La^{3+}]$ is held constant, the value of D will decrease with increasing acidity.

III. RELATIONSHIPS BETWEEN EXTRACTANT STRUCTURE, REACTIVITIES, AND PHYSICOCHEMICAL PROPERTIES

Correlations between the structures of the extractant and its reactivities and physicochemical properties are important because they can be used to guide the extractant design.

Quantitative structure–activity/property relationship studies (QSAR/QSPR) have been exploited extensively in the design of drugs and pesticides, but few such studies have been applied to the design of extractants. The method of topological indexing of molecular structures has also been used widely in recent years in connection with QSAR/QSPR. The key step in development of a topological index is the selection of a graph invariant, which is a quantity that can be derived from the graph and which is not affected by its node numbering. Over a hundred topological indices have been described but only a few of these are suitable for molecules that contain heteroatoms. The topological index a_N suggested by Yang and Jiang[6] performs well but can only be used for saturated alkanes. In this study, we have extended the a_N topological index into a general index a_N(GAI) and applied it to systems containing heteroatoms.

Relationships between the GAIs of the neutral and acidic phosphorus-containing extractants and the reactivities of Ce, Y, U, and Th as well as between the various thermodynamic

Table X. Correlation of GAIs, $\sum X$, and Extraction Reactivities of Y and Ce

extractant	GAI	$\sum X$	D_{Ce}	D_Y
$(C_4H_9O)_3PO$	0.2997344	8.75	0.026	0.044
$C_4H_9PO(OC_4H_9)_2$	0.4252615	7.95	0.291	0.419
$(C_4H_9)_2PO(OC_4H_9)$	0.6016426	7.67	1.96	3.61
$(C_4H_9)_3PO$	0.8488629	5.55	4.25	12.3

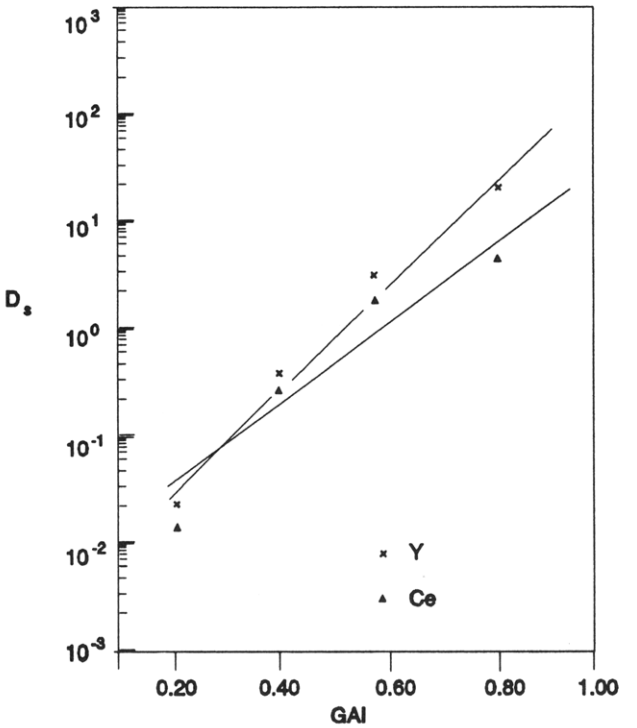

Figure 3. Correlation of GAI and reactivity of Ce and Y with a neutral phosphorus extractant.

properties such as boiling point, density, chromatographic retention index, and enthalpy of formation of aliphatic alkanes have been reported.[7,8] The extraction efficiencies, $\sum X$ values (the sum of the electronegativities of the substituted groups) and GAI for Y and Ce are shown in Table X.

From this table, it can be seen that the GAIs increase as the C-P chain decreases. The charge density at the P═O group will increase as the length of the C-P chain increases, and the tendency of the P═O group to donate an electron will also be increased. This causes the extraction distribution ratios D_s to increase. A plot of D_s versus GAI using the data in Table X is shown in Figure 3. In Table X, $\sum X$ is the sum of the electronegativities of all the substituted groups under the frequency $V_P = 0$, which represents the polarity of a molecule. The correlation between GAIs and $\sum X$ is given by:

$$\text{GAI} = -0.1686 \sum X + 1.18052 \qquad (n = 4; r = 0.97)$$

IV. STUDIES OF EXTRACTION EXPERT SYSTEM

The key step in the extraction process is the selection of a suitable extractant and of extraction conditions. The success of this step depends upon the knowledge and experience of the chemist, and some effort has been made to design an expert system for the planning of the extraction process. The program, which is written in LISP, consists of three parts: (1) separation of the individual rare earth and choice of frequently used extractants, (2) development of recommended procedures for the separation of mixed rare earths with P_{507}, one of the most widely used extractants, and (3) evaluation of the economics of the technical process. As an example, if a mixture of rare earths is to be separated using P_{507}, the optimum process depends upon the composition of the mixture. Figure

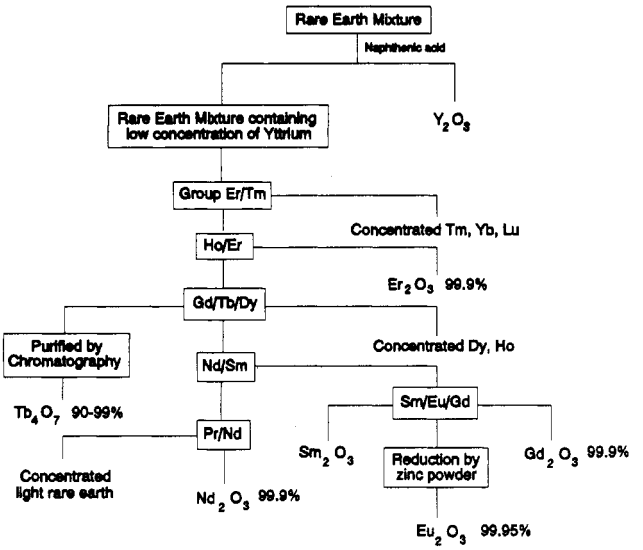

Figure 4. Flow chart of the separation procedure for heavy ore.

Table XI. Example of the Components of a Heavy Rare Earth Ore

oxide	% contents	oxide	% contents
Y_2O_3	65	Tb_4O_7	1.1
La_2O_3	2.2	Dy_2O_3	7.5
Ce_2O_3	<0.1	Ho_2O_3	1.6
Pr_6O_{11}	1.1	Er_2O_3	4.3
Nd_2O_3	3.5	Tm_2O_3	0.6
Sm_2O_3	2.3	Yb_2O_3	3.3
Eu_2O_3	<0.1	Lu_2O_3	3.5
Gd_2O_3	5.7		

4 shows the optimum process for the mixture shown in Table XI. The content of yttrium is higher and its atomic weight smaller than those of the other rare earths, thus the mole percent of yttrium is larger than that of the other rare earths. Consequently, naphthenic acid is used in the first step to extract the yttrium, and the P_{507} extractant is used for the other rare earths.

V. THERMODYNAMIC CALCULATIONS AND PATTERN RECOGNITION

(1) Heat Capacity. The heat capacities of various ions of solid inorganic compounds of the rare earths can be computed using the equation:

$$CP = a + bT^{-3} \times 10 + cT^{-2} \times 10^5$$

where the constants a, b, and c are defined as follows:

$$a = \frac{Tm \times 10^{-3}(\sum\Theta + 1.125n) - 0.298nTm^{-2} \times 10^5 - 2.16n}{Tm \times 10^{-3} - 0.298}$$

$$b = \frac{6.125n + nTm^{-2} \times 10^5 - \sum\Theta}{Tm \times 10^{-3} - 0.298}$$

$$c = -n$$

Here, Θ denotes the contributions of positive and negative ions in a compound, Tm is the absolute value of the melting point, and n is the number of atoms in the compound.

(2) Entropy. The Latimer algorithm[9] was used for the computation of the entropy of ionic compounds. Just as in the case of heat capacity, contributions from both positive and negative ions were considered.

(3) Pattern Recognition. Pattern recognition was used in this work for the classification and prediction of the properties of new compounds and was carried out with MAPP, a multivariate analysis program package developed in this laboratory and providing support component analysis, multiple regression, and so on. The relationship between the structures of complex fluorides and the spectral structure of the Eu(II) ion in complex fluorides (AB_mF_n) has been investigated by pattern recognition techniques such as k nearest neighbors, factor analysis, alternative k nearest neighbors, Bayesian schemes, LLM, SIMCA, and principal component analysis.

Table XII. Variables on Objects AB_mF_n[a]

no.	feature	meaning
1	Z_a/Rk_a	ratio between the number of valence electrons of atom A and its atom core radius
2	Z_b/Rk_b	ratio between the number of valence electrons of atom BA and its atom core radius
3	X_a	electronegativity of atom A
4	X_b	electronegativity of atom B
5	Rc_a/Rc_b	ratio between ionic radii of atoms A and B
6	Z_a/RC_a	ratio between the number of valence electrons of atom A and its covalent atomic radius
7	Z_b/RC_b	ratio between the number of valence electrons of atom B and its covalent atomic radius
8	m	mole ratio of atom B
9	$Z_{a'}$	valence of atom A
10	$Z_{b'}$	valence of atom B
11	Rc_a	ionic radius of A
12	Rc_b	ionic radius of B

[a] A, B: atoms; F: fluorine.

Table XIII. Pattern Recognition Results[a]

	$M = 12$		$M = 6$	
	recognition (%)	prediction (%)	recognition (%)	prediction (%)
KNN	88.9	88.9	87.3	92.6
K = 1	(11,13,18,23,35,41,42)	(3,15,27)		(3,15)
ALKNN	92.1	85.2	92.1	85.2
K1 = K2 = 2	(13,18,27,35)	(3,10,11,15)	(13,18,27,35)	(3,10,11,15)
SIMCA	(A = 3)		87.3	81.5
BAYES	95.2	92.6	96.8	92.5
	(18,23,35)	(3,15)	(18,23)	(3,4)
LLM	85.7	88.9	79.4	85.2

[a] The numbers of misclassified samples are given in parentheses. M = number of variables.

Ninety complex fluorides (AB_mF_n) from the literature and our own measurements were gathered for use as a sample set. These complexes were divided into two categories, 45 with f-f transition emission (class 1) and 45 possessing no f-f transition emission (class 2). Thirty-two samples were selected randomly from class 1 and 31 from class 2 to serve as the training set, and the remainder were treated as unknowns, i.e., the test set. The host of each complex fluoride is characterized by 12 crystal structure variables, shown in Table XII.

The reduction of variables was carried out by combination of the changing variance weight,[10] feature evaluation in Bayesian analysis,[11] and the relevance of variables in SIMCA.[12] Features 2, 5, 7, 8, 9, and 12, which are highly correlated to the nature of the transition emission of the Eu(II) ion in complex fluorides, were selected. A recognition rate between 79.4% and 96.8% and a prediction capability between 85.2% and 92.6% were obtained (see Table XIII[13]).

VI. CONCLUSION

The CIAC Comprehensive Information System of Rare Earths has been built during a 4-year period beginning in 1987. In the semiquantitative areas of this system, namely, the extraction, physicochemical properties, and reference, various search functions and artificial intelligence modules have been completed and made available to users. A material subsystem is under development, but is not yet complete. Spectral and resources subsystems will be undertaken in the near future with an expected completion date in 1993.

ACKNOWLEDGMENT

Financial support for this work from the Center for Scientific Databases, Academia Sinica, is gratefully acknowledged.

REFERENCES AND NOTES

(1) Gschneider, K. A., Jr.; Eyring, L. *Handbook on Physics and Chemistry of the Rare Earths*; North-Holland Publishing Co.: Amsterdam, 1979; Vol. 3. *Ibid.* 1982; Vol. 5.
(2) Marcus, Y.; Kertes, A. S.; Yanir, E. *Equilibrium Constants of Liquid-Liquid Distribution Reactions*; Pages Bros., (Norwich) Ltd.: Norwich, 1974.
(3) *Handbook of Extraction.* Rosen, A. M., Ed.; translated by Yuan Chengye; Atomic Energy Press: Beijing, 1981; Vol. 1. *Ibid.* 1983; Vol. 2. *Ibid.* 1988; Vol. 3.
(4) Zhong Shan University. *Chemicophysical Constants of Rare Earths*; Metallurgical Industry Press: Beijing, 1978.
(5) Mayer, I. P.; et al. *J. Phys. Chem.* **1962**, *66*, 693.
(6) Yang, Jiaan; Jiang, Yuansheng. The Graphic Character and Thermodynamic Properties of Aliphatic Alkanes. *Acta Chim. Sin.* **1983**, *41*, 884.
(7) Wang, Huayun; Lü, Tianxung; Xu, Lu; Wang, Erkang; Su, Qang. General a_N Index and its Applications. I. Extraction Reactivity of Y, Ce, U and Th. *Acta Chim. Sin.* **1990**, *48*, 1159.
(8) Wang, Huayun; Xu, Lu; Su, Qang. General a_N Index and its Applications. II. Properties of Phosphorus Compounds. *Acta Chim. Sin.* **1991**, in press.
(9) Kubaschewski, O.; Alook, A. B. *Metallurgical Thermochemistry.* Butterworth-Spring, Pergamon: Oxford, 1979.
(10) Chen, Nianyi; Xu, Zhihong; Liu, Hongling; Hu, Hua; Wang, Leshan. *Computational Chemistry and Application.* Shanghai Sciences Press: Shangchai, 1987.
(11) Jurs, P. C.; Kowalski, B. R.; Isenhour, T. L. Computerized Learning Machine Applications to Chemical Problems. *Anal. Chem.* **1969**, *41*, 21.
(12) Wold, S.; Sjostrom. SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy. *ACS Symp. Ser.* **1977**, *No. 52*, 243.
(13) Xiao, Yunde; Xu, Lu. Search for the Basic Regularities of the Transition Emission of Eu(II) Ion Complex Fluorides with Pattern Recognition. *Chemom. Intell. Lab. Syst.* **1991**, in press.

MAPOS: A Computer Program for Organic Synthesis Design Based on Synthon Model of Organic Chemistry

LUDĚK MATYSKA* and JAROSLAV KOČA†

Institute of Computer Science and Department of Organic Chemistry, Faculty of Science, Masaryk University, 611 37 Brno, Czechoslovakia

Received October 29, 1990

The program MAPOS is a logically oriented computer program for computer-aided organic synthesis design suitable both for forward and retrosynthetic synthesis planning. It is based on the synthon model of organic chemistry, introduced by the authors. The fundamentals of the model as well as the basic algorithms are described. Examples of the use of the program are given.

1. INTRODUCTION

Computer-aided organic synthesis design (CAOS) is currently an active field of computer chemistry. Programs for CAOS may be classified in two basic directions—the information-oriented programs and those logically oriented.

The former are based on a database of chemical reactions. They may be specialized for some area of organic synthesis (e.g., synthesis of heterocycles), but there exist attempts to cover organic chemistry as a whole.[1-7]

The logically oriented programs need a formal model of organic chemistry.[8,15] This formal (mathematical) model is usually based on discrete mathematics, and it is not directly related to quantum chemistry models. However, energy computational methods, even those based on quantum mechanistics, often serve as bases of the strong heuristics used. Several such programs have been noted in the literature;[9-13,15] many of them are based on the Dugundji-Ugi model of constitutional chemistry.[8]

Obviously, some CAOS computer programs do not fall strictly into any of the two mentioned groups. The empirical chemical knowledge on different levels together with a mechanistic approach is usually used in them.[14,16,17]

The program MAPOS, presented in this paper, belongs to the logically oriented programs. It is based on the synthon model of realistic constitutional chemistry.[18,19] In this model, the central role is played by the so-called valence states of atoms, the notation almost identical with Pauling's[20] and Van Vleck's[21] idea of "atom in molecule". The model incorporates the notion of the reaction center, mostly represented by one (addition and elimination reactions) or two (substitution reactions) atoms where a primary attack occurs. Only changes initiated by the primary attack are expanded to other atoms of the molecule.

The combinatorial nature of this mathematical model is further restricted by the use of heuristics based on the reaction distance,[18,19,22,23] defined as the minimal number of elementary steps of valence electrons reorganizations (ESRE).[8,24-28] Reaction distance is a notion very similar to chemical distance.[29] However, no one of these metrics may be directly transformed to the other; they represent different aspects of the chemical reality.[22]

A similar approach, based directly on the Dugundji-Ugi work, is used in the IGOR system.[30]

2. MATHEMATICAL MODEL

The synthon is the basic structural unit used in the model. The notion of synthon has been, in the frame of organic chemistry, initially introduced by Corey.[31] Based on Corey's approach, the formal synthon model of organic chemistry has been introduced.[18,19,22] It may be understood as a generalized concept of valence states of atoms and their combinations. From all used meanings of the notion "synthon", the meaning of "substructure reduced in the reaction" is the most closest to our formal definition.

The synthon $S(A)$ over an (arbitrary) atomic set A is defined as one or several molecules and/or their parts, composed of atoms from the set A (all atoms must be used). The most similar notion, introduced so-far in the scope of CAOS, is the definition of Ensemble of Molecules EM(A).[8] The notion of synthon is more general because free valences, i.e., bonds that do not connect two atoms from A but only start from one atom, may be also specified.

Being a generalization of the concept of valence states of atoms, the synthon carries also a strong mechanistic content—conversions of valence states of atoms allow clear

0095-2338/91/1631-0380$02.50/0