(Δ^7-cholesterol), Nes et al.[15] reported that "Bernstein et al. have reported a yield of 24% for pure 7-dehydrocholestryl acetate. However, in repeated experiments in this laboratory by essentially the same procedure as used by these authors we were unable to obtain a yield as high as this. The over-all yield of 7-dehydrocholestryl acetate obtained by the 3,5 dinitrobenzoate technique of Bide et al. failed to give yields comparable with those obtained with the isocaproate.

3. Production of steroids is a matter of know-how ("Gewußt wie") more than a matter of equations, and interpretations and the availability of starting materials play a main role.

4. In the German patent law, changing the conditions of production of a steroid when accompanied by a notable increase in the yield permits a new patent for these critical new conditions.[16]

COMPULSORY LICENSE

[In certain countries (other than the United States, but not in the field of atomic energy) a patentee may under certain conditions be required to grant a compulsory license to a party in that country who wishes to practice the invention for a consideration.]

1. The main conditions managing compulsory licenses as nonworking, insufficient working, interdependence of patents, concerning atomic energy are not fulfilled.

2. The fundamental object of the patent is to encourage the replacement of inferior goods and processes by the superior; as the different synthetic steroids have different biological activity; as every company has its own facilities for the production of a special steroid. Hence, compulsory licenses are of no importance in this field.

3. In general, compulsory license is in contradiction to the "laissez faire laissez passez" principle.

4. Djerassi's description of steroids as a "mine of gold" means more research, more discoveries of steroids with physiological activity, and hence no need for compulsory licenses.

LITERATURE CITED

(1) Klyne, W., "The Chemistry of the Steroids," p 14, Methuen & Co., London (1965).

(2) Arago, M., *Mem. Classe Sci. Math. Phys. Inst. Imp. Fr.*, **121**, 93 (1811).

(3) Biot, J. B., *Mem. Acad. Roy. Sci. Inst. Fr.*, **2** (2), 41 (1817).

(4) Snatzke, G., *Angew. Chem.*, **80**, 15 (1968).

(5) Lowry, T. M., "Optical Rotatory Power," p 105, republication, Dover publications, New York, N. Y., 1964.

(6) Moffitt, W., R. B. Woodward, A. Moscowitz, W. Klyne, and C. Djerassi, *J. Amer. Chem. Soc.*, **83**, 4013 (1961).

(7) Klyne, W., "The Chemistry of the Steroids," p 24, Methuen & Co., London (1965).

(8) Jacques, J., H. Kagen, and G. Ourisson, "Constantes Selectionnees: Pouvoir Rotatoire Naturel 1a. Steroids," 2nd. ed, Pergamon, Oxford, 1965.

(9) Kirk, D. N., M. P. Hortshorn, "Steroid Reaction Mechanisms," p 1, Elsevier, Amsterdam, 1968.

(10) Djerassi, C., "Steroid Reactions, an Outline for Organic Chemists," Hodlen-Day, Inc., p ii, San Francisco, 1963.

(11) Cahn, R. S., *J. Chem. Ed.*, **41**, 116 (1964); cf. also ref. cited.

(12) Gilman, H., Ed., "Organic Chemistry, and Advanced Treatise," Vol **II**, p 1530, Wiley, London, 1948.

(13) U. S. Patent 3,383,393, May 14, 1968; Novel 7-Alkyl-19-Norsteroids; granted to Hendrik Paul de Jongh, OSS, Netherlands, assignor to Organon Inc. West Orange, N. J.

(14) In the high court of Eire, before Justice Kenny, November 20, 1967, Farbwerke Hoechst A. G. vs. Intercontinental Pharmaceutical (Eire), Limited. Fleet Street patent law reports, p 187 (1968).

(15) Nes, W. R., R. B. Kostic, and E. Mosettig, *J. Amer. Chem. Soc.*, **78**, 436 (1956).

(16) Roedl, G., Deutsches Patentamt, Zweibrückenstr. 12, 8000 München 2 (private communication).

Computer-Generated Substructure Codes (Bit Screens)*

CHARLES E. GRANITO**, G. THOMAS BECKER, and SCOTT ROBERTS
Institute for Scientific Information, 325 Chestnut St., Philadelphia, Pa. 19106

WILLIAM J. WISWESSER and KURT J. WINDLINX
Fort Detrick, Frederick, Md. 21701

Received January 5, 1971

Substructure searching has been a timely subject for many years now. This is quite understandable since almost all of the questions a chemist takes to his library or computer involve chemical structures. The ability to economically search large files of compounds is the goal of many research projects.

Unfortunately, serial searches of all the structures contained in a large file (>100,000 compounds) can consume considerable amounts of computer time. Efficient screens can greatly reduce computer search time.

This paper presents a set of binary screens computer-generated from Wiswesser Line Notations (WLNs), which greatly reduce computer time for substructure searching of files consisting of hundreds of thousands of compounds. These screens should, therefore, prove of value to any system which includes WLN records.

ICRS (INDEX CHEMICUS REGISTRY SYSTEM)

In 1968, the Institute for Scientific Information initiated its *Index Chemicus Registry System (ICRS)*[1] which consists of a data base containing information being reported in

* Presented in part at the Fifth Middle Atlantic Regional Meeting at the University of Delaware, Newark, Delaware, April 1-3, 1970.

** To whom correspondence should be addressed.

A method for screening Wiswesser Line Notations has been found to be highly efficient in reducing computer search time for substructure searches. The same technique also has been used to study the structural characteristics of new compounds reported in the literature.

Current Abstracts of Chemistry (CAC). Wiswesser Line Notations or WLNs[2] are included for all the new compounds reported in CAC. A total of over half a million compounds has been covered by this service in the few years since it started. A series of computer programs designated RADIICAL (Retrieval and Automatic Dissemination of Information from Index Chemicus and Line Notations) also has been made available. These programs, discussed in an earlier paper, permit searching of all *ICRS* records—titles, authors, addresses, use profiles, analytical codes, index terms, and molecular formulas as well as WLNs. The ability to search WLN records in the *ICRS* tapes provided the first operational substructure search system based on new compounds appearing in the journal literature.

The RADIICAL programs permit "string" searches for any linear sequence or combination of characters. For example, in the case of WLNs, each symbol may be examined to see if it matches the symbol or symbols in the search question. If one wishes to locate all chlorinated pyridines, the RADIICAL routines search WLNs for a T6NJ string (this is the pyridine part), followed by an unspaced letter G (the chlorine symbol). Although very powerful, the RADIICAL program required a character-by-character search. Consequently, all WLN records had to be read. This was no small amount of reading for the current pace of some 17,000 new compounds added each month.

Investigations into ways of shortcutting this character-by-character approach led to the development of sets of binary screens which strikingly reduce the computer time needed to conduct a search. The binary screens function as high-speed switches that permit search programs to skip past most of the WLN records and select only the small number of records that should be examined in greater detail by the RADIICAL programs. This might seem like nothing more than selecting molecular formulas containing N for nitrogen if the search calls for amido, amino, azo, or other nitrogen-containing fragments. The gain in speed and cost reduction is much greater than this, however. It is so much greater that a brief discussion of the differences in computer processing is appropriate at this point.

BINARY "BIT SCREENS"

Most literature chemists know that computers work with binary signals; but few appreciate the tremendous difference in central processing cost between the direct, very high speed handling of binary marks (or bits) as signals in themselves and the tediously indirect handling when combinations of these bits represent alphanumeric information. It is like the difference between recognizing one flashing warning light at a fixed position on a street corner and recognizing what a whole battery of flashing lights is trying to communicate in the leftward-moving message of a newscasting sign.

In any case, bit screens proved that a computer can recognize fragments of the notation up to fifty times faster after first generating fixed-position bit marks in the tape or disc record for each fragment. The computer can also keep a tally on the "scratched" marks and generate a new set of screens for the atomic group marks that appear more than once.

The computer can turn on a new switch when it comes to a ring-starting mark in the WLN record and turn off this switch when it reaches the ring-closing mark. These ring-enclosing marks function mechanically like parentheses in a FORTRAN (formula-translating) or algebraic statement and can be used to separate fragments on the basis of whether they are part of a ring.

ISI's first step in providing binary screens for the RADIICAL program focused attention on the discriminating power of the 40 symbols which are used in the official WLN records—26 letters, 10 numerals, three punctuation marks, and the blank space. The space plays a very powerful role, exactly like the typewriting shift key, in essentially doubling the character set. That is, characters preceded by a space have meanings different from unspaced characters. The WLN symbol set is designed to get maximum usage from both character sets as proven by the fact that the blank space is used far more frequently than the most frequently printed character.

The computer can use the blank space as a switching signal and, at very high speed, generate one set of bits or binary searching screens for the unspaced marks and a different set for the spaced marks.

"FIELDS" OF BIT SCREENS

It does not matter how the bit marks are packed together to make the searching screen. For example, when the first notation-generated bit screen was made on punched cards at Fort Detrick, the last four columns of the card registered the 12 times four or 48 marks that are distinguished in standard card-interpreting or printing equipment. This was doubled to distinguish spaced marks from unspaced marks (see next section), and doubled again to distinguish those inside from those outside the notation's ring-closing marks—which function like parentheses.

When the first such bit screens were put on magnetic tape in ISI's System 360/30, each distinguished set of the 48 marks was packed together as six characters of eight bits each. The ISI record featured one such field for unspaced marks, a second for spaced marks, and a third for a plural number, or multiplicity, of the unspaced marks. (These are atomic group signals in the notation.)

In summary, the first fields of bit screens generated into sorting-machine cards at Fort Detrick had the easily memorized layout shown in Figure 1. The vertical positions for the numbers and letters actually are the same as those manually punched into a demonstration deck of WLN cards made in 1950, 20 years ago. The "special

<	#	,	&
@	%	.	-
SP	$	*	Ø
/	J	A	1
S	K	B	2
T	L	C	3
U	M	D	4
V	N	E	5
W	O	F	6
X	P	G	7
Y	Q	H	8
Z	R	I	9

Figure 1. Bit screen layout for punched cards

SP = blank space

<	U	#	M	,	D	&	4
@	V	%	N	.	E	-	5
SP	W	$	O	*	F	Ø	6
/	X	J	P	A	G	1	7
S	Y	K	Q	B	H	2	8
T	Z	L	R	C	I	3	9

Figure 2. Screen field for 7 track magnetic tape

SP	,	E	M	U	2
.	%	F	N	V	3
<	#	G	Ø	W	4
&	@	H	P	X	5
$	A	I	Q	Y	6
*	B	J	R	Z	7
-	C	K	S	0	8
/	D	L	T	1	9

Figure 3. Screen field for 9 track magnetic tape

48-Bit Field Number	
1	Unspaced and outside the ring marks (chain symbols)
2	Spaced and outside the ring marks (locants for chains)
3	Unspaced and inside the ring marks (cyclic groups)
4	Spaced and inside the ring marks (ring-system positions)
5	Plural or multiple occurrences of above (1) and (3)

Detail: The present WLN "ring marks" are L...J for "aLicyclic" rings and T..J for "heTerocyclic" rings. All bit designations follow those noted in the corresponding Figures 1, 2, or 3.

Figure 5. Description of five-field screen

			Computer search times (min)	
# Profiles	# Terms	WLNs Searched	Without screens	With screens
21	80	11920	150	7
23	86	17189	...	5
20	83	13279	...	3

Figure 6. Typical computer search times for current awareness searches

IBM 360/30 (64K) DOS

characters" or punctuation marks that have become standardized in IBM tabulating equipment since that time are assigned across the top in another easily remembered sequence; zero is followed by *comma* and *period,* then the 3- stroke *asterisk,* the 4- stroke *sharp mark,* the 5-associated *percent* (or "penta") mark, the 6- associated *dollar mark* (spelling $ix), the 7- reflecting "less-than" mark which interprets as a lozenge, and the 8- associated "at" mark ("@ct@l").

The CDC computer at Fort Detrick automatically converts this card-input from a 4 × 12 field to the 8 × 6 magnetic tape field shown in Figure 2.

The 360/30 computer at ISI packs 8 bits rather than 6 into each character or byte, so its pattern of marks for each field screen would look like Figure 3 if you could look the way a 360/30 does.

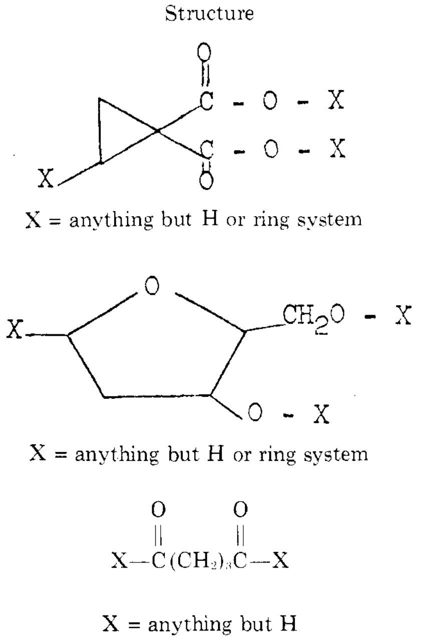
These three explanatory figures emphasize the point previously made: that it really does not matter how the bit marks are packed together by computer, since it alone is concerned with the reading thereof.

THREE-FIELD SCREEN

The three-field screen implemented in ISI's 360/30 computer uses 8 columns or bytes for the first field of *unspaced bits* (UB) in 9-channel tape, or disc records; an identical 8 for the *spaced bits* (SB); and another 8 bytes, or character spaces, for the *plural* or *multiple-occurrence* bits (MB). Thus the ISI tapes are increased by just the equivalent of 18 character spaces when these bit screens are generated; this can be compared with the 123 allowed for WLNs. Figure 4 illustrates these field allotments with an added distinction that the spaced letters are shown here as *lower* case letters, for in the notation diagrams, these ring locants are actually drawn manually (for distinction) as lowercase letters.

	UB						SB						MB					
Track																		
1	A	I	Q	Y	6	#	a	i	q	y	6	#	A	I	Q	Y	6	#
2	B	J	R	Z	7	%	b	j	r	z	7	%	B	J	R	Z	7	%
3	C	K	S	Ø	8	$	c	k	s	Ø	8	$	C	K	S	Ø	8	$
4	D	L	T	1	9	<	d	l	t	1	9	<	D	L	T	1	9	<
5	E	M	U	2	.	@	e	m	u	2		@	E	M	U	2		@
6	F	N	V	3	&	-	f	n	v	3	&	-	F	N	V	3	&	-
7	G	O	W	4	/	,	g	o	w	4	/	,	G	O	W	4	/	,
8	H	P	X	5	*	.	h	p	x	5	*	.	H	P	X	5	*	.

Figure 4. Tape format three-field screen



X = anything but H or ring system

X = anything but H or ring system

X = anything but H

Figure 7. Sample structure questions

= Profiles	= Terms	= WLNs Searched	Total computer search time, min
7	20	153,604	41
14	100	153,604	101
10	46	153,604	57
4	35	185,000	27
1	11	185,000	10
12	65	185,000	45

Figure 8. Typical computer search times for retrospective searches

FIVE-FIELD SCREEN

The five-field screen has a *plural* or *multiple-occurrence* screen identical with that in the three-field screen, but the *spaced* and *unspaced* marks are further differentiated as being *inside* or *outside* the ring-enclosing marks. (These, as previously explained, function exactly like opening and closing parentheses.) These five fields can be characterized as shown in Figure 5.

The five-field screen obviously has greater discriminating power than the three-field screen, but requires more storage space. The two screens are being compared in special tests.

COMPUTER TIME STUDIES

The high value of the three-field screen can be best illustrated by providing a comparison of search times for typical profiles run with and without the screens (*versus* the same WLN files) using the RADIICAL programs. All of these searching times noted in Figure 6 were obtained on an IBM 360/30 (64K) operating under DOS.

Unspaced characters (UB)		Multiple occurrence of unspaced characters (MB)		Spaced characters (SB)	
Symbol	Freq.	Symbol	Freq.	Symbol	Freq.
J	90039	1	45174	Any SC	119203
T	84846	T	37984	B	74566
1	82575	6	37050	D	65757
V	72879	O	33759	C	59176
O	71678	N	32501	E	43152
&	69952	V	31407	A	35490
6	68876	&	31327	F	28007
N	66840	R	24860	&	19100
R	65076	-	18168	G	15696
5	48607	J	15796	H	14968
U	45614	2	14965	I	14502
Y	45104	Y	14211	2	12430
2	39490	U	12385	J	9082
Q	36398	5	11460	O	6270
-	35932	P	10865	K	5664
M	35294	M	9839	L	5550
S	32473	S	7720	M	5039
L	30229	G	7720	N	3948
G	21843	X	5314	P	3285
W	17921	W	5312	1	3236
X	16594	/	5066	3	2151
H	15187	F	4414	R	2043
3	14885	3	2532	Q	1856
/	12422	E	2265	S	1520
Z	11753	L	2207	4	1372
4	10280	Z	2173	T	1093
E	9965	H	1896	X	897
C	8863	4	1779	5	837
I	8653	C	1741	U	661
F	8345	I	1506	0	604
7	8166	0	1136	V	436
P	7825	P	1058	6	296
A	5760	7	504	W	280
K	5199	K	422	8	112
B	3836	A	412	7	77
0	3626	B	371	Z	66
8	2742	8	350	9	51
9	1414	9	160	Y	39
D	1022	D	118	/	34
Peptide	960			-	21

Figure 9. Screen statistics for 153,000 new compounds reported in 1968

Since search times for individual questions vary widely, the times noted in Figure 6 are only generally indicative of the impressive savings to be achieved by using the three-field screen.

Three of the actual search questions included in the data shown as Figure 6 are presented in detail in Figure 7.

RETROSPECTIVE SEARCHES

ISI's RADIICAL programs originally were designed for current awareness SDI (Selective Dissemination of Information) services. However, the considerable savings in computer time for WLN searches provided by the three-field bit screen now makes retrospective substructure searching of *ICRS* tapes an economically attractive possibility.

The computer search times for several one-year accumulations of WLNs in *ICRS* tapes are shown in Figure

8. The fact that large batches of questions could be processed against some 150,000 WLNs on an IBM 360/30 demonstrates that retrospective substructure searches are indeed practical with these linear notations and their self-determined bit screens.

STRUCTURE MAKE-UP

The binary screens discussed above have also been used in studying the structure make-up of new compounds being reported in the literature. For example, Figure 9 gives the binary screen statistics for the 153,000 new compounds reported in 1968.

The symbol J is at the top of the list for primary characters. Nearly 60% of the new compounds reported in 1968 contained a ring system other than benzene. Of these, 33% contained carbocyclic ring systems. About 42% of the new compounds contained an unfused benzene ring. About 47% of the new compounds contained a carbonyl group.

Despite extensive use of contraction rules the "1" (for —CH_3 or —CH_2—) is still the third most frequently used primary character and the most frequent multiple character. Multipliers are used in about 12% of the compounds coded.

A careful review of Figure 9 should be of value to anyone concerned with the make-up of new structures or the design of fragment codes. Studies on screen combinations are planned.

GENERATION TIMES FOR SCREENS

The three-field screens can be generated in less than 5 minutes for the monthly *ICRS* tapes (ca. 17,000 WLNs) and are now provided as part of the *ICRS* system.

AVAILABILITY OF PROGRAMS

The RADIICAL programs referred to above are included in the *ICRS* system provided by the Institute for Scientific Information.

SUMMARY

Substructure searching on files containing hundreds of thousands of compounds is now economically feasible. ISI has developed a series of computer programs for effecting such searches. Part of the efficiency of the programs is directly attributable to the binary screens discussed in this paper.

REFERENCES

(1) Garfield, E., G. S. Revesz, C. E. Granito, H. A. Dorr, M. M. Calderon, and A. Warner, "Index Chemicus Registry System: Pragmatic Approach to Substructure Chemical Retrieval," *J. Chem. Doc.*, **10**, 54-8 (1970).

(2) Smith, E. G., "The Wiswesser Line-Formula Chemical Notation," McGraw-Hill, New York, N. Y., 1968.

Alternatives to Searching Semantic Surrogates of Chemical Structures*

RICHARD I. RUBINSTEIN and ARLENE QAZI
Research and Development Division, BioSciences Information Service
of Biological Abstracts, 2100 Arch Street, Philadelphia, Pa. 19103

Received February 22, 1971

Chemical parameters in BIOSIS' data base are derived from edited, author-specified nomenclature, necessitating "semantic synthesis" of keywords and fragments analogous to chemical synthesis in the laboratory. To obviate this complex synthesis, a pilot file of chemical toxicants was created from the biological literature for study of alternate techniques for chemical information handling. Using synonym indexing, CAS Registry Numbers, and Wiswesser Line-Notation, Toxitapes, a computer file of general, industrial, and pharmaceutical toxicology, was initiated.

BIOSIS' HANDLING OF CHEMICAL INFORMATION

Chemical references in BIOSIS' file are in a biological context rather than about chemical or physical properties. Therefore, chemical information is usually of the nature of pharmacology, chemotherapy, biochemistry, or toxicology rather than organic, inorganic, or physical chemistry.

We determined the extent of chemical information in our file by computer. BIOSIS maintains on-line the indexing assignments made to all references announced in its publications since late 1959. Using C.R.O.S.S. (Computer Rearrangement Of Subject Specialties) we obtained a count of all items indexed in the categories Toxicology, Pharmacology, Chemotherapy, Pollution, and Pest Control for *Biological Abstracts*, Volumes 45-51 (1964-70). This did not include purely endogenous biochemistry.

Biological Abstracts has increased by 31% from 107,000 to 140,000 articles per volume during this 7-year period. The coverage of chemical information, however, increased by 72% from 19,024 to 32,764 articles per volume. It now comprises 23.4% of *Biological Abstracts* whereas 7 years ago it was only 17.8%.

* Presented at the Middle Atlantic Regional Meeting, ACS, Baltimore, Md., February 5, 1971.