Reaktion (25) antwortet daher nicht auf die genannten falschen Kombinationen.

Durch eine wesentliche Verfeinerung des Codes könnte man den sehr kleinen Anfall nicht zutreffender Karten noch weiter vermindern. Dadurch würde jedoch die Verschlüsselung langwieriger und der Code kompliziert; kaum ein Benutzer wäre dann noch in der Lage, durch eigene Verschlüsselungen die Kartei nach seinem Bedarf zu ergänzen. Überdies würde die getrennte Verschlüsselung unabhängiger Reaktionen im Fall von Parallelreaktionen wie (25) wieder die Fragen nach solchen Reaktionen [z.B. Aminocarbonsäureamide aus Nitrocarbonsäuren bei (25)] unmöglich machen.

Die beschriebene Reaktionenkartei ist eine Gemeinschaftsarbeit von vierzehn Firmen des Dokumentationsringes der chemisch-pharmazeutischen Industrie. Unser Dank für ihre Mitarbeit richtet sich besonders an die folgenden Damen und Herren: O. Dold (Boehringer Mannheim), G. Hallermann (CIBA A.G., Basel/Schweiz), R. Payer (Chemie Grünenthal, Stolberg), H. Lebrecht und M. Scheublein (Knoll A.G., Ludwigshafen), K.-H. Bork (Merck, Darmstadt), U. de la Vaissiere (Roussel-Uclaf S.A., Romainville/Frankreich), H. Seidel (Schering A.G., Berlin), I. Schuler (Thomae GmbH, Biberach) und H. Lehwald (Troponwerke, Köln-Mülheim).

Eingegangen am 13. Mai 1970 [A 772]

CCBF – Ein System zur Computerbearbeitung chemischer und biologischer Forschungsergebnisse[**]

Von Gerhard Ohnacker und Werner Kalbfleisch[*]

Es wird eine firmeninterne, computerorientierte Dokumentation für Ergebnisse aus der Arzneimittelforschung beschrieben. Das System speichert die zahlenmäßigen Resultate aus standardisierten biologischen Testen und verarbeitet chemische Formeln mit topologischen Methoden. Es liefert maschinell gedruckte Karteien und Listen mit chemischen und/oder biologischen Sachverhalten aus dem gespeicherten Material. Seine Programme erlauben Recherchen nach beliebigen Substrukturen und nach pharmakophoren Gruppen; damit kann die Suche nach Beziehungen zwischen chemischer Konstitution und biologischer Aktivität wirksam unterstützt werden.

1. Einleitung

Die Entwicklung einer neuen, biologisch aktiven Substanz bis zu ihrer Einführung als Arzneimittel dauert heute im Mittel sechs bis acht Jahre. Dabei müssen in der Regel einige Tausend neue chemische Verbindungen synthetisiert und biologisch in vielen Richtungen geprüft werden. Der Beweisstandard, der zum Nachweis der biologischen Aktivität, der Brauchbarkeit als Heilmittel und der relativen Ungefährlichkeit gefordert wird, steigt ständig, und es gibt ernstzunehmende Prognosen[1], nach denen die Zeit für die Entwicklung eines neuen Medikamentes, das aus einem größeren Forschungsprojekt resultieren soll, bis auf 15 oder 20 Jahre zunehmen könnte.

Während dieser langen Dauer arbeiten Wissenschaftler vieler Fachrichtungen gleichzeitig oder nacheinander an zahlreichen erfolgversprechenden Substanzgruppen, und mit dem Fortschreiten des Projekts wächst die Zahl der Befunde, die beurteilt werden müssen. Dabei sind Einzelaussagen erst dann verwertbar, wenn sie durch andere, ähnliche Ergebnisse bestätigt werden, und sehr viele Resultate, die zu einer Substanz synchron oder in längeren Zeitabständen eintreffen, sind immer wieder mit anderen Einzelwerten oder auch mit den Wirkungsprofilen anderer Substanzen zu vergleichen.

Solche Wirkungsvergleiche werden um so mehr Hinweise für die Optimierung der biologischen Aktivität durch Substanzgruppen- oder Substituentenvariation liefern,

je größer das kooperierende Team ist, d.h. je mehr und je unterschiedlicheren Testen die erfaßten Einzelergebnisse entstammen,

je sicherer die biologischen Aussagen sind, d.h. je mehr „standardisierte Teste" benutzt werden können, deren Ergebnisse über längere Zeit vergleichbar bleiben und die von äußeren Einflüssen weitgehend unabhängig sind;

je umfangreicher und genauer die physikalisch-chemischen Daten sind, die für Regressions- und Korrelationsanalysen herangezogen werden können.

[*] Dr. G. Ohnacker und W. Kalbfleisch
Chemische Forschung der Dr. K. Thomae GmbH
795 Biberach an der Riß

[**] CCBF = Computerbearbeitung chemischer und biologischer Forschungsergebnisse.

[1] Die Medizin um 1990 – eine technische Prognose des Office of Health Economics, London. Med. Pharm. Studienges. e.V., Frankfurt 1969.

Zur rationellen Speicherung und Bearbeitung der ständig wachsenden Datenfülle bieten sich heute Großcomputer an. Ihre Hilfe wird um so wirkungsvoller sein, je schneller und je zielsicherer der Zugriff zu den angesammelten Formeln und Ergebnissen ermöglicht werden kann und je übersichtlicher das Recherchenprodukt ausgegeben wird. Um diesen Forderungen weitgehend gerecht zu werden, kann die Dokumentation firmeninterner Forschungsergebnisse andere Wege gehen und spezifischere Methoden benutzen als eine allgemein zugängliche Literaturdokumentation zum gleichen Themenkreis.

Jede Literaturdokumentation strebt nach Vollständigkeit. Sie soll alle Publikationen einer wissenschaftlichen Disziplin und alle Veröffentlichungen aus deren Randgebieten erfassen. Sie muß außerdem versuchen, für alle Dokumente die gleiche Erfassungsbreite und -tiefe zu wahren. Dies gelingt aber häufig nur unzureichend, weil sie nicht nur vom tatsächlichen Inhalt der Publikationen, sondern auch von einem mehr oder minder ausführlichen Thesaurus der dokumentationswürdigen Begriffe bestimmt werden und außerdem noch von subjektiven Augenblickentscheidungen der Abstraktoren abhängen. Demgegenüber ist die Dokumentation eigener Forschungsergebnisse im Vorteil. Sie darf bewußt auf verschiedenen Ebenen mit unterschiedlicher Erfassungstiefe arbeiten, weil ihre spätere Nutzung entsprechend den Erfordernissen eines begrenzten Teams sehr genau geplant werden kann.

Wenn jedoch die Forschung von der Dokumentation wirksam unterstützt werden soll, sind die jeweils optimalen Dokumentationsmethoden nicht nur von Art und Umfang der Sachverhalte abhängig, die gespeichert und verarbeitet werden sollen, sondern auch von den spezifischen Wünschen der Wissenschaftler. Deshalb müssen die Wissenschaftler selbst mitarbeiten, sowohl beim Entwurf der Hilfsmittel für die Datenerfassung als auch bei der Festlegung der Formate für die SDI[*] und bei der ständigen Systempflege.

Auf dieser Basis wurde das CCBF-System zur Computerbearbeitung chemischer und biologischer Forschungsergebnisse entwickelt. Wie bei allen Dokumentationen ist auch bei diesem System die Datenerfassung ein Engpaß, und es wurde ganz besonders darauf geachtet, daß die geforderten Auswertungsmöglichkeiten mit einem minimalen Aufwand für die Datenerfassung angeboten werden können.

2. Aufgabe des CCBF-Systems

Das System soll

chemische Strukturen und biologische Testergebnisse speichern;

Daten für die Suche nach Beziehungen zwischen chemischer Struktur und biologischer Wirkung bereitstellen;

dem Wissenschaftler im Abonnement oder auf besondere Anforderung maschinell erzeugte Karteien und Ergebnislisten liefern;

statistische Daten für das Management bereitstellen.

Um diesen Forderungen zu entsprechen, erlauben die Programme des Systems im chemischen Speicher die Suche nach Einzelverbindungen, beliebigen Substrukturen und allgemeinen Formeln. Im biologischen Speicher kann nach Einzelbefunden oder Wirkungsprofilen gefragt werden. Die zutreffenden Formeln und Sachverhalte werden maschinell gedruckt. Dabei kann das Druckbild der Listen den Wünschen des Fragestellers weitgehend angepaßt werden. Tabelle 1

Tabelle 1. Übersicht über Fragen und Antworten beim CCBF-System.

Fragestellung	Recherchenergebnis
Chemische Fragen	
Suche nach einer bestimmten Struktur	Druck des Formelbildes
Suche nach einer bestimmten Verbindungsklasse Suche nach einer allgemeinen Formel Suche nach beliebigen Substrukturen	Druck der Formelbilder aller gefundenen Verbindungen — Die Liste kann nach Deckbezeichnungen oder nach Substituenten geordnet sein
Kombinierte Fragen	
Suche nach der biologischen Wirkung einer bestimmten Substanz	Druck aller biologischen Daten, die zu dieser Substanz gespeichert sind
Suche nach Verbindungen, die in einem oder mehreren Testen wirksam sind; es können Dosislimits gesetzt werden	Druck der Deckbezeichnungen und/oder Formeln der gefundenen Substanzen zusammen mit den pharmakologischen Daten. Die Reihenfolge der Substanzen kann auch nach der Wirkungsstärke geordnet werden
Suche nach Verbindungen mit einer oder mehreren bestimmten Nebenwirkungen	Druck der Deckbezeichnungen und/oder Formeln der gefundenen Verbindungen, aufgelistet nach Dosis, Tier- und Applikationsart
Statistische Fragen	
Zusammenstellung aller in einem bestimmten Zeitraum neu synthetisierten Verbindungen	Liste der Deckbezeichnungen und/oder Formeln, geordnet nach Substanzklassen
Zusammenstellung aller in einem bestimmten Zeitraum berichteten Ergebnisse	Liste geordnet nach den Testnummern und/oder dem Auftraggeber

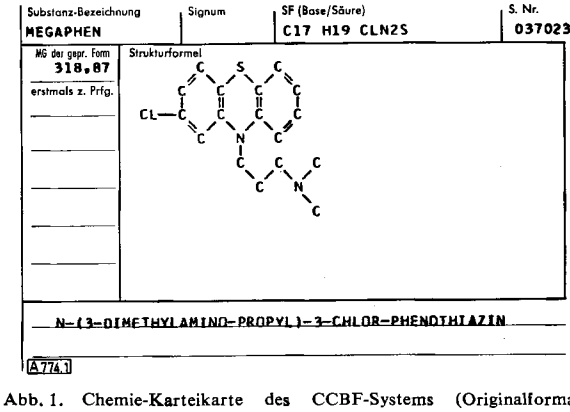


Abb. 1. Chemie-Karteikarte des CCBF-Systems (Originalformat DIN A 6).

[*] SDI = *S*elective *D*issimination of *I*nformation.

nennt einige Beispiele für mögliche Fragestellungen zusammen mit dem jeweils erzielbaren Recherchenergebnis. Neben solchen Recherchen auf besondere Anfrage bietet das CCBF-System auch einen ständigen Dienst in Form chemischer (Abb. 1) und biologischer (Abb. 2) Tochterkarteien. Diese Karteien werden laufend ergänzt und an die Wissenschaftler verteilt.

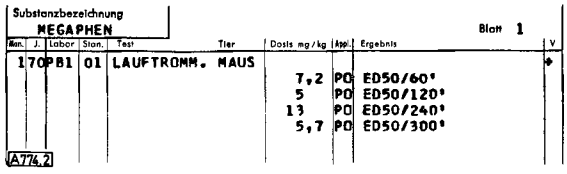


Abb. 2. Biologie-Karteikarte des CCBF-Systems (Originalformat DIN A 6).

3. Umfang der CCBF-Speicher

3.1. Biologische Sachverhalte

Es werden alle Ergebnisse, die beim Screening mit standardisierten biologischen Testmethoden in der Pharmakologie, der Biochemie und der Mikrobiologie anfallen, zahlenmäßig gespeichert. Resultate aus vertieften Untersuchungen einzelner Substanzen werden ebenfalls aufgenommen, allerdings nicht immer mit der gleichen Erfassungstiefe, da sie keine Bedeutung für die Suche nach Struktur/Wirkungs-Beziehungen haben, die sich nur mit standardisierten Testen an mehreren strukturverwandten Verbindungen auffinden lassen. Sie sind jedoch wichtig, wenn nach den Wirkungsprofilen einzelner Substanzen gefragt wird. Dann genügt aber als Antwort ein Hinweis darauf, wann und von wem die entsprechende Untersuchung durchgeführt wurde und wo der ausführliche Bericht zu finden ist.

3.2. Chemische Sachverhalte

Es werden die Strukturen aller neuen Substanzen gespeichert, die im Firmenverband für biologische Prüfungen synthetisiert werden. Außerdem werden die Formeln literaturbekannter Verbindungen aufgenommen, wenn diese Stoffe für eigene biologische Untersuchungen dienten.

4. Dokumentationsverfahren

4.1. Biologische Sachverhalte

Die indirekte Codierung biologischer Sachverhalte mit Zahlenschlüsseln oder Thesauri [2] ist zwar universell anwendbar, aber auch außerordentlich aufwendig. Sowohl die Übersetzung der Sachverhalte in die Zahlenkombinationen oder Schlagwörter als auch die ständige Ergänzung der Nummernsysteme und Register kann nur von Fachleuten nach längerer Einarbeitungszeit mit der notwendigen Treffsicherheit erledigt werden. Das gleiche gilt für direkte Codierungen, die mit „Verschlüsselungsbögen" arbeiten, auf denen die zutreffenden Sachverhalte markiert werden müssen.

Weitaus die meisten biologischen Ergebnisse, die vom CCBF-System gespeichert werden, stammen aus Testen, die nach standardisierten Methoden durchgeführt werden. Daher ist die äußere Form aller Resultate nach jeweils einer der Methoden immer gleich; deshalb wurden für jeden Test ablochfähige Ergebnis-Protokolle entworfen, die am Arbeitsplatz während des Versuches ausgefüllt werden können. Von diesen Protokollen werden die Daten im Klartext oder als Zahlenwerte in den biologischen Speicher übernommen. Abbildung 3 zeigt als Beispiel das Protokoll für

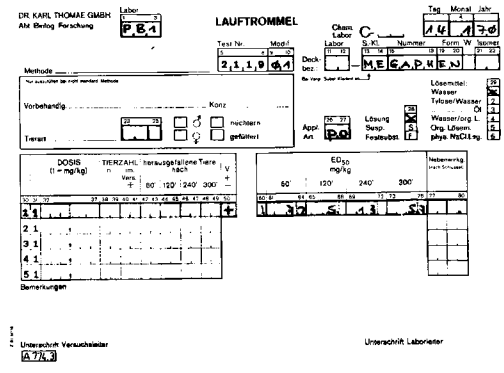


Abb. 3. Ablochfähiges Ergebnisprotokoll zur Aufnahme biologischer Daten in das CCBF-System. Schlüssel werden nur in den Feldern „Test-Nr." und „Modif." für die Testrichtung, im Feld „Tierart" und im Feld „Nebenwirkungen" benutzt. Falls der Versuchsleiter „Bemerkungen" eingetragen hat, werden diese im Klartext übernommen. Solche Bemerkungen sind nicht suchfähig; sie können aber bei biologischen Recherchen mit ausgedruckt werden (Originalformat DIN A 5).

den Motilitätstest mit der Lauftrommel. Die Dateneingabe wird noch einfacher bei Testen, deren Ergebnisse mit biometrischen Methoden maschinell berechnet werden. In diesen Fällen werden auch die Ergebnis-Protokolle (Abb. 4) maschinell erzeugt, die zunächst der Versuchsleiter bekommt, damit er unbrauchbare Versuchsergebnisse eliminieren kann. Nur

BA4 GLUCOSE COLOR. 192001 24.03.70

B -DF 0001 NA

DOSIS 12,500 MG/KG PO RAB

W 180 0 S 7,5 PK

TIERE	ZEIT	MITTELW.	SXQ	AEND.	
21	0	82,5	1,53		
21	60	68,4	2,37	- 17%	SIGN.
20	120	72,5	2,96	- 12%	SIGN.
20	180	73,3	3,34	- 11%	SIGN.
20	240	75,4	3,30	- 9%	N. SIGN.

A774.4

Abb. 4. Maschinell erzeugtes Ergebnisprotokoll für biometrisch berechnete biologische Daten.

[2] Beispielsweise der Biology Code des Chemical-Biological Coordination Center; Herausg. *P. G. Seitner, G. A. Livingston* u. *A. S. Williams*; National Academy of Sciences, National Research Council, Washington, D.C., 1960.

die dokumentationswürdigen Daten werden vom Ergebnisband in den CCBF-Speicher übernommen. Zuerst werden sie maschinell auf formale Fehler und auf Plausibilität geprüft. Formale Fehler werden z. B. angezeigt, wenn die getestete Substanz noch nicht in den chemischen Speicher aufgenommen wurde oder die Satzfolge (Kartenfolge) nicht stimmt. Mit der Plausibilitätsprüfung wird ermittelt, ob die eingegebenen Daten sinnvoll sind; das Programm meldet z. B. unlogische Ergebnisse oder unsinnige Dosisangaben. Der Datensatz aller Ergebnisse, die zu einer Substanz in den biologischen Speicher aufgenommen werden, beginnt mit den Kenndaten der Verbindung, die auch die Verknüpfung zum chemischen Speicher geben. Dahinter stehen als Satzteile in der Reihenfolge des Eingangs die Testresultate, und zwar für jeden Test die Kenndaten (Testnummer, Tierart, Applikationsart usw.), die Dosierungen zusammen mit den zahlenmäßigen Ergebnissen und schließlich eventuelle „Bemerkungen" des Prüfers im Klartext.

4.2. Chemische Strukturen

In den chemischen Speicher des CCBF-Systems werden nur eindeutig definierte Substanzen aufgenommen, Markush-Formeln [*] kommen nicht vor. Die chemischen Formeln können deshalb topologisch verarbeitet werden. Dieses Verfahren, dessen Grundzüge erstmals *Mooers* [3] beschrieb, speichert die Strukturen unverwechselbar und erlaubt Recherchen nach beliebigen Substrukturen. Bei der Entwicklung der Formelcodierung für das CCBF-System wurden die Erfahrungen von *Horowitz* und *Crane* [4], *Waldo* [5] sowie *Ernst Meyer* [6] genutzt. Die Formeln werden von Hand in einen Raster (Abb. 5) übertragen. Dabei

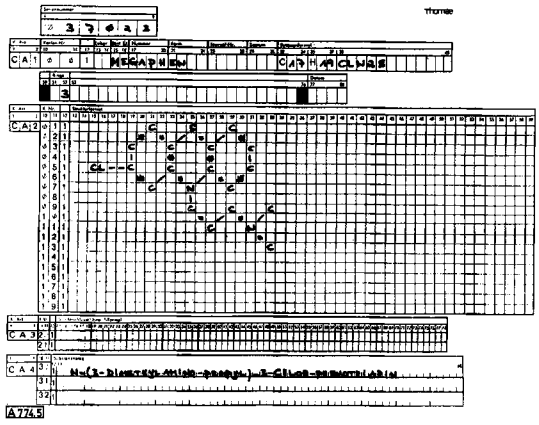


Abb. 5. Codierungsbogen für die Aufnahme chemischer Formeln in das CCBF-System (Originalformat DIN A 4).

müssen einige Regeln beachtet werden, die hauptsächlich durch die Datenerfassungsgeräte bedingt sind.

Für den maschinellen Druck der Formeln wird ein Schnelldrucker IBM 1403/N 1 mit einer besonderen Chemie-Druckkette benutzt. Diese Kette enthält als zusätzliche Sonderzeichen die Einfachbindung \, die Doppelbindungen ||, \ und / sowie die Dreifachbindungen ≡ und |||, außerdem das Zeichen □ zur Kennzeichnung aromatischer Ringe. Um ein ansprechendes Formelbild zu erhalten, werden abweichend von der üblichen Zeilendichte acht Zeilen pro Zoll gedruckt. Da auf der Normaltastatur eines Lochers diese Sonderzeichen nicht vorhanden sind, werden bei der Codierung der aromatische Ring als @, die Einfachbindung \ als ⊁, die Doppelbindungen aller Richtungen als # und die Dreifachbindungen als ⌶ gezeichnet. Das Verschlüsselungsprogramm sorgt dafür, daß die korrekten Symbole in den Speichersatz gelangen. Für die Bindungen sechsgliedriger aromatischer Systeme generiert das Programm den Bindungstyp „aromatisch". Im Hinblick auf ein ansprechendes Druckbild können auch punktierte Bindungen zur Kennzeichnung einer Ionenbindung oder der sterischen Lage, ferner Ladungsvorzeichen eingezeichnet werden. Diese Angaben zu einzelnen Atomen und auch radioaktive Markierungen usw. werden im Sonderschlüssel mitgeteilt. Dieser Schlüssel nennt die „Raster-Koordinaten" (Zeilen- und Spaltennummern des betreffenden Atoms im Raster) zusammen mit einem Schlüsselwort für den Sachverhalt.

Für die topologische Verschlüsselung sind nur die Nicht-H-Atome wesentlich, Wasserstoffatome sind daher nicht zu zeichnen, es sei denn, sie sollen zur Verbesserung des Druckbildes in den Satzteil für die druckfähige Formel übernommen werden. Neben den beschreibenden Daten wie Substanzbezeichnung, Synthesedatum und dem chemischen Namen oder bibliographischen Angaben zur Substanz werden für Prüfzwecke auch die Summenformel und die Anzahl der Ringe, die im Molekül vorhanden sind, angegeben.

Im Mittel können 40 Formeln pro Stunde codiert werden. Die Codierung ist optisch leicht zu kontrollieren und außerdem zukunftssicher, da sie später eventuell auch von Klarschriftlesern [7] verarbeitet werden könnte. Außerdem wird bei der Eingabe die Information für den maschinellen Druck der Formeln ohne zusätzliche manuelle Arbeit miterfaßt. Die Eingabedaten werden maschinell auf formale Richtigkeit geprüft. Ungültige Atom- oder Bindungssymbole, fehlende Zeilen der Formelmatrix und ähnliche Formfehler werden gemeldet.

Zur Erzeugung der topologischen Bindungstafel (Connectivity Table) werden die Atome einer Formel maschinell numeriert. Außerdem ermittelt das Programm die Zahl und die Art der Bindungen, die von jedem Atom ausgehen. Wasserstoffatome, die nicht mit einem Sonderschlüssel gekennzeichnet sind, werden nicht berücksichtigt. Zur Erzeugung der kompakten Darstellung [8] (Compactlist) entfernt das Programm die redundante Information aus der Bindungstafel. Im Interesse höherer Selektivität und damit höherer Geschwindigkeit bei der Recherche werden dann jedoch abweichend von den topologischen Verfahren, die *Gluck* [8] und *Morgan* [9] beschrieben haben, vom CCBF-Verschlüsselungsprogramm weitere Informationen über das einzelne Atom ermittelt und in die Compactlist aufgenommen. Es sind dies die Infor-

[*] Eine Markush-Formel ist eine allgemeine Formel, z.B. R^1NH_2, in der R^1 z.B. die Bedeutung Alkyl von C_1 bis C_{10} haben kann.

[3] *C. N. Mooers*, Zator Techn. Bull. *59*, 1 (1951).

[4] *P. Horowitz* u. *E. M. Crane*: Hecsagon: A System for Computer Storage and Retrieval of Chemical Structure. Eastman Kodak Company, Rochester, N.Y., 1961.

[5] *W. H. Waldo*, J. chem. Documentation *2*, 1 (1962).

[6] *E. Meyer*, Angew. Chem. *77*, 240 (1965); Angew. Chem. internat. Edit. *4*, 347 (1965).

[7] *W. E. Cossum, M. E. Hardenbrook* u. *R. N. Wolfe*, Proc. Amer. Documentation Inst. *1964*, 269.

[8] *D. J. Gluck*, J. chem. Documentation *5*, 43 (1965).

[9] *H. L. Morgan*, J. chem. Documentation *5*, 107 (1965).

mationen, ob das Atom Glied eines Ringes ist oder nicht sowie die Zahl der mit einem Atom verbundenen Wasserstoffatome (diese Angabe verkürzt die Recherchenzeit, wenn in der angefragten Struktur oder Substruktur H-Atome an bestimmten Stellen gefordert worden sind).

Im Speicher ist jedes Atom einer Struktur beschrieben durch

den Code für die Atomart,

die Angabe „Ringglied oder Nicht-Ringglied",

die Angabe „Bindung von Atom Nr. . . .",

den Code für die Bindungsart,

den Code für die Zahl der H-Atome, mit denen das Atom verbunden ist,

die Codes für eventuelle Sonderangaben.

Einige dieser Codes sind Bitketten, damit logische Bedingungen (und/oder/nicht) mit einfachen Booleschen Operationen abgefragt werden können.

In einer rechnerischen Prüfung der Formel ermittelt das Programm die Summenformel und die Zahl der Ringe und vergleicht beides mit den eingegebenen Werten. Fehlerhafte Codierungen werden zurückgewiesen.

Auch bei den leistungsfähigen Computern der dritten Generation erfordert der iterative Atom-für-Atom-Vergleich nach *Ray* und *Kirsch* [10] noch immer viel Maschinenzeit. Deshalb arbeitet auch das CCBF-System mit einigen Vorprüfungen. Diese Screens werden maschinell erzeugt. Zur Zeit benutzt das System

einen Summenformelscreen;

einen Zweier-Fragment-Screen, der angibt, wie oft Mehrfachbindungen zwischen zwei C-Atomen sowie Einfach- und Mehrfachbindungen zwischen C- und Heteroatomen oder zwischen zwei Heteroatomen vorkommen; Einfach-, Doppel-, Aromaten- und Dreifachbindungen werden getrennt gezählt;

einen Ringscreen, in dem jeder Basisring des Moleküls beschrieben ist. Dieser Ringscreen entspricht etwa den Ringfragmenten des Ringcodes [11].

Durch diese Vorprüfungen werden mindestens 90%, in manchen Fällen über 99% der gespeicherten Verbindungen sehr schnell als „nicht zutreffend" ausgeschieden. Nur beim Rest ist zur Beantwortung der Frage ein topologischer Vergleich notwendig.

Der Speichersatz für eine Strukturformel enthält schließlich

die Kenndaten (bei firmeninternen Substanzen die Deckbezeichnung, bei literaturbekannten Verbindungen den Namen im Klartext),

die maschinell erzeugten Screeninformationen,

die maschinell erzeugte Compactlist,

die Information für den Druck der Formel (eine maschinell verdichtete Darstellung der Eingabedaten),

den chemischen Namen und/oder die Bibliographie.

Durchschnittlich hat ein solcher Speichersatz 330 Stellen, so daß eine Magnetbandspule etwa 120000 Formeln faßt. Das CCBF-Verschlüsselungsprogramm verarbeitet im Durchschnitt 150 Formeln pro Minute.

5. Recherchierprogramm

Neben dem ständigen Dienst zur Erzeugung von Tochterkarteien (Abbildungen 1 und 2) ermöglicht das CCBF-System auch Recherchen nach biologischen und chemischen Sachverhalten sowie nach deren Verknüpfung in komplexen Fragestellungen (Tabelle 1).

5.1. Codierung der biologischen Fragen

Bei biologischen Recherchen kann jedes Feld des biologischen Datensatzes angesprochen werden. Innerhalb gleicher Satzteile sind logische Verknüpfungen möglich. Alle Fragen nach Dosierungen und Ergebnissen müssen mit den betreffenden Testnummern verbunden sein. Nach einigen Daten, z.B. nach Angaben zur Toxizität (Prozent der Tiere, die während des Versuchs gestorben sind) und nach Nebenwirkungen, kann auch global gefragt werden. Die biologischen Fragen werden weitgehend in freier Form in der Folge Schlüsselwort – Operator – Suchargument codiert. Mit den Schlüsselwörtern werden diejenigen Datenfelder des biologischen Speichersatzes bezeichnet, die in der Recherche angesprochen werden sollen. Operator und Suchargument nennen alle Bedingungen, die zur Beantwortung der Frage erfüllt sein müssen oder ausgeschlossen werden können. Als Operatoren sind „gleich", „ungleich", „größer als" und „kleiner als" erlaubt. Als Suchargument können auch Zahlenbereiche, z.B. ED 50 „zwischen 10 und 50 mg/kg", benutzt werden.

5.2. Codierung der chemischen Fragen

Im Formelspeicher kann nach bestimmten Strukturen, nach allgemeinen Formeln und nach Substrukturen beliebiger Art und Größe gefragt werden. Substrukturen sind Teile der Atomfolgen aus vollständigen Strukturen; sie können geradkettige, verzweigte und ringförmige Elemente oder Ringteile allein oder in beliebiger Kombination enthalten.

Für eine chemische Recherche wird zunächst die Bindungstafel der angefragten Struktur oder Substruktur manuell codiert. Wenn dabei diejenigen Teile der Struktur möglichst niedrig numeriert werden, die im Bezug auf alle gespeicherten Formeln „selten" sind, kann dies die Suchzeit verkürzen [12]. Außerdem werden alle Screens codiert, die für die angefragte Struktur oder Substruktur sinnvoll sind. Für topologische Recherchen nach bestimmten Verbindungen wird zusätzlich die Zahl der H-Atome angegeben, mit denen jedes Nicht-H-Atom der Formeln verbunden sein soll. Dadurch wird erreicht, daß die Recherche

[10] *L. C. Ray* u. *R. A. Kirsch*, Science (Washington) *126*, 814 (1957).

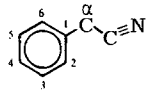
[11] *W. Steidle*, Pharm. Ind. *19*, 88 (1957).

[12] *W. E. Cossum, M. L. Krakiwsky* u. *M. F. Lynch*, J. chem. Documentation *5*, 33 (1965).

tatsächlich nur die explizit angefragten Formeln trifft.

Bei topologischen Recherchen nach Substrukturen und allgemeinen Formeln ist es möglich, anstelle eines oder mehrerer bestimmter Atome „Dummy-Symbole" einzusetzen, z.B. „beliebiges Halogen", „beliebiges Heteroatom", „N oder O oder P oder S" und „C oder O oder N oder S". Damit können z.B. Gruppen isosterer Verbindungen in einer Frage gesucht werden. Bei der Codierung der Bindungsarten können für jede Bindung „oder"- oder „nicht"-Kombinationen gewählt werden, z.B. „Einfach- oder Doppelbindung", „Einfach- oder Aromaten- oder Doppelbindung (= nicht-Dreifachbindung)" und „beliebige Bindung".

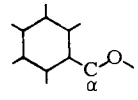
Durch die Angabe der Zahl der Wasserstoffatome, die mit einem Nicht-H-Atom einer Substruktur oder allgemeinen Formel verbunden sein sollen, kann auch verlangt werden, daß z.B. in der Substruktur *(1)* das



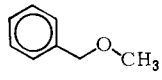
(1)

α-C-Atom mit zwei H-Atomen und die C-Atome 3, 4, 5 und 6 des Phenylrestes mit je einem H-Atom verbunden sein sollen. Dadurch werden bei diesem Beispiel nur Phenylacetonitril und diejenigen Mono-*ortho*-Derivate gefunden, die in α-Stellung nicht substituiert sind.

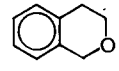
Durch die Angabe, ob ein Atom Glied eines Ringes sein soll oder nicht, kann z.B. gefordert werden, daß in der Substruktur *(2)* das α-C-Atom ein Ringglied ist; dann trifft nur die Formel *(4)* zu. Die nicht eingeschränkte Frage würde auch von der Struktur *(3)* positiv beantwortet werden.



(2)



(3)



(4)

5.3. Recherche

Das Recherchierprogramm übersetzt zuerst die auf Lochkarten eingegebene Codierung der biologischen und/oder chemischen Fragestellung in interne Fragesätze. Dabei wird – analog zu den Prüfungen bei der Dateneingabe – auf formale und inhaltliche Fehler geprüft. Bei kombinierten Recherchen kann gewählt werden, ob der chemische oder der biologische Speicher zuerst abgefragt werden soll. Enthält die Frage z.B. einen seltenen biologischen Test, so wird zuerst der biologische Speicher durchsucht. Andererseits wird die chemische Frage vorangestellt, wenn sie Strukturteile betrifft, die im Speicher selten vorkommen.

Das Recherchenprogramm bringt die Datensätze für alle zutreffenden Substanzen auf einen Zwischenspeicher, aus dem das Druckaufbereitungsprogramm diejenigen Teile selektiert, die für das gewünschte Ergebnis (Tabelle 1) erforderlich sind. Die Anweisungen für die Druckaufbereitung werden zusammen mit der Anfrage eingegeben. Es ist vorgesehen, daß Anweisungen für die Druckaufbereitung auch noch nach der Recherche eingegeben werden können, damit die gesamte Information für alle zutreffenden Substanzen mehrfach und unterschiedlich nutzbar wird.

6. Erfahrungen

Mit dem CCBF-System wurde 1968 nach zweijähriger Entwicklungsarbeit eine Forschungsdokumentation modernisiert, die vorher etwa acht Jahre lang chemische Formeln mit dem Ringcode[11] und biologische Sachverhalte mit direkter Codierung erfaßt hatte. Unsere Erfahrungen mit diesen Dokumentationssystemen erlauben die Aussagen,

daß eine Forschungsdokumentation, die ein wirksames Hilfsmittel sein und bleiben will, ihre Methoden laufend der Entwicklung auf dem Gebiet der Datenverarbeitung und dem Fortschritt bei den Dokumentationsverfahren für die Sachverhalte anpassen muß;

daß bei einer Dokumentation von chemischen Formeln aus der Arzneimittelforschung reine Fragmentierungscodes nicht selektiv genug arbeiten und daß vor allem im Hinblick auf die Recherchen nach beliebigen Substrukturen nur topologische Methoden allen Ansprüchen genügen;

daß eine Dokumentation biologischer Sachverhalte nur dann rationell arbeitet, wenn die Datenerfassung so einfach ist, daß für die Codierung keine Fachkräfte benötigt werden;

daß eine Forschungsdokumentation für die Wissenschaftler nur dann attraktiv ist, wenn sie neben der Möglichkeit, individuell zu recherchieren, auch einen ständigen Dienst in Form einer SDI bieten kann;

daß eine Forschungsdokumentation nur dann aktuell bleibt, wenn die Dokumentation ein Instrument der Forschung ist und die Wissenschaftler, deren Forschungsergebnisse gespeichert werden, bei Systemerweiterungen und bei der Systempflege ständig mitarbeiten.

Die Programme des CCBF-Systems wurden in Zusammenarbeit mit der Abteilung Zentrale Datenverarbeitung der Firma C. H. Boehringer Sohn, Ingelheim, teils in Assembler, teils in PL/I für eine IBM 360/40 geschrieben. Herr Dr. B. Braun, Ingelheim, koordinierte die Arbeiten an den Teilen des Programms und unterstützte neben der Systemanalyse auch die Systempflege und die ständigen Systemerweiterungen mit vielen wertvollen Anregungen. Die umfangreichen Planungen für die Computerbearbeitung biologischer Befunde hat Fräulein U. Zech mit vielen praktikablen Ideen sehr wirkungsvoll unterstützt. Herr J. Becker schrieb das Verschlüsselungsprogramm für die chemischen Formeln und steuerte sehr nützliche Details bei. Den biologischen Teil programmierte Herr J. Gruber; seine scharfsinnigen Vorschläge kamen der Entwicklung des Systems sehr zustatten. Die Änderungsdienste und die Druckanforderung hat Herr P. Oppitz mit vielen förderlichen Einfällen programmiert.

Eingegangen am 25. Mai 1970 [A 774]