(66) Lederberg, J., "DENDRAL-64." Part II. NASA CR-68898, December 15, 1965.
(67) Sorter, P. F., C. E. Granito, J. C. Gilmer, A. Gelberg, and E. A. Metcalf, J. CHEM. DOC. 4, 56 (1954).
(68) Granito, C. E., A. Gelberg, J. E. Schultz, G. W. Gibson, and E. A. Metcalf, *ibid.*, 5, 52-5 (1965).
(69) Granito, C. E., J. E. Schultz, G. W. Gibson, A. Gelberg, R. J. Williams, and E. A. Metcalf, *ibid.*, 5, 229-33 (1965).
(70) Landee, F. A., Abstracts of Papers, p. 3F, 147th Meeting, ACS, Philadelphia, Pa., April 1964.
(71) Landee, F. A., "Computer Methods of Handling Files of Chemically Oriented Information," Dow Chemical Co., Midland, 1965.
(72) Bowman, C. M., F. A. Landee, and M. H. Reslock, J. CHEM. DOC. 7, 43-7 (1967).
(73) Bowman, C. M., F. A. Landee, N. W. Lee, and M. H. Reslock, Abstracts of Papers, 155th Meeting, ACS, San Francisco, Calif., April 1968.
(74) Gibson, G. W., C. E. Granito, D. E. Renard, and E. A. Metcalf, "The Wiswesser Line-Notation: An Introduction," CRDL Tech. Memo 7-3, Edgewood Arsenal, Md., 1965.
(75) Mitchell, J. P., Editor, "Proceedings of the Wiswesser Line Notation Meeting of the Army CIDS Program, 6-7 Oct. 1966," EASP 400-8, Edgewood Arsenal, Md., January 1968.
(76) Barnard, A. J., Jr., C. T. Kleppinger, and W. J. Wiswesser, J. CHEM. DOC. 6, 41-8 (1966).
(77) Horner, J. K., *ibid.*, 7, 85-8 (1967).
(78) Hyde, E., F. W. Matthews, L. H. Thomson, and W. J. Wiswesser, *ibid.*, 7, 200-4 (1967)
(79) Thomson, L. H., E. Hyde, and F. W. Matthews, *ibid.*, 7, 204-9 (1967).

Use of the IUPAC Notation in Computer Processing of Information on Chemical Structures*

H. F. DAMMERS and D. J. POLTON
"Shell" Research Limited, Woodstock Agricultural Research Centre, Sittingbourne, Kent, England.

Received May 21, 1968

A computer-operated storage and retrieval system for chemical structures based on the use of the IUPAC notation has been in operation at Shell Research Limited, Sittingbourne, Kent, England, since 1965, involving a file of nearly 50,000 compounds. Use of the IUPAC cipher has proved advantageous as regards speed and cost of both input and searching. For most searches, scanning of the information explicit in the cipher has proved adequate. Our computer programs also enable conversion of ciphers into atom-connection tables and generation of fragmentation codes. The integrated use of these facilities and their merits relative to other approaches are discussed.

Research activities in the agricultural chemicals and public health fields carried out by Shell companies have involved the synthesis and testing of large numbers of organic chemicals. The main part of this work has been carried out partly by Shell Research Limited at its Woodstock Agricultural Research Centre at Sittingbourne, Kent, U.K. (Table I) and partly by Shell Development Company in its Agricultural Research Division at Modesto, California.

In 1962, it was decided that computer processing would be necessary if we were to achieve the fullest possible exploitation of our large file of compound data; in addition, it was considered essential that a complete structural description of the compounds be provided to the computer system rather than one based on a fragmentation code. It was agreed to pursue two different approaches as regards the method of inputting the structural information with the aim of integrating the two methods at a later date. One of these methods was based on the use of the chemical structure typewriter.[1] This approach was followed by Shell Development Company and led to the machine described by J. M. Mullen,[2] which has been in use since late 1965 at Modesto and Sittingbourne.

The other method, followed at Sittingbourne, makes use of the IUPAC notation.[3] The main reason for adopting a notation that converts graphic structures into a linear graph (cipher) suitable for input into the computer system was that it presented the most convenient and

Table I. Shell Research Laboratories at Sittingbourne, Kent

Laboratory	Subject field
Woodstock Agricultural Research Centre	Chemicals, in particular pesticides, for use in agriculture and public health
Tunstall Laboratory	Toxicology; environmental health aspects of Shell products/processes
Milstead Laboratory	Chemical enzymology; natural products of biological significance

Total staff ca. 500
Graduate staff ca. 150

Technical Information Services (staff: 11 + 3 part-time)

responsible for provision of:
library services, literature searches,
research data storage and retrieval,
computer services,
translations, notification, etc.

* Presented before the Division of Chemical Literature, Symposium on Notation Systems, 155th Meeting, ACS, San Francisco, Calif., April 4, 1968.

economical method of input available at the time (1962), thus providing us with the means of establishing an operational system relatively quickly. Our experience has confirmed this view. The system became operational for computer searches on our large file of compounds by mid-1965.

The main reasons in 1962 for choosing the IUPAC notation rather than the current main contenders—i.e., Wiswesser[4] and Hayward[5]—were essentially pragmatic rather than theoretical. They were:

1) It had been published by IUPAC in a definitive form; it therefore appeared to command the maximum of international agreement.[6]

2) It was, at the time, extensively used by the Research and Development Department of *Chemical Abstracts*.

3) The Wiswesser notation, although widely used in the United States, was, as far as we knew, not in use in Europe. Moreover, it was very much in a state of flux as regards agreement on rules.

4) The Hayward notation, although basically an attractive approach, had as yet been inadequately explored.

There were, however, also some theoretical reasons which attracted us to the use of the IUPAC notation—e.g., its suitability for dealing with even the most complex ring systems in our file, its relationship with systematic nomenclature, etc.

Nonetheless, the choice was, in essence, a pragmatic one and this is perhaps understandable considering the fact that any system development in our case had, and still has, to be done by a staff dealing with a variety of day-to-day operational information handling tasks. In connection with this, Table II gives some indication of our over-all operation and development activities. Hence our need for a system that first of all could be developed as a side activity by our operational staff, secondly would become operational at an early stage, thus allowing us to abandon various labor intensive conventional operations, and last but not least would be capable of further improvement once operational.

CIPHERING AND INPUT CONVERSION

The most common objection made to the use of a notation system such as the IUPAC is that it requires highly skilled staff for its operation, and hence not only puts an additional burden on such staff, but also tends to be expensive to operate. In view of this, the information on ciphering performance given in Figures 1 and 2 is of interest. It means that a load of say 5000 new compounds per annum or ca. 100 per week is likely to occupy ca. 3% of the working time of one experienced cipherer. Similarly, we find with regard to keypunching of the cipher that, once the operator is familiar with the cipher, speed is as high (5 to 6 strokes per second) as for the more usual keypunching operations. Hence input has proved economical and fast in particular as ciphering and keypunching can overlap. Table III gives a comparison of the cost of this type of input compared with structure typewriter input. The cost advantage becomes more marked if one considers that topological code input needs, as a rule, substantial computer processing before it is suitable for storage and searching, whereas the cipher input can be used as such for search purposes or after minor processing only.

Table II. Developments in Mechanized Information Storage, Retrieval, and Dissemination at "Shell" Research Ltd., Sittingbourne

	Main Developments
Phase I (1962–64)	Preparatory phase
	Start with research data coding (structures according to IUPAC notation)
	IBM 870 tabulation
	IBM 1401 and Autocoder use
	Feature card system development
Phase II (1964–66)	First computer-operated search system operational (chemical structures)
	KWIC indexing
	Variety of computer-systems (processors) used
Phase III (1966–68)	On-line computer use operational (Univac 1108 system)
	Setting up of inverted files for rapid access
	SDI-Data acquisition on magnetic tape
	Tape typewriters for text and graphic structure input
	Fully mechanized feature card input
Phase IV (1968–70)	Extension of access points throughout site
	On-line use mainly for selected files, usually inverted
	Start of computer aided typing
	Start of real on-line mass storage
Phase V (1970–72)	Multi-access system fully extended (probably 30 access points)
	I/O improvements (CRT, Rand tablets etc.)
	Capture of most site-produced texts for computer editing/storage
	Sizeable on-line mass storage of literature information
	Gradual disappearance of most manual retrieval tools

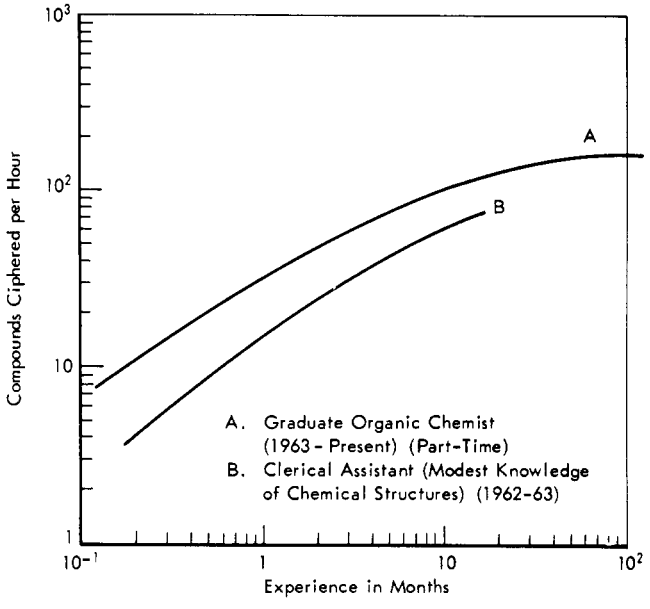


Figure 1. Ciphering performance

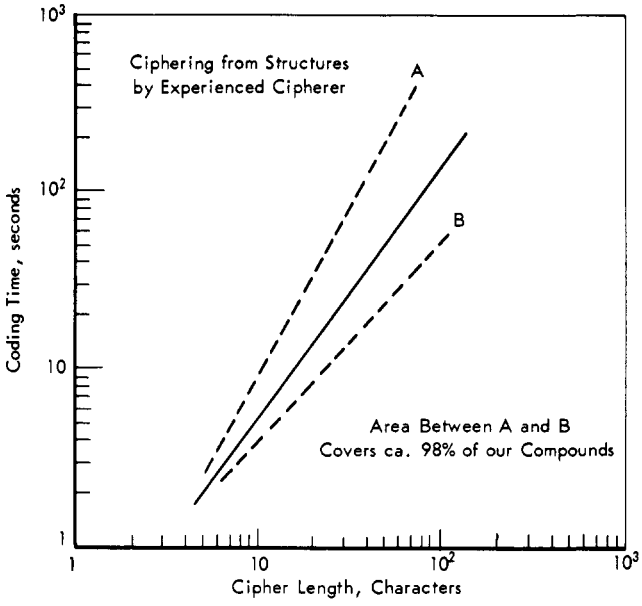
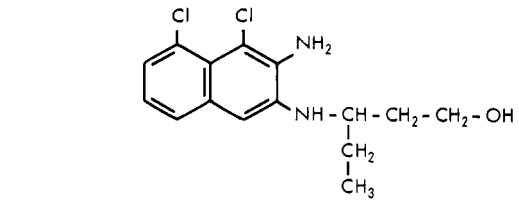


Figure 2. Coding time as a function of cipher length

Figure 3 gives an example of a compound with its notation and its representation after keypunching. The choice of symbols/shift signals in this representation was very much governed by the use of the IBM 870 during the period 1963–65 for tabulation of the ciphers.

The cipher in this form has been used for custom tape searches since mid-1965. More recently however, we have been devising a rearrangement of the internal representation with the aim of simplifying search programs and increasing search speed. The internal representation is illustrated in Figure 4 and will be discussed more fully in a forthcoming paper.[7] An analysis of the composition of our cipher file has shown that this rearrangement does not increase the average length of the cipher record. This is mainly due to the fact that shift signals tend to account for ca. 25% of the characters in a cipher record (Table IV).

An analysis of the file, then containing ca. 45,000 cipher records, indicated a distribution of the cipher length (including shift signals) as indicated in Figure 5. The average cipher length corresponding to this distribution was 25.6 characters per record.



Cipher: Normal B6₂Ch310N4:5N/3C₅Q

Cipher: As Keypunched B6=2*C=H*3-1-0N4,5N/3C=5*Q=(

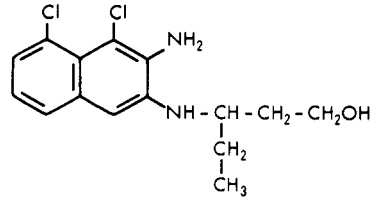
= Lower Case Signal

* Normal Case Signal

(In Lower Case Indicates End of Cipher

, Indicates Colon (Not Available on Keypunch)

Figure 3. Cipher input: use of shift signals



Cipher: Normal B6₂Ch310N4:5N/3C₅Q

Cipher: Internal Representation B661,L3.10.N4:5N/3CCCCC,Q1

Figure 4. IUPAC cipher: example of internal representation

1. Alphabetic and numeric characters subscripted by numerals are written out in full
2. The substituents are separated from the parent group by a comma
3. All omitted "1" locants are reinserted
4. Common chemical symbols of two characters, one upper and one lower case, are replaced by one character—L for chlorine, R for bromine
5. Numerals consisting of more than one digit (underlined in the cipher) are enclosed between full stops (periods). The same applies to most elements—e.g., .CD. for Cd.

Table III. Effort Involved in Structure Input via Cipher and via Structure Typewriter

Optimum performance

	Per 100 structures	
	Cipher	Typewriter
Labor effort professional	0.8 hour	
clerical (key-punching, tape typewriter)	0.6 hour	5 hours
total	1.4 hours	5 hours[a]
Total labor cost	ca. 30 shillings	ca. 50 shillings
Time lapsed (start to finish)	0.9 hour	5 hours

[a] This effort provides of course typescript copies of the structure at the same time.

Table IV. Proportion of Shift Signals in Cipher

Average Cipher Length (Including Shift Signals)	% of Shft Signals in Cipher
3	67
8	35
13	28
18	27
23	25
28	25
33	25
38	25
43	25
48	25
53	25
58	26
63	25
68	25

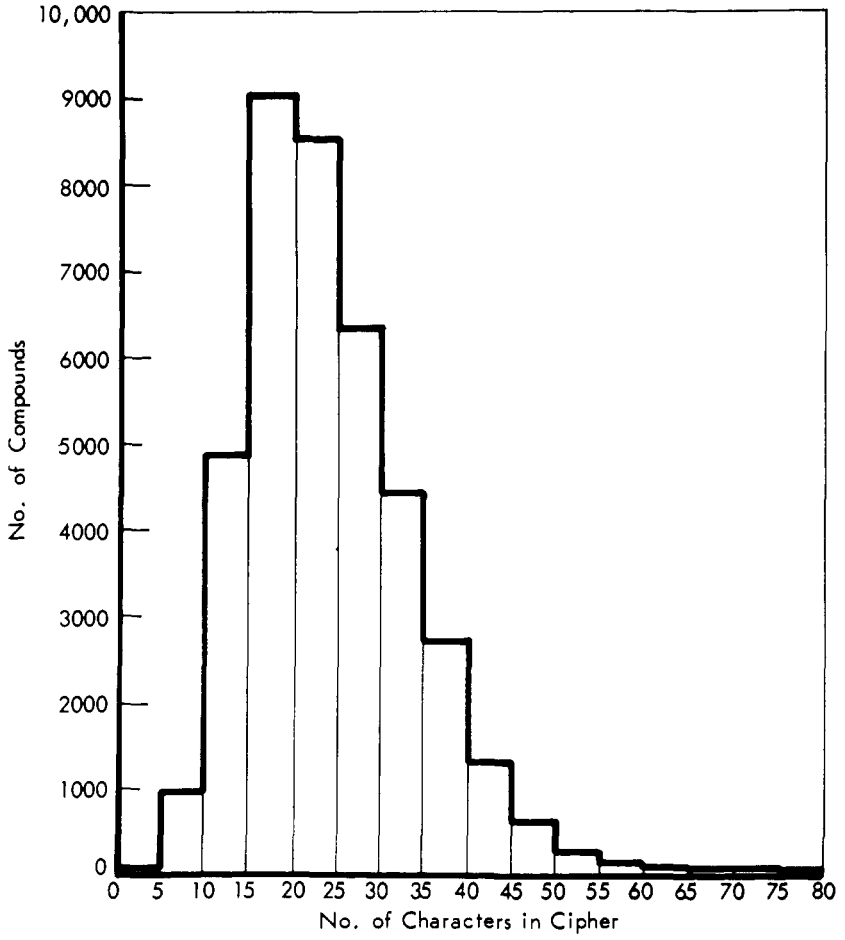


Figure 5. Distribution of ciphers according to record length

Before leaving the subject of cipher input, the question of error-rate should be commented upon. No particular care has been taken in devising error-detection at the manual input stage. In fact, for a long time we used simple visual perusal instead of punch verifying to maximize output from our limited keypunching facilities. The main aim was to avoid, as much as possible, constraints on input speed and to eliminate any errors later by machine detection via operational searches and by checks against manual searches. Such checks and examination of tape-dumps have indicated that error-level is surprisingly low—i.e., of the order of 0.5%. This should eventually be reduced significantly by the use of machine operated error-detection procedures.

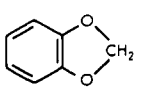
SEARCHES USING INFORMATION EXPLICIT IN THE CIPHER

From early 1965 till late 1967, all our computer searches were carried out via simple scanning of the characters displayed in the cipher. A relatively simple example of this type of search has been given in an earlier paper.[8]

As shown in Table V, a succession of search conditions of increasing specificity can effect retrieval with a very high degree of relevance, without the need for a topological search requiring the conversion of the cipher into an atom-connection table. In general, the type of searches carried out by us (Table VI) tend to confirm this; however, the fact that we are fairly well informed on the overall composition of the file (mainly as a result of feedback from searches) may well have influenced our search results favorably, as we can roughly predict the response likely to be produced by a given search condition or combination of search conditions.

Table V. Example of Structure Retrieval via Scanning of Cipher as Such

Search for methylenedioxy compounds



Cipher: B65ZQ79
42,000 compounds in file

Search Conditions	Compounds Retrieved: % of total file	Compounds Retrieved: Compared with no. of relevant compounds
1. A or B present followed by 5	25	
2. (1) + ZQ present	2.8	
3. (2) + 2 locants following ZQ	1.6	2.75
4. (3) + ZQ locants to differ by 2 (n and n + 2)	1.04	1.86
5. (4) + locant n + 1 to be absent	0.64	1.10

Table VI. Examples of Computer Searches Carried Out at "Shell" Research, Sittingbourne

	No. of Compounds, Found/Relevant	% Relevance
A specific heterocyclic system without phosphorus containing groups	140/140	100
Specific oxadiazinediones	0	...
Certain *N*-alkyl chloroacetamides *O*-derivatives	13/11	84.5
Alkyl nitrophenols and their *O*-derivatives	105/89	84.7
Hydroxythiophenols and their *O*- and *S*-derivatives	478/408	85.3
A specific OP. compound with any substituent on the ring	18/4	22.2
All phosphonates	743/	unchecked
Compounds with 7-membered rings	383/376	98

As stated in the introduction, our need was for a system that could become operational at an early stage and be further developed while in operation. Use of the notation has met the requirement admirably as the notation, read in from punched cards, provided us immediately with a suitable compact search medium. Search programs were written for each specific case, and this proved quite feasible as we could show, even under these initial conditions, that searches could be carried out with less effort and at a lower total cost than if executed conventionally by manual means. Table VII gives some cost estimated for computer searches carried out using IBM 1460, IBM 7094, and Univac 1108 facilities, whereas Table VIII shows that those searches that are now carried out by computer would have cost on average £35 (range £15–£100) to carry out conventionally. Our main guiding principle has been to ensure that computer use should:

a) Free information scientists for development work, hence reduce labor effort involved in day-to-day operations.
b) Lead to a total cost not in excess of the cost of carrying out the searches by conventional means.

In accordance with this approach, we have not attempted to develop a generalized search system at an early stage. The main reasons for avoiding such a generalized system were in our view:

a) The considerable labor effort and hence cost involved in developing it (at least equal to total *ad hoc* search programming for one year).
b) The delay in getting such a system operational (with our part-time effort at least a year).
c) The chance that its running cost would be relatively high compared with special purpose searches.
d) The expected need for frequent and perhaps cumbersome modifications when in operation.

Instead, we have used the various *ad hoc* searches to build up gradually a set of subroutines.

Some Examples of Subroutines Developed as Product of Searches

All in use with unmodified IUPAC cipher

1. Standard introduction to search.
2. Test whether a character is numerical or otherwise.
3. Count of number of locants following a substituent.
4. Test for a specific cipher fragment.
5. List of all atoms connected to a specific system.
6. Test for subscript numerals.

These subroutines could, at a later date if desired, be integrated into a generalized system, based on the experience gained in the special purpose searches. We are now cautiously moving toward that stage.

The above searches all concern the retrieval of compounds with given substructures or Markush formulas. Search for specific complete structures is as yet faster and more economical via our conventional molecular formula index in card form, at least when only a few compounds are involved. However, if a substantial number (more than say 100 at any one time) is involved, it is quicker to cipher, keypunch, and process the structures via the computer, involving a sort and match against a file which holds a sorted list of all our compound ciphers. Such cases arise when a relatively large number of compounds are supplied to us from outside sources.

Table VII. Approximate Costs Associated with Special Purpose Cipher Searches

Records in cipher file ca. 40,000

Year	1965	1966
Computer used	IBM 1460	IBM 7094
Programming language	1401 autocode	Fortran IV
Search time (1 to 3 queries)[a]	24 minutes	6 minutes
Cost computer time[b]	ca. £16	ca. £30
Labor effort (programming, etc.)	average 8 hr	(range 2 to ca. 30 hr)
Cost per query	ca. £10–£25	ca. £15–£40

[a] Search time very much inflated in particular on the 7094 due to record length being greatly in excess of cipher length as it also included nomenclature, molecular formula, etc.; moreover, the 1460 tape was used on the 7094 hence requiring unpacking and conversion. [b] Use of suitable storage of the cipher file on fast tapes or drum (Fastrand) in conjunction with Univac 1108 systems should bring cost of computer time per search down to less than £10.

Table VIII. Chemical Structure Queries

A, Preferred search medium	B, Frequency per Annum	C, Average Labor Effort Required for a Manual Search/Consultation	D, Total Labor Effort Involved in Searches if Done Manually ($B \times C$)
1. Molecular formula index[a]	1000–1200 (ca. 70%)[b]	ca. 2 min. (1–5 min.)	ca. 40 hr./annum
2. Punched feature-card system	300–500 (ca. 25%)	45 min. (20 min.–2 hr.)	ca. 300 hr./annum
3. Computer search system	50–100 (ca. 5%)	25 hr. (10–60 hr.)[c] equiv. to ca. £35	ca. 2000 hr./annum

[a] Card index system arranged according to Fletcher system. [b] Includes simple consultations—e.g., to verify structure or name. [c] Assumes searches done via the molecular formula card index and feature cards (where feasible).

An example of the sorted cipher file is given in Figure 6; it is used only within the computer system. We have also considered the creation of a file of sorted rotated ciphers—i.e., ciphers which have been rotated around strokes. However, as yet we have not pursued this mainly because of the increase in file size involved.

CIPHER CONVERSION TO ATOM-CONNECTION TABLE

Despite the frequent emphasis in the literature on the need for atom-connection tables and/or topological coding in structure searching, we have found, during the two years of actual computer use of our cipher file involving a great number of searches, very little need or economic justification for such type of coding.

Our own desire to convert the cipher into an atom-connection representation arises from the following reasons, listed in order of increasing importance:

a) Usefulness in some searches.
b) Use in error-detection.
c) Compatibility with structure typewriter input.

As a result, we have given only low priority to the job of writing the programs necessary for this conversion.

Use of such a conversion for error-detection has been in our mind since 1962, but error-rate in actual operation has not worried us sufficiently to divert part of our effort to the conversion. We have, however, started to write the conversion programs, mainly in order that our cipher-based system might be capable of linking up with the system developed by Shell Development Company at Emeryville, California, which is based on the use of chemical structure typewriter input.[2] Our connection-table closely follows the structure of the cipher (is cipher-directed) and resembles the type outlined by Dyson *et al.*[9] It makes use of an internally rearranged cipher such as illustrated in Figure 4.

As far as we have been able to assess, the conversion will work with any ring configuration however complex, expressed in IUPAC cipher notation. The program and its applications will be discussed in detail in a forthcoming paper,[10] hence it may suffice to indicate here only some of the operational desiderata which have guided its development.

As this type of structure representation is expensive to search, it will only be used for those compounds which have been selected via scanning of the information explicit in the cipher. This means that as a rule, it will only be applied to 0.1 to 0.2% of all the structures present in our file. In such cases, the connection-table will be generated from the cipher, examined, and used for the generation of suitable visual display.

It will be used for all new input to calculate the molecular formula, which can be matched against the molecular formula provided with the original cipher input, and to match cipher input against input from the chemical structure typewriter. However, it is not intended to store the table either as such or even in abbreviated form. Instead, the table is designed to display, when required and as comprehensively as possible, information on every non-hydrogen atom present in the structure (its type, enumeration, attachments, bond types, valency, stereo considerations, etc.).

Simple structures only have been used in the example for clearness of representation.

```
A 5 C 1 E 1 E Q 3 Q 2 = (
A 6 C E Q = A * 1 E 3 = (
A 6 C Q 1 E 3 = (
A 6 Z Q 1 C E Q = A * 2 E 5 = (
A 7 E Q = (
E 6 C = 2 (
B 6 C 1 F 3 : 4 Q / 6 C = 6 * C = H 2 (
B 6 C 1 F 3 : 4 Q / 6 C = 6 * C = H 3 (
B 6 E Q 1 4 = (
B 6 Q C = (
B 6 Q 1 2 = (
C = 2 * B = R (
C = 2 * E : Q / 2 C = 2 * C = H (
C = 2 * E 1 B = R (
C = 2 * Q : 2 / = 2 * Q = (
C = 3 * C 2 E Q ; C = H (
C = 3 * E 1 = C * C = H * 1 3 = (
C = 3 * Q C = 2 * 1 Q 2 3 = (
C = 3 * Q C = 2 * 2 Q = (
C = 3 * Q C 1 C = H * 2 = (
C = 3 * Q C 2 = 2 (
C = 3 * X 1 C = H * 2 = (
C = 4 * C 2 C = H * 4 = (
C = 4 * Q C 2 Q 3 = (
C = 4 * Q 2 Q = H * 3 = (
C = 5 * C = H * 3 = (
C C = H * / = 2 * Q = (
```

Notes: = Lower case/subscript signal is indicated by =
* Return to normal case is indicated by *
(End of cipher is indicated by (in lower case
H Chlorine, ciphered as Ch, is therefore represented by C = H

Figure 6. Part of an ordered cipher file

CIPHER FRAGMENTATION AND ITS APPLICATION

While the cipher can be expanded to display all its implicit information in a connection-table, it can conversely also be broken up to provide a fragmentation code (Figure 7). The latter process obviously involves information loss but is nonetheless of considerable value.

The main objects of generating a fragmentation code in our case were, in the first instance, to provide research workers with a suitable manual structure search tool (feature card system) and in the second place to use it for file subdivision in the computer system, leading to a simple system for on-line compound selection.

Basically, the fragmentation procedure selects, from the cipher, units indicated by element or ring symbols, while ignoring the relational information (Figure 8).

The feasibility of this fragmentation was established about three years ago when we carried out a test involving ciphers on 1000 of our compounds. We found that the number of distinct fragments generated increased as indicated in Figure 9, suggesting that the following relationship exists:

$$K = a \log_2 M$$

in which K = number of different fragments, M = number of compounds in system, and a = average number of fragments generated per compound (ca. 11).

The frequency of fragment occurrence obviously provides a guide to the use of certain structural features in compounds of, for example, given biological charac-

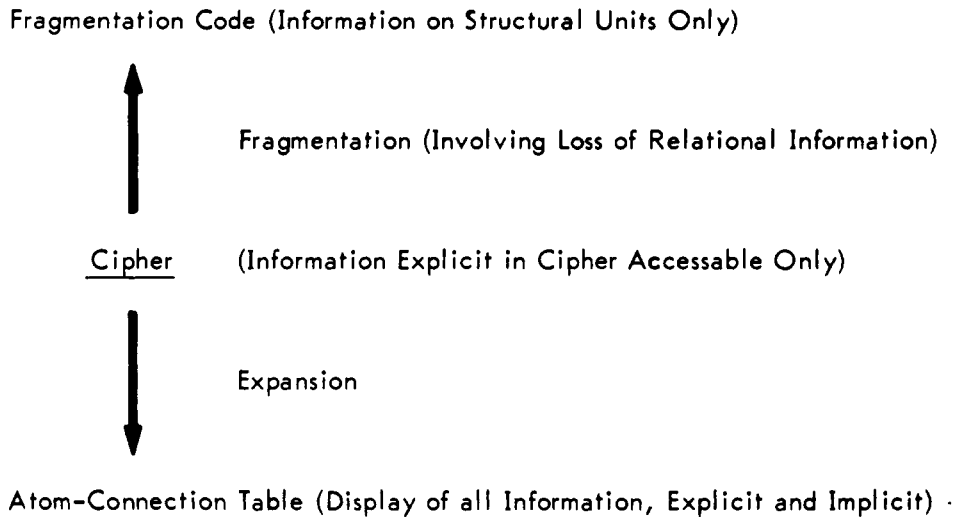
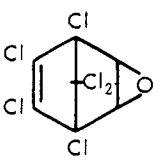


Figure 7. Cipher conversion



Cipher $A5_231-34ZQ8Ch12_2367E6$

Fragments: 3 Membered Ring(s)
5 Membered Ring(s), More than One
3 Ring Aggregate
3 Rings Present
Bridged Ring
Six or More Chlorine Atoms Present
Heterocyclic Oxygen, One Only
Non-Aromatic Double Bond

Figure 8. Example of cipher fragmentation

terization. For the list of the 1000 compounds mentioned, such a distribution is indicated in Table IX.

It is obviously not feasible to discuss this approach here in detail, but we hope to provide fuller information on cipher fragmentation and its use in a separate paper.[11] We should mention, however, the very effective use that can be made of such fragments in a coordinate index type search. To assess this, the features were entered into a feature card system (one card per fragment) which was searched via superimposition and optical coincidence for various classes of compounds. The search results are indicated in Table X. A manual search of the set of 1000 compounds used in this experiment showed that all relevant compounds had been retrieved by the feature card searches except for two compounds; failure to retrieve these compounds proved to be due to an error made when entering the compounds into the feature card system. Such errors are hardly avoidable in a manually operated system and this, together with the need to minimize effort in setting up feature card systems for manual consultation,

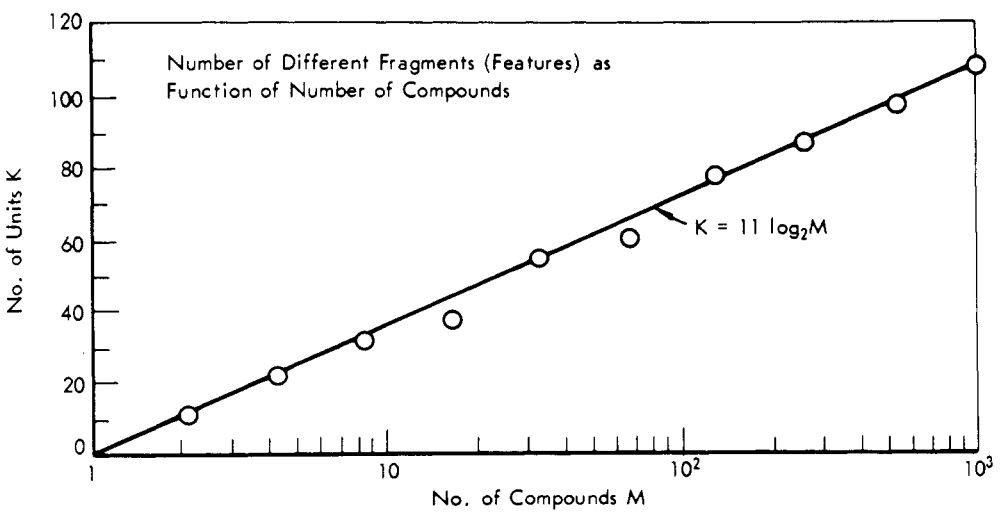


Figure 9. Cipher fragmentation

Table IX. Occurrence of Fragments in a Group of 1000 Compounds

Occurrence[a]	No. of Fragments
1 ×	9
2 ×	9
3–4	8
5–8	9
9–16	8
17–32	8
33–64	19
65–128	18
129–256	14
257–512	15
	119

[a] Average occurrence of fragments (11 × 1000) 119 = ca. 100.

Table X. Coordinate Index Type of Search Using IUPAC Cipher Fragments (Features)

Results obtained in 10 experimental searches on a set of 1000 compounds

	No. of Features Used	Compounds Retrieved		
		Total	Relevant	% Relevant
1. Ureas (aliphatic)	4	22	16	73
2. Ureas (with N as part of cycle)	4	20	4	20[a]
3. Nitrodiphenyls	4	20	3	15[a]
4. Dinitrophenols	4	16	11	70
5. Methylenedioxy compounds	4	11	5	45[b]
6. Anthraquinones	5	8	7	85
7. Thiophenes	4	2	2	100
8. Phenazines	5	2	2	100
9. Indanes	4	1	1	100
10. Vinylnaphthalenes	5	1	0	0
		103	51	50–60% on av.

[a] Low % relevance due to general nature of features and nondistinction (at this stage) between ring and nonring substituents. [b] Low % relevance due to a batch of freak structures in the 1000 compounds studied.

has led us to aim for complete mechanization of fragment generation and their input into a feature card system. Accordingly, we have written computer programs which will fragment ciphers along the lines indicated. One such program, writted in FORTRAN IV, selects an average 5 to 6 features per compound according to a permitted set of 83 features. Processing time involved using the IBM 7094 was ca. 30 ms. per compound. In addition, we hope to have by mid-1968 a feature card punch which will accept input on paper tape, produced by the feature generating program, and enter the information automatically into a set of feature cards. This should enable us to produce rapidly and at low cost sets of feature cards, which can be used by chemists at the bench for rapid manual searches.

Within the computer system, the features produced by cipher fragmentation can be used to set up inverted files for rapid on-line consultation of our structure file. A system such as that shown here should allow interrogation of the structure file virtually in real time.

ON-LINE CONSULTATION OF CHEMICAL STRUCTURE FILE

Target Date Mid-1969

Univac 1108 system operating under multi-programming executive.

File of ca. 50,000 compounds held as inverted files, according to cipher fragments, on Fastrand random access storage drum.

Total size of file ca. 4 million characters.

Access via teleprinter.

1. Call for search routine and entry of features causes relevant inverted files to be matched.
2. Numbers of matches is provided as output.
3. If necessary, a further feature is specified and phases 1 and 2 repeated.
4. On request, compound numbers are given as output as such or after further matching against name/cipher file.

Main frame time involved ca. 100 ms., equivalent to less than 1 shilling per search.

Speed of response governed mainly by input/output facilities used.

As access to the computer system improves via installation of on-line consoles, this should eventually tend to displace manual search tools.

SYSTEM INTEGRATION AND SEARCH STRATEGY

The constraints of manpower, equipment, and costs at any one time lead to the obvious requirement of using the various search facilities mentioned earlier in a well integrated manner.

To achieve in each case the right balance between search speed and cost, one needs a variety of search media. The way in which we are meeting this need is schematically indicated in Figure 10. It shows that the matching with chemical structure typewriter input was expected to be operational by mid-1968, whereas conversion to structural formulas should be in use by late 1968, likewise the mechanized input into a feature card system for chemical structure searches. Although the composite card indexes are now frequently used for rapid consultation (Table VIII), we hope that we will be able to dispense with these centrally held indexes gradually, starting early 1969 and use on-line consultation of the computer-stored files instead. However, as it will take several years before an adequate number of consoles are available throughout the laboratories on the site, we will need to maintain card indexes and feature card systems for use in the laboratories for some time to come.

Cipher and structure typewriter input will be used side-by-side for several years to come, not only because of the error-detection requirements but also in view of the specific merits of each method of input.

The system as outlined provides us with a very large measure of flexibility and adaptibility and therefore meets the two main objectives mentioned—i.e., (*i*) a search strategy with a variety of options, allowing in each case optimum choice of search speed and cost, and (*ii*) gradual and continuous development towards a more highly mechanized storage and retrieval system.

For each of the subsystems, the search parameters are, as a result of extensive use and experience, reasonably

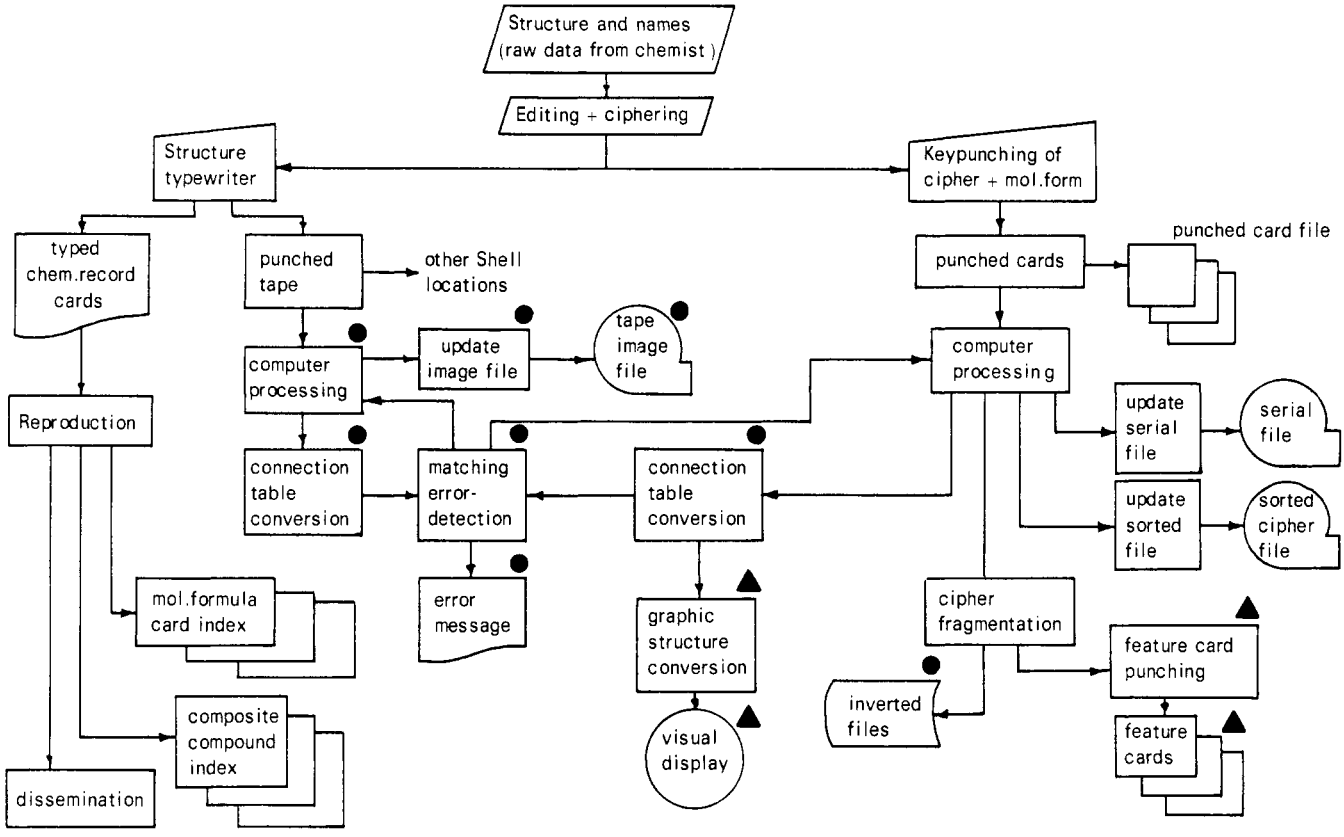


Figure 10. Chemical structure processing

Operational before end of 1966 except:

● target data mid-1968

▲ target date late 1966/early 1969

well-known to us. This, in conjunction with a fairly accurate knowledge of the over-all file composition, allows us in each case to decide which approach is likely to provide the answer in the time and at the cost level appropriate for the case concerned. As yet, cost and speed may vary over a wide range, some searches may warrant an expenditure of say $1000 whereas others should not exceed $1, similarly as regards the response time required, the range may vary from several weeks to several minutes, again a range of roughly 1000 to 1.

Development in the design and performance of the computer operated systems associated with their steadily improving cost to performance ratio should eventually enable us to carry out all searches via the computer system. To achieve this as soon as feasible and at an acceptable cost level is one of our goals.

SYSTEM/USER INTERFACE

It may be appropriate to discuss in greater detail the manner in which we expect the user to interact with the system which contains information on a file of 50,000 research compounds. First of all, it should be pointed out that one of our objectives is to bring the user into direct contact with the system, thus eliminating as much as possible the specialist information worker as an intermediary. Secondly, we do not expect the users in general to become fully acquainted with the formalisms of a notation system; to expect this would be as unrealistic as to assume that the average chemist is fully familiar with systematic nomenclature.

The means whereby our users can interact at present with the system, and are expected to interact with it in the near future, have been indicated earlier in Figure 10 and can briefly be outlined as follows:

1. Consultation of card indexes—i.e., a molecular formula index (in Fletcher notation[12]) and a serial file according to compound code number. These indexes are used for simple look-ups and queries, which however constitute a very large part of all queries handled. The completeness and order of the molecular formula index can be verified by matching against lists produced from our computer-stored file which constitutes our most complete reference file.
2. Queries similar to those for which the card indexes are now used might, by early 1969, be done via direct access of the computer-stored files and molecular formula or compound number as input.
3. Printed lists; our collection is too large to make these feasible for the total file. However, the user receives printouts of compounds relating to his specific query; he can, if desired, be kept aware of structures of interest to him by running his structure profile against all new compounds entered into the system. These searches would be done by Technical Information Services (TIS) using the IUPAC cipher.
4. Feature card (optional coincidence card) system for chemical structures. This we hope to be able to provide to our chemists by mid- or late 1968. The features would correspond as indicated earlier to specific IUPAC cipher fragments

which are automatically selected as well as entered into feature cards. These systems can, if required, be tailor-made to specific user requirements.

5. On-line computer consultation using random access inverted files arranged again according to cipher fragments. This search system, aimed at being available for use by TIS staff in 1968, should, we hope, become available to some users in the laboratories via consoles late in 1968 or early in 1969.

6. Queries entered via a chemical structure typewriter using the type of programs developed by the staff of Shell Development Company. Output via typewriter or CRT. Target data for this facility is late 1968, but CRT output not before 1968. This type of query system is likely to be more expensive to operate than the others mentioned.

7. Query entered via Rand tablet type of input (Grafacon[13]); this would be handled rather like chemical structure typewriter input; output probably via CRT. This might become operational during 1969.

Hence the user would have very little contact with the cipher representation as such. Nevertheless, for several years to come the bulk of the chemical structure processing is likely to be done internally in cipher form and via cipher fragments.

COMPARISON/RELATION WITH ALTERNATIVE CHEMICAL STRUCTURE

In comparing the IUPAC cipher-based system we intend to consider only those systems which deal with complete structural representation (so called unambiguous codes)—i.e., notations and so called topological coding (atom-by-atom coding) systems. Fragmentation coding involves the use of lower level coding languages, which can be derived as sub-sets from the above complete structural representations.

With regard to notations, the Wiswesser and IUPAC notations are perhaps the main contenders, the first one because of its widespread use in the United States, the second one because of its adoption by the IUPAC.

The Hayward notation, which might well have become a third contender because of its origination at the U.S. Patent Office, is still, to our knowledge, in the experimental stage.

With regard to the Wiswesser notation, its strongest point seems to be its American origin resulting in a relatively widespread use in the United States, albeit a use mainly oriented towards mechanical punched card handling and fragmentation coding.[14] The choice of Wiswesser symbols makes it relatively easier to handle on punched card equipment, and this may well have been of significance in the days of mechanical punched card handling. However, this point is now of less significance as notation handling will mainly be done by computer. Moreover, the length of the cipher and its keypunching speed do not differ significantly from those for the IUPAC notation. The same goes for ciphering speed.

With regard to the IUPAC notation, its relationship to graphic structure representation and to systematic nomenclature, and also its handling of ring systems, are worth stressing. After all, we shall still need to use names however much the practicing chemist tends to confine himself to the use of structural formulas.

The structure representation in the IUPAC cipher has, in our experience, proved very effective in computer searches involving scanning of the cipher as such. As a result, computer searches are relatively cheap to operate in the great majority of cases; it seems less clear whether the same will hold for the Wiswesser notation. The efficiency of the IUPAC cipher as a search medium has meant that we could use a computer operated search system in competition with manual searching procedures as far back as mid-1965.

It may be inferred from the above that the main drawback with regard to the IUPAC notation has perhaps been its non-American origin, and the consequent lack of acceptance in the United States. This, associated with the relatively limited computer facilities outside the United States, has tended to restrict the computer application of the notation, which we consider to be essential if the notation is to display its main merit. To our knowledge, the only other recent experimental work with IUPAC has been in Sweden[15] and Japan.[16]

We are therefore of the opinion that the present relative status of the two notations is quite unrelated to their inherent technical merits.

The urgent requirement would now appear to be for research aimed at providing conversion programs and relative testing of the main notations.

The considerable effort on development of topological coding systems during the past five years is in a way unfortunate, in that it has diverted much needed effort from the development of computer systems using notations, which, it is now acknowledged, have at least short term greater economic and operational merit.

The topological coding approach tends to ignore the chemist's view of a structure as an assembly of substructures. In its effort to depict a chemical structure simply as a graph, it abandons much specific chemical interpretation that later on has to be grafted back on to the generalized system. Topological searching is inherently expensive and requires effective screens; the expectation that such screens can be most effectively developed by purely statistical methods seems mainly to have arisen with nonchemists and has yet to be proven.

A subject-oriented coding system such as a notation has, even for computer processing, a considerable advantage over the generalized approach. The type of search problems which have been mentioned as justification for the topological approach are in practice rare; moreover, they can, if necessary, be handled by expansion of the subject oriented coding—i.e., the notation.

The development of topological search programs for chemical structure handling is, of course, of importance for such uses as:

a. Final assessment of the relevance of structures selected in a chemical search.

b. Assessment of structural similarity (presence of largest common substructures).

c. Handling of chemical information presented as structural formulas—e.g., in text handling or during interaction of the user chemist with the search system.

However, for the bulk of present day file handling operations involving large collections of chemical structures, the use of a notation system would appear to offer the advantages of efficiency and economy.

CONCLUSIONS

We should like to end with the following summarizing comments and suggestions for further development.

Use of the IUPAC notation has proved very valuable to us. Its speed in coding and input, efficiency as a search medium, and versatility has enabled us to establish a computer operated search system involving nearly 50,000 compounds, which has been operational since mid-1965.

The total manpower devoted to the various cipher related activities described in this paper has, however, amounted to not more than ca. three man-years in all over a period of five years, hence a very modest effort. Of this effort, ca. 45% was devoted to system development, ca. 30% to input (coding, keypunching), and ca. 25% to actual use of the system for searches.

Despite the fact that we have covered a fairly wide range of applications, we are fully aware that as yet we have only very inadequately explored the potentialities of the IUPAC cipher. Apart from the various aspects described in this paper, we should have liked to be able to devote some effort to applications such as coding of chemical reactions, of Markush formulas (in patent literature), nomenclature/notation conversion, spatial arrangements of atoms, etc.

Although the IUPAC notation has merits which, in our view, might make it preferable to its rivals (relation to systematic nomenclature, to the chemist's view of a structure, search efficiency, etc.), it may well be that the chance of any one notation becoming widely accepted in any form has now passed. None is likely to carry adequate authority. Hence we may expect to see a variety of systems designed to meet the needs of specific environments.

This is perhaps not as serious as it sounds; eventually the user will communicate with the computer system predominantly via structural formula representation—e.g., using CRT/light pen, Rand tablet, etc. How this representation is handled internally will largely depend on the design and facilities provided by one's own system.

In view of the above, there would seem to be an obvious need for research leading to effective computer programs for interconversion of the main notations as well as topological coding. This should prove of considerable value not only for practical but also theoretical purposes.

LITERATURE CITED

(1) Feldman, A., D. B. Holland, and D. P. Jacobus, "The Automatic Encoding of Chemical Structures," Division of Chemical Literature, 141st Meeting, ACS, Washington, D. C., March 1962.

(2) Mullen, J. M., "Atom-by-Atom Typewriter Input for Computerized Storage and Retrieval of Chemical Structures," J. CHEM. DOC. **7**, 88–93 (1967).

(3) IUPAC Commission of Codification, Ciphering and Punched Card Techniques, "Rules for IUPAC Notation for Organic Compounds," Longmans, Green and Co., London, 1961.

(4) Wiswesser, W. J., "A Line-Formula Chemical Notation," T. Y. Crowell Co., New York, 1954.

(5) Hayward, H. W., "A New Sequential Enumeration and Line Formula Notation System for Organic Compounds," U.S. Patent Office Research and Development Report No. 21, 1961.

(6) Lord Todd, Presidential Address, XIXth International Congress of Pure and Applied Chemistry, London, 1963, *Chem. Eng. News* **41** (31), 1956 (1963).

(7) Polton, D. J., "The IUPAC Cipher: Optimising Its Internal Representation for Computer Processing," to be published in *Information Storage and Retrieval* Vol. **4**, 1968.
System," *New Scientist* **31**, 325–27 (1966).

(9) Dyson, G. M., W. E. Cossum, M. F. Lynch, and N. L. Morgan, "Mechanical Manipulation of Chemical Structure: Molform Computation and Substructure Searching of Organic Structures by the Use of Cipher-Directed, Extended and Random Matrices," *Information Storage and Retrieval* **1**, 69–99 (1963).

(10) Polton, D. J. and H. F. Dammers, "Computer Programs for the Conversion of IUPAC Ciphers into Atom-Connection Tables and Their Use in Chemical Structure Searches," unpublished data, 1968.

(11) Dammers, H. F. and D. J. Polton, "IUPAC Cipher Fragmentation and Its Application," unpublished data.

(12) Fletcher, J. H. and D. S. Dubbs, "Quick Access to Research Records," *Chem. Eng. News* **34** (48), 5888 (1956).

(13) Davis, M. R. and T. O. Ellis, "The Rand Tablet. A Man-Machine Graphical Communication Device," Rand Memo No. RM-4122-ARPA, Rand Corp., Santa Monica, Calif., 1964.

(14) National Academy of Sciences-National Research Council, Survey of Chemical Notation Systems, Report of the Committee of Modern Methods of Handling Chemical Information, Publication 1150, Washington, D. C., 1964.

(15) Wurm, B. R., "The Discriminatory Power of the Biological Terms of U.S. Pharmaceutical Patents for Information Retrieval Purposes," Proceedings of the 3rd Annual Meeting of the Committee for Internation Cooperation in Information Retrieval among Patent Offices (ICIREPAT), Vienna, Austria, pp. 277–305, Spartan Books, Inc., Baltimore, Md., 1964.

(16) Uchida, H., Kikuchi, and K. Murayama, "An Algorithm for Mechanical Generation of Matrices from the IUPAC Cipher," Presented before the 2nd National Conference on Documentation in Japan, Tokyo, 1965.