möglichkeiten existierten in den ersten Jahren nicht; manche Fragen müssen deshalb an verschiedene Bereiche des Speichers verschieden gestellt werden.

4.4. Zeitbedarf

Verzögerungen bei der *Einspeicherung* waren früher zuweilen recht spürbar. Der Rückstand ist aber fast völlig aufgeholt. Die IDC rechnet für 1970 mit einem Zeitraum von 8–10 Wochen zwischen Eingehen der Referate und Auslieferung des Magnetbandes mit der Codierung an die IDC-Gesellschafter.

Der Zeitbedarf für die *Codierung einer Frage* ist sehr unterschiedlich. Er kann Minuten betragen, manchmal aber auch eine halbe Stunde ausmachen, wenn die Frage mit der Facettenklassifikation schlecht vereinbar ist. Zeitraubend sind fast immer Fragen mit pauschalen Substitutionsangaben, z.B. „0 bis 2 OH-Gruppen und 0 bis 3 Halogenatome im Molekül".

Die *Recherchenzeit* pro Frage wurde bereits mit 20 bis 90 Sekunden angegeben. Die in Assembler geschriebenen Programme haben einen Kernspeicherbedarf von 54 K bytes. Da hiervon nur 12 K auf die eigentlichen Recherchenprogramme entfallen, kann man auch mit kleineren Kernspeichern arbeiten, wenn auch nicht ganz so schnell.

Die Zeit für die Nachbearbeitung der Recherchen hängt naturgemäß von der Zahl der Antworten ab. In aller Regel erhalten die Chemiker aber die Referatkopien binnen 24 Stunden nach Stellung der Frage.

Um die hochgezüchtete Recherchentechnik genau zu kennen, bedarf es einer gründlichen *Einarbeitung*. Da das System aber logisch und gut durchdacht ist, kann es nach einer Anlaufzeit von 4 bis 8 Wochen auch von begabten Chemotechnikern beherrscht werden.

4.5. Kosten

Die Recherchekosten hängen vom jeweils benutzten Computer und dem Abrechnungsmodus des Rechenzentrums ab. Die angegebenen CPU-Zeiten geben jedem Interessenten genügende Anhaltspunkte; die Kosten für die Recherche fallen aber ohnehin gegenüber dem Mitgliedsbeitrag für die IDC wenig ins Gewicht. Hier sind zur Zeit von jedem Gesellschafter je Chemiker rund 1000 DM/Jahr zu zahlen, wobei aber eine Mindestzahl von 50 Chemikern zugrunde gelegt wird. Dieser Beitrag verringert sich, wenn ein Gesellschafter nur Literatur oder nur Patente recherchieren will, er sinkt natürlich auch, wenn sich die Gesellschafterzahl der IDC erhöht. Gesellschafter ohne eigenen Computer können bei der IDC recherchieren lassen.

5. Ausblick

Abschließend kann festgestellt werden, daß uns im IDC-System eine nicht gerade billige, aber zuverlässige und zukunftssichere Recherchenmethode zur Verfügung steht. Anfragen können recht bequem und flexibel verschlüsselt werden. In naher Zukunft soll auch die makromolekulare Chemie in den Erfassungsbereich einbezogen werden, ein weiterer Schritt zum noch weit entfernten Endziel der vollständigen und schnellen Dokumentation der chemischen Literatur der Welt.

Eingegangen am 13. Januar 1970 [A 765]

Vielseitige maschinelle Suchmöglichkeiten nach Strukturformeln, Teilstrukturen und Stoffklassen[1]

Von Ernst Meyer[*]

Das Prinzip der topologischen Formelcodierung und maschinellen Recherche wird kurz erläutert. Durch den Ausbau dieser Methode ist es jetzt möglich geworden, nicht nur nach beliebigen Teilen von Strukturformeln maschinell und ballastfrei zu suchen, sondern dabei auch Fragebedingungen an die Strukturformel zu stellen, die zwar wohldefiniert sind und vom Chemiker gern benutzt werden, sich aber nicht ausschließlich durch Elementsymbole und Bindungsstriche darstellen lassen.

1. Einleitung

Eine der wichtigsten Aufgaben der Dokumentation für die organische Chemie ist das Auffindbarmachen von Strukturformeln und Stoffklassen, die durch gemeinsame Partialstrukturen gekennzeichnet sind. Das geschah bisher meist mit „Fragmentcodes"; mit ihnen verschlüsselt man eine Auswahl von Strukturmerkmalen durch bestimmte Codesymbole, die dann einzeln oder in Kombination miteinander abgerufen werden können. Der Vorteil dieser Codierung liegt vor allem in der leichten Abfragbarkeit. Nachteile sind der intellektuelle Aufwand beim Verschlüsseln und vor allem die Beschränkung auf eine begrenzte Zahl von im Schlüssel vorgesehenen Merkmalen: Sucht man Stoffklassen mit Partialstrukturen, für die es kein eigenes Code-Symbol gibt, so muß man die Frage verfälschen und bei den Antworten Ballast

[*] Dr. E. Meyer
Ammoniaklaboratorium C 6
der Badischen Anilin- & Soda-Fabrik AG
67 Ludwigshafen

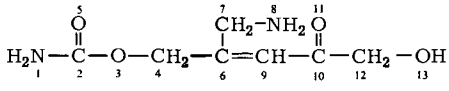
[1] Die Arbeiten wurden zum Teil von der IDC Internationale Dokumentationsgesellschaft für Chemie mbH, Frankfurt, unterstützt.

in Kauf nehmen. Das kann oft sehr lästig werden und viele interessante Fragestellungen aussichtslos machen.

2. Die topologische Methode

Der Einsatz elektronischer Rechenanlagen ermöglichte ein neuartiges Verfahren: Die topologische Methode. Sie wurde für die Strukturformel-Dokumentation 1951 von *Mooers* [2] vorgeschlagen und hat den Vorzug, daß man mit ihr nach ***beliebigen*** Partialstrukturen ballastfrei recherchieren kann, auch wenn deren Wichtigkeit zur Zeit der Verschlüsselung noch nicht bekannt war. Ein Beispiel soll das Prinzip dieser Methode erläutern (Abb. 1).

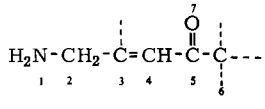
Speicherstruktur:



1	N	1–2.
2	C	1–1, 1–3, ②–5.
3	O	1–2, 1–4.
4	C	1–3, 1–6.
5	O	②–2.
6	C	1–4, 1–7, ②–9.
7	C	1–6, 1–8.
8	N	1–7.
9	C	②–6, 1–10.
10	C	1–9, ②–11, 1–12.
11	O	②–10.
12	C	1–10, 1–13.
13	O	1–12.

Bindungstypen:
1 = Einfachbindung
②= Doppelbindung

Fragestruktur:



1	N	1–2.
2	C	1–1, 1–3.
3	C	1–2, ②–4, ...
4	C	②–3, 1–5.
5	C	1–4, 1–6, ②–7.
6	C	1–5, ...
7	O	②–5.

Übereinstimmung:							
Frage:	1	2	3	4	5	6	
Speicher:	8	7	6	9	10	12	1

Abb. 1. Topologische Strukturformel-Codierung (vereinfachtes Beispiel): Oben ist eine gespeicherte Formel, unten eine geforderte Partialstruktur dargestellt und codiert.

Bei der Einspeicherung wird eine Formel codiert, indem zunächst jedem Atom (außer Wasserstoff) eine Nummer zugeteilt wird, ohne daß man dabei eine bestimmte Reihenfolge beachten muß. Dann wird für jedes dieser Atome eine Zeile gebildet, in der neben seinem Elementsymbol die Nummern seiner „Liganden" (Nachbaratome) angeführt sind. Jeder dieser Ligandennummern unmittelbar vorangestellt wird eine Codezahl für den Typ der Bindung, mit der dieser Ligand am betrachteten Atom hängt. So entsteht eine Verknüpfungstabelle, deren äußeres Bild zwar stark von der zufälligen Reihenfolge der Atomnumerierung bestimmt ist, aus der man aber die Strukturformel eindeutig rekonstruieren kann.

In ähnlicher Weise kann man die Fragestruktur für die Recherche verschlüsseln, wobei freilich die Atomnummern höchstwahrscheinlich anders gewählt sind als in den entsprechenden Speicherstrukturen, in denen sie fündig werden soll. Trotzdem kann der Computer äquivalente Strukturen „erkennen". Im Schema von Abbildung 1 geschieht das etwa in folgender Weise:

Atom Nr. 1 der Frage ist Stickstoff, das erste N-Atom der Speicherstruktur trägt gleichfalls die Nummer 1. Beide haben nur einen (einfach gebundenen) Liganden, in beiden Fällen Kohlenstoffatome. Bei weiterem Vergleich ergibt sich aber nun ein Unterschied: Nummer 2 der Frage soll nur zwei einfach gebundene Liganden tragen, während Nummer 2 des Speichers zusätzlich einen dritten besitzt. Nummer 2 und folglich auch Nummer 1 entsprechen also nicht den Fragebedingungen. Deshalb unterstellt die Maschine nunmehr, daß das nächste N-Atom der Speicherstruktur (Nr. 8) dem Frageatom Nr. 1 entspricht. Wenn das stimmt, muß sein Ligand (Nr. 7) der Frage-Nummer 2 entsprechen. Beide sind Kohlenstoff mit zwei einfach gebundenen Liganden. Dem nächsten Frageatom (Nr. 3) müßte dann der noch nicht belegte Ligand von Speicheratom Nr. 7 entsprechen, nämlich Nr. 6. Beide sind Kohlenstoff, und das Speicheratom hat – wie gefordert – mindestens einen einfach und einen doppelt gebundenen Liganden. Letzterer, nämlich Nr. 9, müßte dann dem Frageatom Nr. 4 entsprechen (vgl. Gegenüberstellung in Abb. 1 unten), und so fort. In dieser Weise wird eine Liste von möglicherweise zutreffenden Atomnummern aufgebaut. An Verzweigungsstellen kommen dabei u.U. für ein Frageatom zunächst mehrere Speicheratome in Betracht und werden vorsorglich notiert. Stößt die weitere Abfrage dann auf Widersprüche, so springt das Programm zur letzten Verzweigungsstelle zurück und verfolgt den dort abzweigenden Weg. Dies wird wiederholt, bis die geforderte (Partial-)Struktur gefunden ist oder alle möglichen Wege vergeblich untersucht worden sind.

Eine ähnliche topologische Recherche wurde etwa gleichzeitig mit der unsrigen beim Chemical Abstracts Service [3] programmiert [4]. Erwähnt sei, daß es auch andere Methoden für den topologischen Strukturvergleich gibt [5], doch hat die oben beschriebene Art der „iterativen" Suche den Vorteil, daß sie weitgehend die Denkvorgänge des Chemikers beim Strukturvergleich simuliert und deshalb auch für höhere Anforderungen ausbaufähig war. So haben wir von Anfang an die Speicherung von Markush-Formeln (d.h. solchen, die eine oder mehrere Stellen aufweisen, die jeweils alternativ mehrere Substituenten tragen können) berücksichtigt und die Recherche auch für diese Formeln programmiert, ohne daß sie in alle denkbaren Einzelverbindungen aufgelöst werden mußten; andernfalls wäre in der Patentdokumentation der Speicher viel zu umfangreich geworden. Auch die unbestimmte Stellung von Substituenten an einem bestimmten Ring konnten wir berücksichtigen.

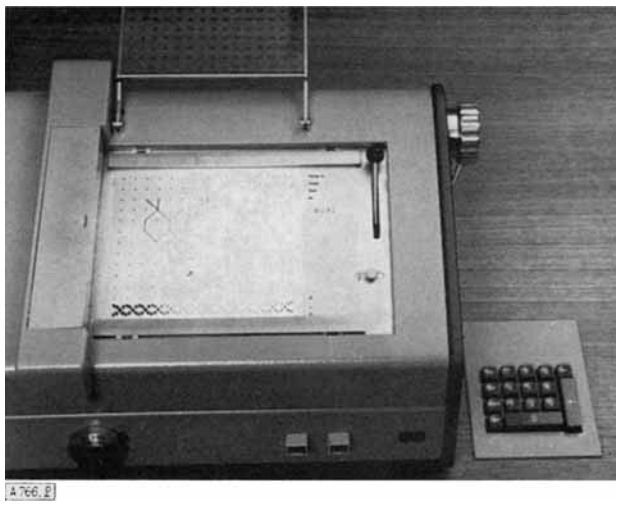
Zwei Nachteile hat die topologische Dokumentationsmethode: Einmal wären bei der Einspeicherung der Schreib-

[2] *C. N. Mooers*, Zator techn. Bull. *59*, 1 (1951).

[3] Für anregende Diskussionen im Rahmen unseres Erfahrungsaustausches danken wir den Mitarbeitern des Chemical Abstract Service (CAS), insbesondere Frau *Myrna Krakiwsky*.

[4] *W. E. Cossum, M. L. Krakiwsky* u. *M. F. Lynch*, Amer. chem. Soc. Meeting, Philadelphia, Pa., April 1964.

[5] *G. Salton* u. *E. W. Sussenguth jr.*, ADI Annual Meeting 1963, Short Papers, Part 2, S. 143; Scientific Report Nr. ISR-6 to The National Science Foundation, Harvard University, April 1964.



A766.2

Abb. 2. Formellesemaschine. Die Struktur wird auf ein transparentes Rasterblatt gezeichnet, das in einen Schlitten eingelegt und von zwölf Photozellen abgetastet wird; dabei entsteht ein Lochstreifen, der auf dem Computer weiterverarbeitet wird.

aufwand und die Fehleranfälligkeit recht groß, wollte man die Verknüpfungstabellen „von Hand" erstellen; zum anderen würde es viel zu viel Maschinenzeit kosten, wenn man alle Strukturformeln eines großen Speichers auf diese Weise abfragte. Wie wir beide Schwierigkeiten umgingen, wurde bereits an anderen Stellen beschrieben[6]: Für die Codierung gerechten und deshalb schnell abfragbaren Überlagerungscode[10]. Nur diejenigen Formeln, die diese beiden Vorselektionsstufen passiert haben — und das ist nur ein sehr kleiner Teil des Speichers — müssen nun noch topologisch untersucht werden, wenn nicht sogar schon die GREMAS-Recherche treffsicher genug war. Die Maschinenkosten dafür fallen selten noch ins Gewicht.

Einer Beschränkung unterlagen die topologischen Verfahren aber bisher noch: Man konnte nicht nach solchen Formelteilen recherchieren, die zwar klar definiert, aber nicht durch Strukturen aus einer bestimmten Anzahl von Atomen und festliegenden Bindungen darstellbar sind, beispielsweise „Carbonsäurederivat", „Alkyl", „Alkinyl", „Cycloalkyl", „Heterocyclyl" oder „Aryl". Derartige Fragebedingungen kommen relativ häufig vor, weil der Chemiker gewöhnt ist, in solchen Begriffen zu denken, und weil er daher z. B. seine Patentansprüche entsprechend formuliert. Wir haben deshalb in unseren Abfrageprogrammen die Möglichkeit geschaffen, eine große Zahl derartiger Fragebedingungen zu stellen und sie auch mit strukturell fest definierten Partialstrukturen zu verknüpfen. Dazu bedienten wir uns zweier Kunstgriffe:

Zunächst errechneten wir uns aus der Verknüpfungstabelle eine Anzahl „redundanter" Angaben. Wir führten dabei besondere Zeilen ein, die nicht ein Atom,

Parameter-Nr. (Adresse)	Elementsymbol	Bindungswert	Ringwert	Zahl d. H-Atome	Zahl d. Liganden	offene Bindungen (Ladung)	Heteroorientierung	Komplex-Bindungen	Flags	1.Ligand		2.Ligand		3.Ligand		4.Ligand		5.Ligand		6.Ligand		Ringzugehörigkeiten			
										Bindgs.-Typ	Nr.	Bindgs.-Typ	Nr.	Bindgs.-Typ	Nr.	Bindgs.-Typ	Nr.	Bindgs.-Typ	Nr.	Bindgs.-Typ	Nr.	Nr.	Nr.	Nr.	Nr.
	a	b	c	d	e	f	g	h	i	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
HAUPTMATRIX MAT																									
MAT	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	5	1	5	1	2	0	0	0	29	3	2	3	3	0	0	0	0	0	0	0	0	19	0	0	0
2	5	1	5	1	2	0	0	0	40	3	1	3	4	0	0	0	0	0	0	0	0	19	0	0	0
3	5	1	5	1	2	0	0	0	42	3	1	3	5	0	0	0	0	0	0	0	0	19	0	0	0
4	5	1	5	1	2	0	0	0	52	3	2	3	6	0	0	0	0	0	0	0	0	19	0	0	0
5	5	1	5	1	2	0	0	0	54	3	3	3	6	0	0	0	0	0	0	0	0	19	0	0	0
6	5	1	5	0	3	0	1	0	65	3	5	3	4	0	7	0	0	0	0	0	0	19	0	0	0
7	7	0	0	1	2	0	0	0	77	0	6	0	8	0	0	0	0	0	0	0	0	0	0	0	0
8	7	1	0	0	2	0	0	0	89	0	7	4	9	0	0	0	0	0	0	0	0	0	0	0	0
9	5	1	0	1	2	0	2	0	101	4	9	0	10	0	0	0	0	0	0	0	0	0	0	0	0
10	5	1	5	0	3	0	0	0	113	0	9	3	11	3	12	0	0	0	0	0	0	20	0	0	0
11	5	1	5	1	2	0	0	0	124	3	10	3	13	0	0	0	0	0	0	0	0	20	0	0	0
12	5	1	5	1	2	0	0	0	126	3	10	3	14	0	0	0	0	0	0	0	0	20	0	0	0
13	5	1	5	1	2	0	0	0	136	3	11	3	15	0	0	0	0	0	0	0	0	20	0	0	0
14	5	1	5	1	2	0	0	0	138	3	12	3	15	0	0	0	0	0	0	0	0	20	0	0	0
15	5	1	5	0	3	0	1	0	149	3	14	3	13	0	16	0	0	0	0	0	0	20	0	0	0
16	7	2	0	0	3	0	0	0	161	4	17	0	15	4	18	0	0	0	0	0	0	0	0	0	0
17	6	1	0	0	1	0	0	0	162	4	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	6	1	0	0	1	0	0	0	160	4	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	4	3	5	0	0	0	6	0	41	0	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0
20	4	3	5	0	0	0	6	0	125	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0
21	160	19	1	2	4	6	5	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	160	20	10	11	13	15	14	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

A766.3

Abb. 3. Topologische Verknüpfungstafel mit redundanten Angaben („Hauptmatrix"), wie sie im Kernspeicher zur Erzeugung des GREMAS-Codes und zur Recherche dient. „Flags" bedeutet die ursprüngliche Punktnummer auf dem Rasterblatt, u.a. zur leichteren Rekonstruktion der Formel.

bauten wir eine besondere „Formellesemaschine"[7] (Abb. 2), die auf Rasterblätter gezeichnete Strukturen mit Photozellen abtastet und für den Computer verarbeitbar macht.

Für die Recherche dagegen entwickelten wir ein zweistufiges Vorselektionssystem[8]: Aus der topologischen Darstellung erzeugt der Computer bei der Einspeicherung zusätzlich die Codierung nach dem bewährten, von *Fugmann* entwickelten GREMAS-System[9] (einem besonders leistungsfähigen Fragmentcode), und daraus wieder einen besonders maschinen-

[6] *E. Meyer*, Angew. Chem. *77*, 340 (1965); Angew. Chem. internat. Edit. *4*, 347 (1965).

[7] *E. Meyer*, Nachr. Dokumentation *13*, 144 (1962).

[8] *E. Meyer*, J. chem. Documentation *9*, 109 (1969).

[9] *R. Fugmann*, Proc. IUPAC Congr., München 1959, S. 331; Classification Research. Munksgaard, Kopenhagen 1965, S. 341; *W. Braun*, *R. Fugmann* u. *W. Vaupel*, Angew. Chem. *73*, 745 (1961); Nachr. Dokumentation *14*, 179 (1963).

[10] *E. Meyer*, Proc.FID/IFIP Joint Conf., Rom 1967, S. 280.

sondern einen Ring beschreiben[11] (beide Begriffe zusammen wollen wir im folgenden mit dem Oberbegriff „Parameter" benennen); und für jeden Parameter hielten wir in unserer „Hauptmatrix", d.h. in der erweiterten Verknüpfungstabelle (Abb. 3), eine Anzahl charakteristischer Angaben fest. Wir benutzten dazu Begriffe wie „Ringwert", „Bindungswert", „Grad der Heteroorientierung" sowie die Zahl der Ringglieder, der anhängenden Wasserstoffatome oder anderen Liganden, der Ladungen etc.; wir notierten sie in der Hauptmatrix und machten sie dadurch rasch abfragbar.

So gibt uns der Ringwert an, zu welchen Typen von Ringen ein Atom gehört (Heterocyclus, aromatischer oder nichtaromatischer Kohlenstoffring). Der Bindungswert dagegen zeigt, an wieviel Doppel- und/oder Dreifachbindungen das Atom beteiligt ist, gleichgültig zu welchen Nachbarn sie führen; diese Angabe ist besonders für Tautomeriefälle wichtig. Der Grad der Heteroorientierung gibt in Anlehnung an GREMAS an, wieviel nicht-ringliegende Bindungen zu Heteroatomen führen. In besonderen Spalten sind ferner die Nummern der Ringe aufgeführt, zu denen das betrachtete Atom gehört oder mit denen der betreffende Ring kondensiert ist. Bei Ringparametern führt zudem eine Verweisadresse zu einer Aufzählung der Ringglieder hin. Ferner unterscheiden wir acht Bindungstypen: Einfach-, Anderthalbfach-, Doppel- und Dreifachbindungen, jeweils in Ring oder Kette liegend.

Außerdem führten wir eine Anzahl von Pseudoelementsymbolen (Abb. 4) für die Fragecodierung ein; sie bewirken, daß bei der Recherche entsprechende Unterprogramme aufgerufen werden, die bestimmte, im Programm jeweils festgelegte Bedingungen prüfen.

Code	Bedeutung	
$A	beliebig, aber kein H	
$B	Heteroatom	
$C	nichtmetallisches Heteroatom	
$D	metallisches Heteroatom	
$I	nichtaromatischer Ring	im Speicher
$J	aromatischer Ring	„Stellung unbestimmt"
$K	Kohlenwasserstoffrest allg.	
$L	Alkyl	unsubstituiert
$M	Cycloalkyl	unsubstituiert
$N	Arylrest	unsubstituiert
$O	Isocyclischer Rest	unsubstituiert
$P	Alkenyl (Lage der Doppelbindung gleichgültig)	
$Q	Alkinyl (Lage der Dreifachbindung gleichgültig)	
$R	Ungesättigte Kohlenwasserstoffkette (Lage gleichgültig)	
$S	hetero- od. ring-substituiertes Alkyl	
$Z	Ringschluß	
HL	beliebiges Halogen	

Bindungswerte (sind vollständig additiv):

Code	Bedeutung
0	nur Einfachbindungen
1	eine (ggf. delokalisierte) Doppelbindung
2	zwei Doppelbindungen (usf. additiv)
6	eine Dreifachbindung

Abb. 4. Liste von Pseudo-Elementsymbolen und Bindungswerten, die bei der topologischen Recherche benutzt werden. Die Maschine verfolgt die angegebene Richtung weiter und prüft, ob keine anderen als die hier angegebenen Reste an dieser Stelle stehen.

So sucht beispielsweise das Unterprogramm für „Alkenyl" in der angegebenen Richtung weiter und prüft, ob die Speicherstruktur dort nur Kohlenstoff- (und Wasserstoff-)atome und neben Ketten-Einfachbindungen nur (mindestens) eine Ketten-Doppelbindung enthält. Durch Einbau weiterer solcher Unterprogramme können wir bei Bedarf noch beliebige andere wohldefinierte, aber strukturell nicht beschreibbare Suchbegriffe einführen.

[11] Dadurch ist es möglich, allgemeine Fragebedingungen an Ringe zu richten wie Gliederzahl, Art und Anzahl der Mehrfachbindungen, Zahl der Substituenten und/oder ankondensierten Ringe etc.

3. Anwendungsbeispiele

Eine große Vielfalt an Fragemöglichkeiten kompliziert zwangsläufig auch die Codierung der Anfrage. Durch Gebrauch eines Formulars versuchten wir, die Frageverschlüsselung möglichst einfach und wenig fehleranfällig zu halten und ihre Kontrolle zu erleichtern; überdies prüft der Computer die Frage auf formale Fehler und innere Widersprüche. Um einen Eindruck von der Vielfalt der Fragemöglichkeit zu vermitteln, wollen wir hier die Codierung der Frage näher erläutern. Sie beginnt – ähnlich wie die von Speicherformeln – zunächst mit dem Zeichnen der geforderten (oder verbotenen) Partialstrukturen und einer Numerierung der Atome (außer Wasserstoff) und Ringe (Abb. 5 und 6).

Die Reihenfolge ist dabei beliebig, doch kann man durch Beachtung einiger Regeln u.U. viel Maschinenzeit einsparen; so empfiehlt es sich beispielsweise, mit einem möglichst seltenen Heteroatom zu beginnen und zunächst in Richtung auf ein weiteres, seltenes Strukturmerkmal zu numerieren. Wie schon angedeutet, kann man an die gleiche Formel im selben Frageformular mehrere Partialstruktur-Fragebedingungen richten, die durch die logischen Verknüpfungen „UND", „ODER" und „NICHT" verbunden werden (Abb. 7); dabei kann man Identität („Überlappung") oder Nichtidentität einzelner Atome verschiedener Teilfragen fordern (Sp. 20–22) und auch die Zugehörigkeit einzelner Atome zu bestimmten Ringparametern verlangen oder verbieten (Sp. 23–25). Darüber hinaus können mehrere Anfragen zu einem „Verbund" zusammengefaßt werden (Sp. 5); sie müssen dann im gleichen Dokument fündig werden, z.B. im Ausgangsmaterial und im Endprodukt.

Die Ausfüllung des Formulars beginnt dann mit einer „1" in Spalte 11 (Gruppenlogik); weitere Teilfragen beginnen in dieser Spalte mit einer ihrer Logik entsprechenden Ziffer. In Spalte 12 und 13 wird dann die Nummer des zu beschreibenden Atoms oder Ringes, hier also zunächst „01", eingetragen, und in Spalte 15–16 das Element- oder Pseudoelementsymbol. Weitere Kennzeichen dieses Parameters trägt man in die entsprechenden Spalten der gleichen Zeile ein.

Die vier letzten Doppelspalten müssen freilich vorerst noch frei bleiben, da unser Programm bisher diese Eigenschaften (Komplexbindungen, Reaktionen, Chiralität und Masse) noch nicht berücksichtigt. Den meisten Spalten, die die Parameter-Eigenschaften beschreiben, ist eine Logik-(L-)Spalte vorangestellt; sie gibt an, ob die betreffende Eigenschaft gefordert oder verboten ist oder ob es sich um eine Minimal-, Maximal- oder Identitätsforderung handelt.

Bei den folgenden Atomen wird – sofern in Spalte 11 nichts eingetragen ist – angenommen, daß sie mit dem jeweils vorangehenden verbunden sind, und der Typ dieser Bindung wird in Spalte 17–19 beschrieben.

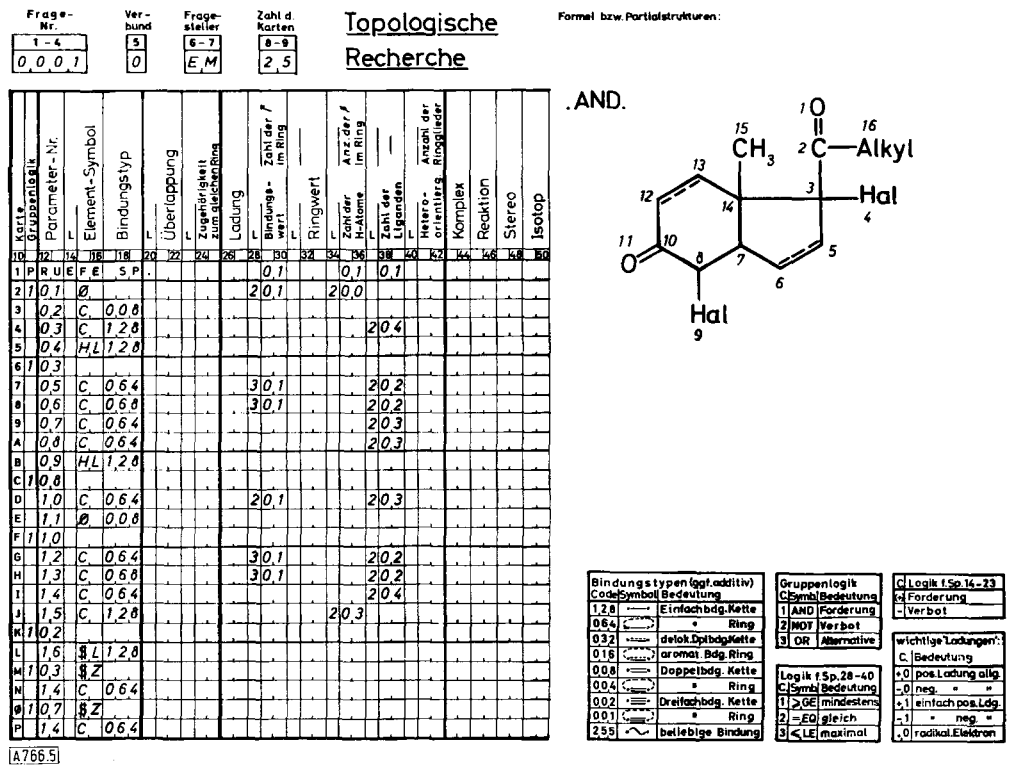


Abb. 5. Beispiel für eine Frage: Gefordert ist die durch die Formel dargestellte Verbindungsklasse; die Ringe dürfen an den bezeichneten Stellen Doppelbindungen, aber keine Substituenten tragen und nicht weiter kondensiert sein.

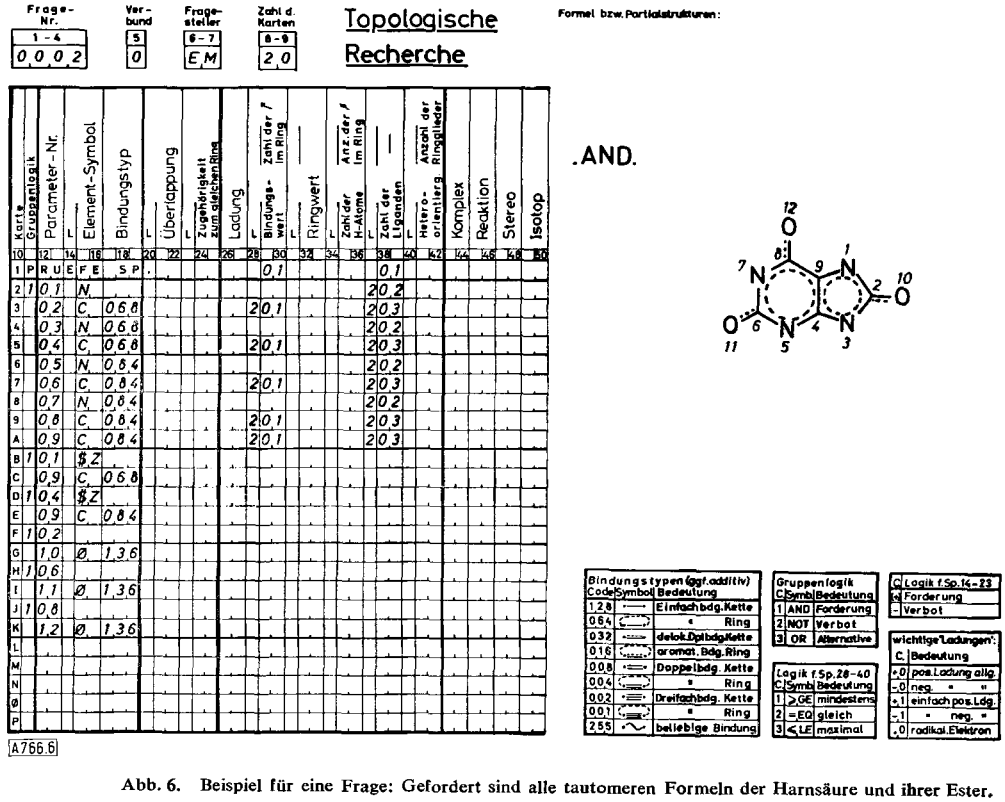


Abb. 6. Beispiel für eine Frage: Gefordert sind alle tautomeren Formeln der Harnsäure und ihrer Ester.

Werden dabei mehrere Bindungstypen im Speicher zugelassen, so werden ihre Codezahlen einfach addiert und die Summe eingetragen. (An den Zweierpotenz-Codezahlen kann der Fachmann schon erkennen, daß wir hier mit Bitketten und Maskenoperationen arbeiten, daß die Angabe von Alternativen also keinen zusätzlichen Aufwand an Maschinenzeit bringt.) Ist eine durchlaufende Kette von Frageatomen auf diese Weise codiert und will man dann zu einem Verzweigungspunkt zurückspringen, von dem aus eine Seitenkette weiternumeriert ist, so muß man zunächst in einer eigenen Zeile in Spalte 11 eine „1" eintragen und in Spalte 12–13 die Nummer des Verzweigungspunktes wiederholen; der Rest dieser Zeile bleibt leer. In ähnlicher Weise werden Ringschlüsse codiert, nur trägt man hier noch in die Elementsymbolspalte „$Z" ein,

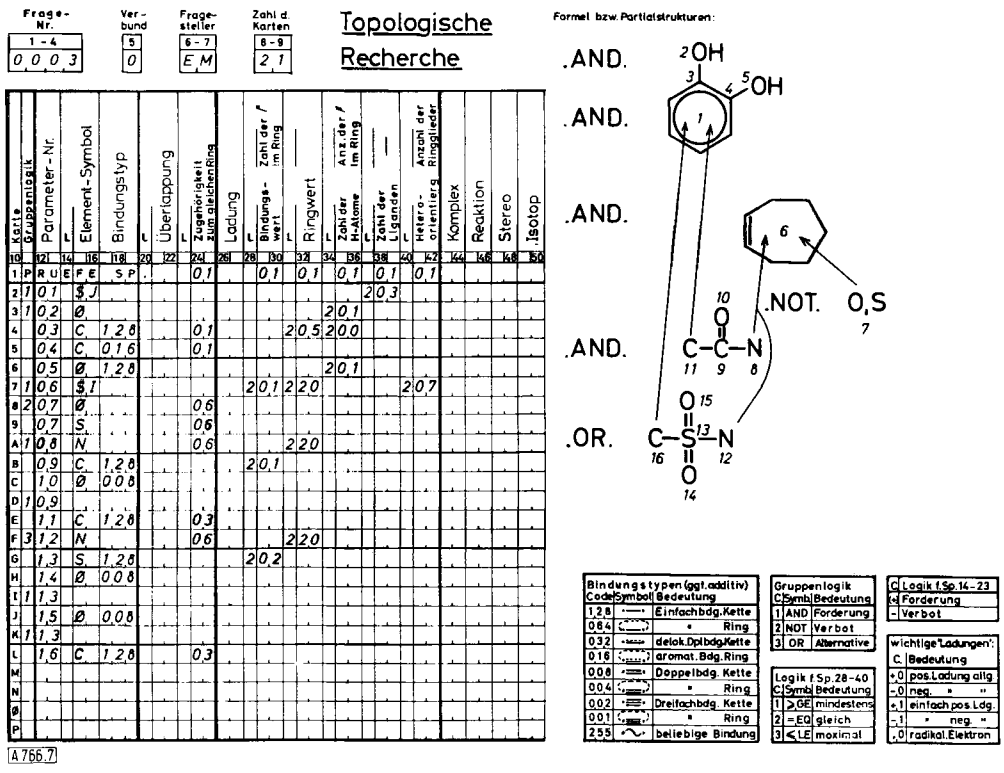
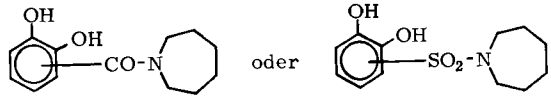


Abb. 7. Beispiel für eine Frage: Gesucht sind alle Verbindungen der Formel



in denen der Brenzcatechinring keinen weiteren Substituenten oder ankondensierten Ring trägt und der siebengliedrige Heterocyclus eine Doppelbindung, aber weder Sauerstoff- noch Schwefelglieder enthält.

und in der folgenden Zeile werden die Atomnummer des Ringschlußpartners, sein Elementsymbol und der geforderte Typ der ringschließenden Bindung angegeben.

Die Spalte „Überlappung" wird verwendet, wenn ein Parameter einer später angeführten Partialstruktur mit einem einer vorher codierten identisch (oder nicht identisch) sein soll; die Nummer des früher zitierten wird hier eingetragen (in der zweiten Teilstruktur war dem Atom oder Ring eine neue Nummer zugeteilt worden). Ebenso wird in Spalte 24–25 die Nummer eines früher zitierten Ringes oder eines Ringatoms eingetragen, wenn das gerade codierte Atom zum gleichen Ring gehören soll (oder nicht gehören darf). Bei Ringparametern haben einige Spalten eine andere Bedeutung als bei Atomen: Statt des Bindungswertes kann die Zahl der Doppelbindungen, statt der Zahl der Wasserstoffatome die der Dreifachbindungen im Ring, und anstelle der Heteroorientierung die Zahl der Ringglieder eingetragen werden, sofern sie gefordert sind. Dadurch werden Fragestellungen möglich wie „Heterocyclus mit mindestens sieben Gliedern und höchstens zwei Doppelbindungen". (Als Ringparameter sind im Speicher allerdings nur Ringe mit maximal neun Gliedern codiert; größere Ringe sind nur als Ketten mit ringschließender Bindung abfragbar.)

Alternativ-Forderungen können, wie erwähnt, als eigene Partialstrukturen mit entsprechender Gruppenlogik codiert werden. Sind nur einzelne Atome oder Pseudoelemente alternativ gefordert, so ist ihre Codierung noch einfacher: Sie erhalten die gleiche Atomnummer und werden mit dieser in aufeinanderfolgenden Zeilen des Formulars beschrieben (Abb. 7, Zeile 8 und 9).

Redundante Angaben, d.h. solche, die sich zwangsläufig aus der übrigen Fragecodierung ergeben, brauchen nicht gemacht zu werden. Oft ist ihre Eintragung trotzdem zweckmäßig, weil die Maschine dann u.U. schneller merkt, daß sie einen Irrweg verfolgt; so kann mitunter Maschinenzeit eingespart werden, doch sollte man berücksichtigen, daß mit der Zahl der redundanten Angaben auch die Gefahr fehlerhafter Fragecodierungen wächst.

Bei einfacheren Anfragen müssen keineswegs alle Spalten des Formulars benutzt werden. Um diese Bedingungen gar nicht in die Abfrage einzubeziehen und somit Maschinenzeit zu sparen, haben wir einen „Spaltenvektor" eingeführt: In der ersten Formularzeile muß in denjenigen Spalten, die abgefragt werden sollen, jeweils eine „01" eingetragen werden; die anderen werden nicht geprüft. Reicht ein Formular für die Codierung der Frage nicht aus, so können weitere angehängt werden, in denen die erste Zeile gestrichen wird. In den Abbildungen 5 bis 7 sind drei konstruierte Fragebeispiele verschlüsselt, die einige wichtige Recherchebedingungen und ihre Codierung zeigen sollen. Auf weitere Möglichkeiten und Einzelheiten kann hier nicht eingegangen werden.

Nachdem die Formulare fertig ausgefüllt und kontrolliert worden sind, dienen sie als Lochbelege. Jede

Zeile wird in eine eigene Lochkarte übertragen. Diese Fragekarten und eine auf Magnetband oder -platte verzeichnete Liste von Formelnummern, die die Vorselektion passiert haben, werden der Maschine mit dem Rechercheprogramm eingegeben und von ihr mit dem topologischen Speicherband verglichen. Dieses enthält – um Einlesezeit zu sparen – die Hauptmatrizen in stark gekürzter Form: Zwar sind die redundanten Angaben noch explizit vorhanden, aber die weitaus meisten Nullelemente der Matrizen sind durch einfache Kunstgriffe unterdrückt; so wird nur wenig Rechenzeit für das Wiederaufblähen im Kernspeicher benötigt.

Als Antworten druckt der Computer entweder die Literaturhinweise oder die Nummern von Referaten aus, die wir gegebenenfalls kopieren und dem Fragesteller zuschicken. Auf Wunsch kann aus der Hauptmatrix maschinell auch die Strukturformel rekonstruiert und ausgedruckt werden. Das geschieht beispielsweise schon während der Einspeicherung zu Kontrollzwecken; allerdings liefern die gebräuchlichen Schnelldrucker kein sehr schönes Formelbild.

4. Schlußbemerkung

Die Computerprogramme für die hier beschriebene Recherche sind für das IBM-System 360 (OS) geschrieben und ausgetestet[12] – mit Ausnahme der erwähnten vier Fragebedingungen. Auch diese werden noch einprogrammiert: Mit „Komplex" sollen koordinative Bindungen und auch wichtige Wasserstoffbrücken abgefragt werden; durch „Reaktion" wird im Rahmen einer Reaktionenfolge ein Atom angerufen, das reagiert hat oder reagieren wird; „Stereo" soll nach einem von *Petrarca, Lynch* und *Rush*[13] vorgeschlagenen System prüfen, ob ein Atom die gewünschte Chiralität oder *cis-/trans*-Stellung aufweist. (Voraussetzung für die praktische Anwendung dieses Verfahrens ist freilich eine wirtschaftliche Einspeicherungsmethode; Programme, die diese mit Hilfe unserer Formellesemaschine erreichen sollten, wurden für die IBM 7090 schon geschrieben, konnten aber noch nicht auf das System 360 umgestellt werden.) Mit „Isotop" sollen schließlich Verbindungsklassen auffindbar gemacht werden, die an bestimmter Stelle markiert sind. Wir hoffen, mit diesem Dokumentationssystem besonders dem präparativ arbeitenden Chemiker ein Werkzeug in die Hand gegeben zu haben, mit dem er nicht nur seine Fragen an den schon erarbeiteten chemischen Wissensschatz schneller und zielsicherer als bisher beantwortet bekommt; wir glauben vielmehr, daß gerade die Möglichkeit zu einer großen Zahl neuartiger Fragestellungen die chemische Forschung befruchten und anregen kann, wenn der Chemiker diese Möglichkeiten erkennt und nützt. Dies könnte seinen Arbeitsstil noch rationeller gestalten, und obendrein würde die Gefahr, längst Bekanntes unfreiwillig nachzuarbeiten, sicher gemindert.

Eingegangen am 11. September 1969 [A 766]

[12] Die topologischen Einspeicherungs- und Rechenprogramme (IBM System /360) schrieb Herr *Peter Schilling*; er steuerte für die Details viele interessante Ideen bei.

[13] *A. E. Petrarca, M. F. Lynch* u. *J. E. Rush*, J. chem. Documentation 7, 154 (1967).

Tosar – ein topologisches Verfahren zur Wiedergabe von synthetischen und analytischen Relationen von Begriffen

Von Robert Fugmann, Herbert Nickelsen, Ingeborg Nickelsen und Jakob H. Winter[*]

Bei mechanisierten Suchsystemen in Literaturspeichern ist ein ständig wachsendes Bedürfnis zu verzeichnen, nicht nur die Fachbegriffe selbst als Suchbedingung formulieren zu können, sondern auch die charakteristischen Verknüpfungen, unter denen diese Fachbegriffe in der Fragestellung erscheinen. Auf diese Weise läßt sich die Treffsicherheit der mechanisierten Literatursuche erheblich steigern. Das System Tosar wurde entwickelt, um die Literatursuche speziell mit elektronischen Rechenanlagen in dieser Hinsicht zu vervollkommnen.

1. Einführung

Mit dem ständigen Anwachsen der chemischen Fachliteratur gewinnen alle Verfahren an Bedeutung, welche es ermöglichen, einschlägige Publikationen zu einer wissenschaftlichen oder technischen Fragestellung wiederzufinden. Die herkömmlichen Registerwerke sind hierfür nur so lange eine gute Hilfe, wie die Fragestellung auf nur einen einzigen Begriff abzielt und nur sofern für dieses Thema ein Schlagwort im Register existiert. Je stärker aber die Spezialisierung in Wissenschaft und Technik sich entwickelt, desto weniger läßt sich ein Fragethema mit nur einem einzigen Begriff, z. B. mit einer Strukturformel, beschreiben. Eine Fragestellung nach Terephthalsäurederivaten oder nach Propylen allein würde heute derartig viel Literatur als Antwort ermitteln, daß sie von niemandem mehr überblickt werden könnte, es sei denn,

[*] Dr. R. Fugmann, Dr. H. Nickelsen, Dr. I. Nickelsen und Dr. J. H. Winter
Farbwerke Hoechst AG
623 Frankfurt/Main-Höchst und
IDC-Internationale Dokumentationsgesellschaft für Chemie mbH
6 Frankfurt/Main