Note

Computer retrieval of carbohydrate references from the current chemical literature

G. G. S. DUTTON AND K. B. GIBNEY*
Department of Chemistry, University of British Columbia, Vancouver, British Columbia (Canada)
(Received March 26th, 1971; accepted June 14th, 1971)

The increase in scientific literature led some years ago to the emergence of publications whose aim was to report, with the minimum of delay, the appearance of papers in certain specific areas. One such current-awareness journal is *Chemical Titles*, which was established in 1961. The advantages of this publication have in practice been somewhat tempered by the lack of reader appeal caused by the special typography necessary and the permuted index system. It is not widely appreciated that *Chemical Titles* and such abstracting services as *Chemical Abstracts Condensates* are also available on tape suitable for machine reading by computer.

The idea of using a computer program for various aspects of information retrieval is not new, but no account of such an application to carbohydrate chemistry has been given. This area of chemistry is fortunate in having a highly stylized and systematic nomenclature[1], and it is this feature that makes mechanical retrieval of information therein attractive and useful. For example, even before their structures have been established, simple carbohydrates are given a name ending in -ose (*e.g.*, mycaminose), and very many compounds may be described by names that are related to glycerose (glyceraldehyde), the two aldotetroses, the four aldopentoses, or the eight aldohexoses; for example, methyl α-D-ribopyranoside, 2-deoxy-D-*ribo*-hexose, and D-*glycero*-D-*manno*-heptose. This aspect of the system is discussed later.

We were invited by our University Library to develop for current carbohydrate literature a retrieval program that would use the CAN/SDI system of the National Science Library, Ottawa. The present note reports briefly the development of our search profile and the results obtained. The program here described refers specifically to the CAN/SDI system. Other systems may use different techniques for searching tapes, but the information available to all systems is the same, and the search methods differ only in technical detail rather than in principle. The interested reader will find the background of the CAN/SDI system reported elsewhere[2–4]. A brief survey of computer searching in the chemical field has been given[5], and a description of the equivalent Danish system has been published[6].

*Present address: Department of Cellulose Research, Columbia Cellulose Ltd., 852 Derwent Way, Annacis Island, New Westminster, British Columbia, Canada.

It must be recognized that the details of any system of retrieval are dependent on the particular interest of the user; hence, the following account does not attempt to be exhaustive. On the other hand, it indicates some pitfalls likely to be encountered in the development of a search profile, and the potential of the method. The final profile presented covers the field of general carbohydrate chemistry. Other, more specialized, topics may then be added by individual users. The profile developed was designed to cover only the chemistry of carbohydrates, not their biochemistry or clinical applications.

Briefly stated, a computer retrieval-program is composed of a series of *profile terms* that are used in particular combinations made according to a group of *search expressions* or *search equations*. The resultant set of machine instructions is known as a *search profile*[7,8].

The profile terms may consist of title words (or parts thereof), authors' names, or journal Codens. The same search profile may also be used in conjunction with *Chemical Abstracts Condensates*, and the profile word-terms then also include key-words. In the United States are several centers that will conduct searches, as well as centers in Canada, Denmark, England, Germany, the Netherlands, and Sweden[9]. Each center has its own subscription rate, but that of the National Science Library, Ottawa, may be taken as representative. At present, the annual charge for searching *Chemical Titles* is $45.50 and that for *Chemical Abstracts Condensates* is $58.50 each for the odd- and even-numbered series. These charges are for a basic profile of 60 terms, and each overterm is charged at the rate of $0.02 per term per tape searched.

Profile terms are given a letter code (for example, A, B,...Z, and then AA, AB,...AZ, and so on), and, although these terms *may* consist of whole words or parts of words, any desired combination of letters may be employed. An important aspect of designing a search profile is the judicious application of "term truncation". Truncation is used for facilitating the retrieval of items containing word fragments that are common to two or more forms of one word. The combination of the systematic nomenclature of carbohydrate chemistry with term truncation is a prime factor in permitting use of a relatively few profile terms in order to cover, successfully, a large segment of the literature. In machine logic, an asterisk(*) following a syllable is an instruction to retrieve words containing *that syllable*, no matter what follows. Symbols may similarly be preceded by an asterisk, or both preceded and followed by one. There are thus four modes of truncation, as illustrated by the following example.

	Truncation	*Retrieval*
1.	GLUCO	GLUCO
2.	GLUCO*	GLUCOMANNAN, GLUCOSE, GLUCOSIDE, GLUCOSIDES
3.	*GLUCO	GALACTOGLUCO, MANNOGLUCO
4.	*GLUCO*	GALACTOGLUCOMANNAN(S)

The position of truncation may be critical in eliminating *excessive* retrieval; this is illustrated by consideration of an appropriate group of carbohydrate terms, such as:

cerebr-o-sides	glyc-o-sides
furan-o-sides	pyran-o-sides
gangli-o-sides	

Truncation as *SIDE* will not only retrieve all of the above, but also: consider, inside, prusside, residence, side, side chain, and sidewise. However, *OSIDE* picks up none of this latter group. This example also serves to illustrate two further points. (*a*) Because of the systematic -OSIDE ending, it is unnecessary to have a special profile word for cerebroside or ganglioside. (*b*) The computer does not automatically retrieve plurals, unless so instructed; thus, glycoside and glycosides are treated as two separate entries. For this reason, it is, in the great majority of cases at least, necessary to follow a profile word by an asterisk. On the other hand, overtruncation may lead to excessive retrieval. In addition to the example just given may be cited *HEX* (which was unsatisfactory because it selected cyclo*hex*ane, *etc.*) and *RIB* (which retrieved dist*rib*ution), whereas HEX* and RIB* have proved useful. In part, this result ensues because, in such a compound word as tri-*O*-methyl-D-ribose, the computer reads "ribose" as a separate word.

Profile terms are transformed* into search expressions by the use of four logic-connectors[7,8]. These are:

AND (+)
Example: A+B implies that A and B *must both* be present.

OR (|)
Example: A|B implies that one (A), *or* the other (B), *must* be present.

AND NOT (¬)
Example: A¬B implies that A *must* be present, but B *must not* be present.

THROUGH (→)
Example: A|B→E, which is read as "A or B through E". The "Through Operator" implies that the first operator is repeated between all profile words.

Each search equation will retrieve a maximum of 99 entries. Therefore, if, in the initial stages of a profile design, it is found that any particular equation is retrieving this number of entries, consideration should be given to breaking down that equation into two smaller ones. As soon as the critical number of 99 entries has been achieved, there is no way of knowing what has *not* been retrieved. Thus, in our first profile, three search equations gave the maximum permitted retrieval. Analysis of the retrievals

*The use of * for truncation, and the symbols given for the logic connectors, are peculiar to the CAN/SDI system. Other systems achieve similar results in various ways.

showed that, in one instance, removal of the term *CELL* to a separate equation would solve the problem. This particular term generated a large number of retrievals based on cellulose and cells, and it was therefore subsequently modified to CELLULOS* in order to eliminate the latter.

In our second profile, it was found that one equation still retrieved too many titles, because of the chance occurrence of two profile terms; thus, (a) "triplet state effects in dye *lasers* at *thresh*old" afforded THRE, ASE, and (b) "binding energy and compressibility of body centered cubic and cl*ose* packed *hex*agonal sodium" afforded HEX, OSE. This problem was overcome by removing certain unneeded common endings, such as *ASE*, and adding a NOT term in order to eliminate such words as adip*ose*, cl*ose*, d*ose*, etc. When it is desired to restrict the number of profile terms used, serious consideration should be given to possible advantages of incorporating a few NOT terms, instead of incorporating words readily found "manually" in *Chemical Titles* (*e.g.*, carrageenan). If there is no restraint on the number of profile terms permitted, this choice does not, of course, have to be made.

A few other minor modifications were made in devising profile 4 (shown in Table I). The journal Coden for *Carbohydrate Research* was added as a check on the performance of the profile. It may be noted that a potential user who regularly reads certain primary journals can easily employ the Codens for these journals in conjunction with NOT, and thus use the machine print-out for retrieval of peripheral articles that might otherwise have been missed.

In the initial stages, retrievals were classified as relevant or not relevant (Yes or No, Table IIA and IIB), and study of Table IIA clearly reveals that the profile had been poorly constructed. However, relatively few changes gave the results shown in Table IIC. At this stage, retrievals were classified as relevant, not relevant, and not relevant but valid. The last category contained carbohydrate terms, but were usually of medical or biochemical, not personal, interest (*e.g.*, effects of ethanol, sorbitol, and thyroid hormones). The non-relevant references usually arose by fortuitous combination of two terms (*e.g.*, purification of *xyl*ene with b*ases*), and occasionally were to important areas of research outside the field of carbohydrates (*e.g.*, mineral composition of herbage browsed by mo*ose* in Alaska!).

Any system for information retrieval must choose between precision and recall[7]. In general, it is better to err on the side of high recall, because the time taken to peruse the print-out and discard unsuitable references is but a matter of moments. In practice, a high percentage of marginal material ("noise") can be tolerated before mechanical retrieval ceases to be competitive with a manual search of the literature. The computer print-out has, in effect, carried out a preliminary screening and has "short-listed" references possibly deserving closer attention. Random checks indicated that the profile in Table I retrieved at least 95% of the current papers of interest to our group. Furthermore, each entry of the print-out was on a separate, perforated sheet of paper (7 × 3.25 in.) which could serve as the basis of a personal filing-system. The time saved in avoiding the necessity of transcribing references of interest should also be considered when assessing the merits of the method.

TABLE I

PROFILE NO. *4*

	Profile words		*Profile words*
A	ALD*	AE	POLYOL*
B	ERYTHR*	AF	*SACCHARI*
C	HEX*	AG	*STARCH
D	KET*	AH	*SUGAR*
E	PENT*	AI	*ARIC*
F	RIB*	AJ	*ASE*
G	TETR*	AK	*FURANO*
H	THRE*	AL	*ITOL*
I	AMYL*	AM	*LACTONE*
J	FUC*	AN	*ONIC*
K	SORB	AO	*OS AMIN*
L	TAL*	AP	*OSE
M	XYL*	AQ	*OSES
N	*ALTR*	AR	*PYRANO*
O	ARAB*	AS	*OSIDE*
P	*FRUCT*	AT	*URONIC
Q	*GALACT*	AU	ADIPOSE
R	GUL*	AV	CLOSE*
S	LYX*	AW	DOSE*
T	MANN*	AX	MANNICH
U	RHAMN*	AY	GLUCAGON
V	*GLUC*	AZ	METABOLI*
W	*GLYCO*	BA	*ENE*
X	CARBOHYDR*	BB	INSULIN
Y	EXUDATE*	BC	GLYCOLY*
Z	GLYCAN*	BD	CORTICO*
AA	GUM*	BE	CELLULOS*
AB	MUCO	BF	CRBRAT
AC	MUCI*	BG	PHOSPH*
AD	POLY HYDROXY	BH	TRANSPORT*

Search expressions

1 T99 (A/B-I) & (AK/AI/AO/AP/AQ/AR/AS/AT) ¬ (AU/AV/AW/AZ/BG/BH)
2 T99 (J/K-M) & (AI/AJ-AT) ¬ (AU/AV/AW/AZ/BG)
3 T99 (N/O-U) ¬ (AX/AZ/BG)
4 T99 (BE/BE)
5 T99 (V/V) ¬ (AY/AZ/BA/BB/BD/BG/BH)
6 T99 (W/W) ¬ (AZ/BA/BB/BC/BG)
7 T99 (X/Y-AC) ¬ (AZ/BB/BH)
8 T99 (AD/AE-AH) ¬ (AZ/BB/BH)
9 T99 (AK/AL/AO/AR/AT) ¬ (AZ/BG/BH)
10 T99 (AP/AQ) ¬ (AU/AV/AW/AZ/BB/BG)
11 T99 (AS/AS) ¬ (AZ/BB/BG/BH)
12 T99 (BF/BF)

The same profile has also been used to search *Chemical Abstracts Condensates* with good results, although the percentage of undesirable retrievals was higher. This is readily understood when it is realized that *Chemical Titles* comprises articles

TABLE II

EVALUATION OF SEARCH PROFILES[a]

Search expression	(A) *Profile 1* *C. T.*[b] *No. 11, 1969*			(B) *Profile 2* *C. T. No. 13, 1969*			(C) *Profile 4* *C. T. No. 20, 1969*			
	Yes	*No*	*Total*	*Yes*	*No*	*Total*	*Yes*	*No*	*No, but valid*	*Total*
1	12	87	99[c]	13	61	74	14	1	4	19
2	8	46	54	9	10	19	4	2	0	6
3	15	34	49	8	22	30	13	1	0	14
4	16	65	81	15	19	19	10	0	7	17
5	30	69	99[c]	12	7	19	10	0	10	18
6	5	94	99[c]	22	20	42	8	2	9	19
7	12	62	74	2	7	9	10	0	2	12
8	4	13	17	11	42	53	10	3	9	22
9	7	11	18	7	15	22	2	3	1	6
10		—		6	17	23	4	3	2	9
11		—			—		2	0	2	4
12		—			—		5	0	1	6
Total	109	481	590	105	220	325	90	15	47	152
Total (%)	18	82		32	68		59	10	31	

[a]Each profile was tested against several issues of *Chemical Titles*. Only 1 representative search is given. [b]C. T. = *Chemical Titles*. [c]Maximum number of retrievals, 99.

from about 800 journals, whereas *Chemical Abstracts* is based on approximately 12,000. The amount of noise may be lessened by citing certain Section numbers of *Chemical Abstracts* as profile terms in conjunction with the logic connector NOT.

Full details of the process leading to the profile given in Table I may be found elsewhere[10] and, although this note has referred only to machine searching of tapes of *Chemical Titles* and *Chemical Abstracts Condensates*, it should be pointed out that other publications are available on tape. These include *Chemical-Biological Activities* (CBAC), *Polymer Science and Technology* (POST), *Basic Journal Abstracts* (BJA), *Information Service in Physics, Electrotechnology, and Control* (INSPEC), and publications of the Institute for Scientific Information (ISI), *e.g.*, *Science Citation Index*. The selection of tapes available depends upon the search center used by a subscriber for his SDI service.

ACKNOWLEDGMENTS

We are grateful to Mr. R. J. Brongers for advice in designing the profiles, and to the National Research Council of Canada for financial support.

REFERENCES

1 Rules of Carbohydrate Nomenclature, *J. Org. Chem.*, 28 (1963) 281.
2 J. E. BROWN, *Special Libraries*, 60 (1969) 501.
3 P. H. WOLTERS AND J. E. BROWN, *Can. Library J.*, 28 (1971) 20.

4 J. HEILIK, *Can. Library J.*, 28 (1971) 120.
5 *Chem. Eng. News*, July 28, 1969, p. 44.
6 H. J. SKOV, *Libri*, 18 (1968) 204.
7 *Preparation of Search Profiles*, Chemical Abstracts Service, Columbus, Ohio, 1967.
8 *Profile Design Manual*, National Science Library, Ottawa, 1970.
9 *Information Services*, Chemical Abstracts Service, Columbus, Ohio, 1971.
10 K. B. GIBNEY, Ph. D. Thesis, University of British Columbia, 1971.

Carbohyd. Res., 19 (1971) 393–399