51. Strukturaufklärung organischer Verbindungen durch computerunterstützten Vergleich spektraler Daten

von **F. Erni** und **J. T. Clerc**

Organisch-Chemisches Laboratorium, Eidgenössische Technische Hochschule, Zürich

Herrn Prof. Dr. *A. Wettstein* zum 65. Geburtstag gewidmet

(10. I. 72)

Summary. A system for computer-aided identification of organic compounds by means of spectroscopic data is described. The binary coded spectral data of the unknown compounds are compared with a reference file containing NMR., IR. and Mass Spectra. A highly flexible software automatically selects the appropriate search strategy and directs the updating of the collection, thereby ensuring continuous adaption to varying needs.

1. Einleitung. – Bei der Strukturaufklärung organischer Verbindungen mit Hilfe spektroskopischer Methoden wird das analytische Instrument heute weitgehend im Sinne einer «Black Box» eingesetzt [1]. Der Wunschtraum des Chemikers, dieses System durch eine weitere «Black Box» zu erweitern, die die Daten zu brauchbaren Strukturvorschlägen verarbeitet (Fig. 1), scheint beim jetzigen Stand der elektronischen Datenverarbeitung nicht mehr unverwirklichbar.

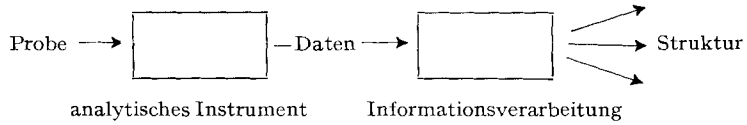


Fig. 1. *Erweiterte «Black Box»-Philosophie: Wunschtraum des organischen Chemikers*

Es zeigen sich heute zwei grundsätzlich verschiedene Wege, um aus den spektralen Daten einer unbekannten Verbindung sinnvolle Strukturvorschläge zu erhalten. Bei dem einen versucht man, den bei der Interpretation durch den Analytiker ablaufenden Denkprozess möglichst getreu auf einen Computer zu übertragen, den Analytiker also auf dem Computer zu simulieren. Dies scheitert aber daran, dass auch die grössten heute existierenden Computersysteme bei der Bewältigung von komplexen Entscheidungsvorgängen dem menschlichen Gehirn noch um viele Grössenordnungen unterlegen sind; dementsprechend sind diese Resultate trotz teilweise beängstigend hohem Aufwand im Vergleich mit der Leistungsfähigkeit eines Analytikers bemerkenswert bescheiden [2]–[5]. Zudem kann das Computersystem nur über explizite ins Programm eingebaute Korrelationen verfügen. Nicht ganz eindeutig mathematisch formulierbare Zusammenhänge, die bei der semiempirischen Deutung spektroskopischer Resultate häufig eine Rolle spielen, sind daher von vornherein nicht verwertbar.

Der zweite Weg besteht darin, unter Verzicht auf eine Simulation des Analytikers ein den Stärken und Schwächen elektronischer Datenverarbeitung besser angepasstes

System zu verwenden, beispielsweise das Aufsuchen von Referenzverbindungen durch computergesteuerten Vergleich mit einer Sammlung von Referenzspektren [3] [6] [7] [8] [9].

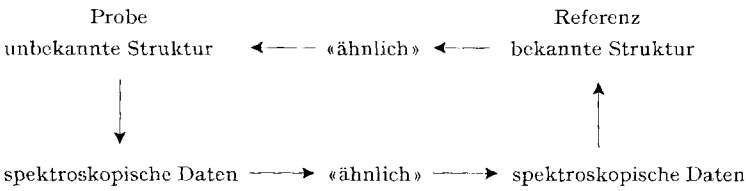
2. Problemstellung. – Das System soll aufgrund eines Vergleichs von Spektren für eine unbekannte Probe Strukturen vorschlagen, die der unbekannten Struktur ähnlich sind, was voraussetzt, dass

1. ähnliche Spektren von ähnlichen Strukturen stammen und
2. ähnliche Strukturen zu ähnlichen Spektren führen (s. Schema 1).

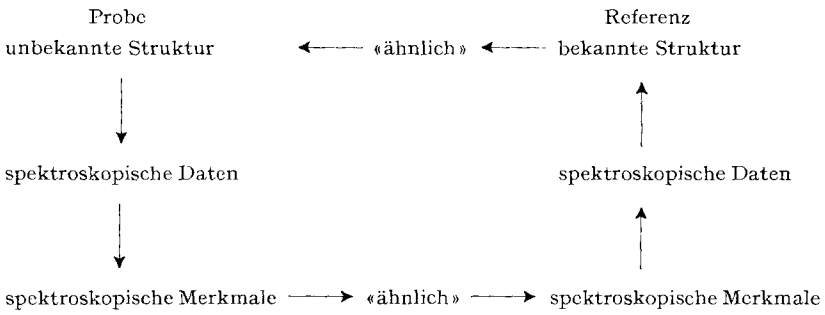
Ein Vergleich vollständiger Spektren ist kaum realisierbar, da Zeitaufwand und Speicherbedarf prohibitiv hoch werden [6]. Ausserdem gilt der in Schema 1 dargestellte Zusammenhang nicht generell, sondern nur für gewisse spektroskopische Merkmale. Dementsprechend muss dieses Schema in der im Schema 2 dargestellten Weise erweitert werden. Die wesentlichen Probleme sind dabei: sinnvolle Auswahl der spektroskopischen Merkmale, sinnvolle Definition des Begriffs «ähnlich», Anwendung effizienter Vergleichsverfahren.

Ein System mit brauchbaren Lösungen für diese Probleme soll im folgenden beschrieben werden.

Schema 1. *Algorithmus zur Gewinnung von Strukturvorschlägen durch Vergleich von Spektren*



Schema 2. *Erweiterter Algorithmus zur Gewinnung von Strukturvorschlägen durch Vergleich von spektroskopischen Merkmalen*



3. Beschreibung des Systems. – 3.1. *Grundprinzip.* Die spektralen Merkmale werden wie folgt dual verschlüsselt: Spektrales Merkmal vorhanden: Code = 1; spektrales Merkmal fehlt: Code = 0. Dieses Prinzip wird hier erstmals konsequent auf die Verschlüsselung mehrerer Spektrenarten angewendet. Lediglich für die Codierung von Massenspektren ist dieses Verfahren schon von anderen Autoren beschrie-

ben worden [8]. Ein einfaches Mass für die Ähnlichkeit zweier Spektren ist dann die Anzahl S der übereinstimmenden Merkmale. Damit lässt sich bereits ein primitives System konstruieren (Fig. 2).

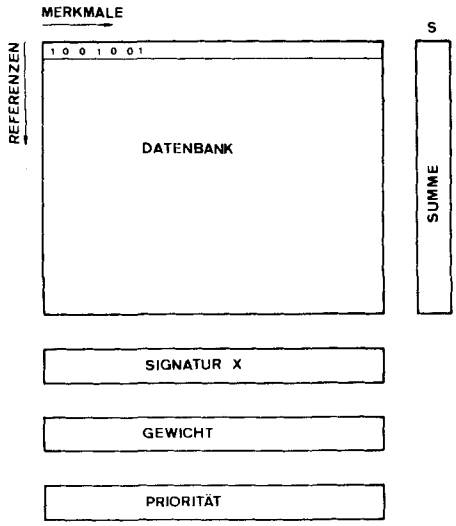


Fig. 2. *Schematische Darstellung des Vergleichssystems mit individuellen Prioritäten und Gewichtungen für Elementarvergleiche*

Die verschlüsselten Merkmale aller Referenzverbindungen werden in einer Datenbank am einfachsten in Form einer Matrix zusammengestellt; jede Zeile dieser Matrix entspricht einer Referenzverbindung und enthält die Codes aller ihrer spektralen Merkmale, d.h. ihre Signatur. Die Signatur einer unbekannten Probe X wird mit allen Signaturen der Datenbank verglichen, wobei jeweils die Anzahl der übereinstimmenden Merkmale in die Kolonne S (Fig. 2) eingetragen wird. Die Verbindungen mit den höchsten S-Werten werden als Referenzen ausgewählt. Bei geeigneter Auswahl der spektralen Merkmale sind ihre Strukturen der Struktur der unbekannten Probe X ähnlich.

Dieses primitive Verfahren kann wie folgt schrittweise verbessert werden.

3.2. *Gewichtung der Vergleichsresultate.* Der Vergleich zweier dual verschlüsselter Merkmale kann zu vier Resultaten mit verschiedener Bedeutung für den Grad der Ähnlichkeit führen:

A. Unbekannte und Referenz weisen das entsprechende Merkmal auf: «positive» Übereinstimmung.

B. Bei beiden Verbindungen fehlt das entsprechende Merkmal: «negative» Übereinstimmung.

C. Das bei der Referenz vorhandene Merkmal fehlt bei der Unbekannten.

D. Das bei der Unbekannten vorhandene Merkmal fehlt in der Referenz.

Diesen vier möglichen Resultaten werden die vier Grössen R1 bis R4 zugeordnet, die den Grad der Ähnlichkeit quantitativ beschreiben (Fig. 3). Diese hier erstmals vorge-

schlagene unterschiedliche Bewertung der möglichen Resultate eines Elementarvergleichs schafft erst die Voraussetzung für ein erfolgversprechendes Aufsuchen von ähnlichen Spektren. Andere bisher beschriebene Vergleichssysteme [3] [6] [7] [8] [9] dienen vorwiegend zum Auffinden von identischen Spektren. Sie zeigen dementsprechend nur dann ihre volle Leistungsfähigkeit, wenn das Spektrum der unbekannten Probe in der Datenbank vorhanden ist. Die Werte, die die Grössen R1 bis R4 für ein gegebenes Merkmal annehmen, sollen den Informationsgehalt des betreffenden Merkmals widerspiegeln. Dieser Informationsgehalt hängt sowohl von der spektralen Bedeutung des Merkmals wie auch von seiner statistischen Verteilung in der Datenbank ab. Die spektrale Bedeutung eines Merkmals muss heute noch weitgehend empirisch festgelegt werden. Eigene Versuche mit selbstlernenden Entscheidungsmaschinen auf der Basis von Entscheidungsvektoren [4] [8] oder Häufungssuchern (Cluster Analysis) haben vielversprechende Resultate gezeigt, über die später berichtet werden soll.

Die statistische Bedeutung eines Merkmals hängt vom Anteil p jener Referenzen in der Datenbank ab, die das betreffende Merkmal aufweisen. Der statistische Informationsgehalt I ist dann durch die folgende Gleichung [11] gegeben:

$$I = -p \cdot \log_{(2)}(p) - (1-p) \cdot \log_{(2)}(1-p).$$

Der statistische Informationsgehalt eines Merkmals hat demnach dann ein Maximum, wenn das betreffende Merkmal in der Hälfte aller Referenzverbindungen vorkommt.

Die Werte für R1 bis R4 werden nun aufgrund der empirisch festgelegten spektralen Bedeutung und des statistischen Informationsgehaltes für jedes einzelne Merkmal festgelegt und so transformiert, dass R4 gleich null wird.

Beim Vergleich der Signatur einer unbekannten Probe X mit den Signaturen der Referenzverbindungen werden die dem Resultat der Elementarvergleiche entsprechenden Gewichte R1 bis R4 (vgl. Fig. 3) zum Ähnlichkeitsmass S aufsummiert

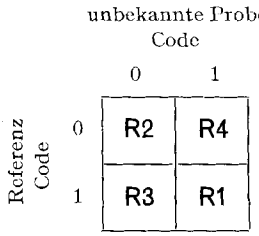


Fig. 3. *Bewertungsmatrix für elementare Vergleichsresultate*

(Fig. 2). Die Strukturen der Referenzverbindungen mit besonders hohen S-Werten sind bei geeigneter Wahl der Merkmale und ihrer spektroskopischen Gewichte der Struktur der unbekannten Probe ähnlich. Dieses durch Berücksichtigung der spektralen Bedeutung und des statistischen Informationsgehaltes verfeinerte Ähnlichkeitsmass wird hier erstmals verwendet. Es ist hervorragend geeignet, analoge und homologe Strukturen zu erkennen. Auch die Probleme, die sich aus der aufnahmetechnisch bedingten Variabilität der Spektren und aus eventuellen Verunreinigungen von Probe und/oder Referenz ergeben, lassen sich damit weitgehend lösen.

3.3. *Vergleichs-Strategie.* Das im Abschnitt 3.2 beschriebene System erfordert in jedem Fall die Durchsicht der gesamten Datenbank, was langwierig und damit teuer ist. Die Effizienz kann ohne Beeinträchtigung der Resultate enorm verbessert werden, wenn bei schlechter Übereinstimmung zweier Signaturen der Vergleichsprozess vorzeitig abgebrochen wird.

Für die Signatur einer unbekannten Probe X kann vor Beginn des Vergleichens aus den jeweiligen Werten von R1 und R2 (vgl. Fig. 3) und dem entsprechenden Signaturelement für jedes Merkmal der höchstmögliche Beitrag an die Summe S bestimmt werden. Diese Werte geben für jedes Merkmal an, wie hoch seine Bedeutung für den Vergleich der Signatur X ist. Damit lässt sich für die Merkmale eine nach abnehmender Bedeutung geordnete Prioritätssequenz P festlegen (vgl. Fig. 2). Werden nun die Elementarvergleiche in der durch die Prioritätssequenz gegebenen Reihenfolge durchgeführt, so steigt S bei voller Übereinstimmung der verglichenen Signaturen maximal an (Kurve a in Fig. 4). Nach jedem Elementarvergleich wird der aktuelle Stand der Summe S mit einem Schwellenwert (Kurve b in Fig. 4) verglichen.

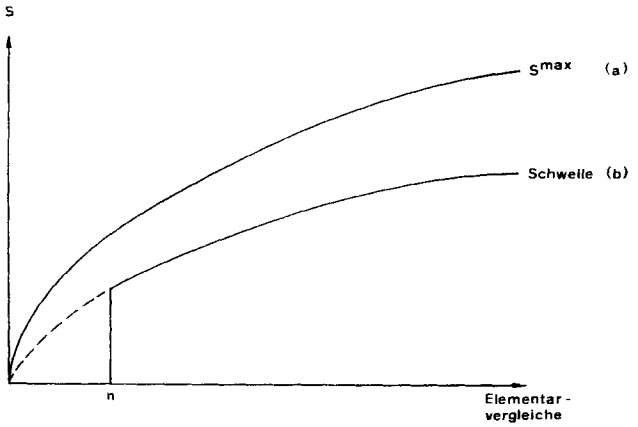


Fig. 4. *Schwelle als Abbruchkriterium*

Sinkt S unter den Schwellenwert, so wird der Vergleich als erfolglos abgebrochen. Damit werden eklatante Unterschiede schon sehr früh erkannt, so dass keine Rechenzeit an hoffnungslosen Fällen vergeudet wird. Um den ersten Elementarvergleichen keine allzugrosse Bedeutung zu verleihen, wird die Schwelle zweckmässig im Bereich der ersten n Vergleiche null gesetzt (Fig. 4). Die Erfahrung hat nun gezeigt, dass der überwiegende Teil aller Vergleiche bei n (Fig. 4) abbricht. Dementsprechend kann die Schwelle durch eine Hürde bei n ersetzt werden, was die Effizienz weiter verbessert (Fig. 5).

Da die Prioritätssequenz automatisch auf die Signatur der jeweiligen unbekannten Probe optimal abgestimmt wird, kann das Abbruchkriterium sehr streng gewählt werden. Es entspricht in seiner Wirkung einer Vorauslese (Filter), die aber im Gegensatz zu den bisher bekannten Verfahren (vgl. z.B. [9]) nach objektiven Kriterien vom System durchgeführt wird, und darum die unvermeidbare Voreingenommenheit des Benützers nicht berücksichtigen kann. Dies ermöglicht eine extrem hohe Effizienz ohne Beeinträchtigung des Resultates.

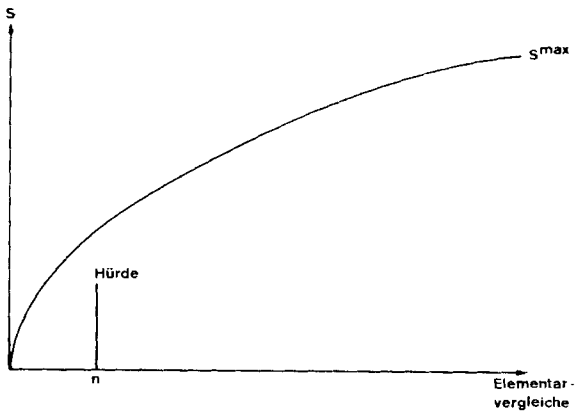


Fig. 5. *Hürde als Abbruchkriterium*

3.4. *Aufbau und Unterhalt der Datenbank.* Die für das Durchsuchen der Datenbank notwendige Rechenzeit wird in hohem Mass auch durch die Grösse der Datenbank bestimmt; diese ist im wesentlichen durch die Anzahl der Signaturen und durch deren Länge gegeben. Dies zwingt zu einer Limitierung der Anzahl der aufzunehmenden Verbindungen. Der Auswahl dieser Verbindungen muss daher besondere Beachtung geschenkt werden.

Die Zusammensetzung der Datenbank soll bezüglich der darin vertretenen Verbindungstypen dem Interessenbereich der Benützer möglichst getreu entsprechen. Die intensiv bearbeiteten Stoffklassen sollen mit einer grossen Anzahl auch spezieller Fälle vertreten sein. Von den seltener vorkommenden Stoffklassen genügen einige typische Beispiele.

Zur Beantwortung der Frage, ob die getroffene Auswahl den Anforderungen der Benützer entspricht, erstellt das System selbständig eine Statistik über die Häufigkeit, mit welcher die Verbindungen als Referenzen gewählt wurden. Selten oder nie gewählte Verbindungen sind entweder falsch codiert oder gehören zu seltenen Stoffklassen. Sie können bis auf einige typische Vertreter ohne Verminderung der Leistungsfähigkeit des Systems weggelassen werden. Sie sollen durch Vertreter jener Stoffklassen ersetzt werden, die überdurchschnittlich oft als Referenzen gewählt wurden.

Eine Statistik über den mittleren Rang, den ein Merkmal in der Prioritätssequenz beim Vergleichen erreicht, gibt Auskunft darüber, wie gut dieses Merkmal diskriminiert. Merkmale, die immer am Ende der Prioritätssequenz stehen, sollen durch Zusammenlegen oder Neudefinition verbessert werden.

Durch diese periodischen Korrekturen wird die Datenbank laufend den wechselnden Bedürfnissen der Benützer angepasst, ohne dass ihre Grösse und damit der Rechenaufwand ohne Grenzen ansteigen.

3.5. *Zusammenstellung.* Das beschriebene Vergleichssystem ist durch folgende Eigenschaften charakterisiert:

1. Durch die Beschränkung auf binär verschlüsselte Merkmale wird auf unnötige Redundanz weitgehend verzichtet, wodurch der Bedarf an Speicherplatz klein gehalten wird.

2. Eine sich dem jeweiligen Problem automatisch anpassende Vergleichsstrategie erlaubt möglichst frühe Erkennung schlechter Übereinstimmungen, so dass durch vorzeitigen Abbruch des Vergleichsprozesses enorm an Rechenzeit gespart wird.

3. Die Gewichtung der Elementarvergleiche ergibt ein sinnvolles und einfach zu errechnendes Ähnlichkeitsmass.

4. Die vom System selbständig erstellte Statistik der Resultate liefert die für eine kontinuierliche Anpassung der Datenbank an die Bedürfnisse notwendigen Angaben.

5. Der hohe Grad an Abstraktion durch binäre Codierung der Merkmale macht das System allgemein verwendbar.

4. Beschreibung des Datenmaterials. – Aus der Sammlung von Übungsbeispielen zur Vorlesung über Instrumentalanalyse am organisch-chemischen Laboratorium der ETH Zürich wurde eine kombinierte Datenbank aufgebaut, die gegenwärtig 953 Verbindungen umfasst. Für alle diese Verbindungen liegen Protonenresonanzspektrum, *Wiswesser*-Notation [12] und Summenformel vor; von 493 Verbindungen sind die IR.-Spektren vorhanden, von 428 Verbindungen auch die Massenspektren. Im ganzen sind 357 Verbindungen mit den drei genannten Spektren dokumentiert. Die Erweiterung der Sammlung durch UV.-Spektren und ^{13}C-Kernresonanzspektren ist im Gange.

Eine weitere, 9020 Massenspektren enthaltende Datenbank (MS.-Datenbank) wurde aus der Massenspektren-Sammlung des Mass Spectrometry Data Centre, Aldermaston, aufgebaut [13].

Für die erstgenannte Datenbank wurden die Spektren in folgender Form in Karten abgelocht:

Protonenresonanz-Spektren: Signalschwerpunkt (in ppm); Signalform (Singlett, Dublett, Triplett, Quartett, höheres Multiplett, breites Signal).

Infrarot-Spektren: Bandenmaximum (in cm^{-1}); Bandenintensität (schwach, mittel, stark, sehr stark, breit).

Massenspektren: Pik-lage in m/e-Einheiten; Pik-intensität bezogen auf Basispik = 100%.

Diese übersichtliche und leicht lesbare Codierung erlaubt es, die Fehlerrate klein zu halten und gestattet eine einfache Kontrolle der Rohdaten.

5. Beschreibung der Programme. – Alle hier erwähnten Programme sind in FORTRAN oder COBOL geschrieben und laufen auf dem CDC-6400/6500-System des Rechenzentrums der ETH Zürich. Auf eine zum mindesten teilweise Codierung in Assembler wurde – trotz dem damit verbundenen Verlust an Effizienz – vorerst verzichtet, um die Transparenz der Programme hoch zu halten und die Kompatibilität mit anderen Rechnersystemen zu gewährleisten.

Ein erstes Programm dient zur Erstellung der Datenbank. Die zu verwendenden spektralen Merkmale können vom Benützer spezifiziert werden. Aus den Rohdaten (codiert wie in 4 beschrieben) erzeugt das Programm selbständig die binären Signaturen und stellt sie zur Datenbank zusammen, die dann auf ein Magnetband geschrieben wird. Mit Hilfe dieses Programms kann die Datenbank jederzeit aus den Rohdaten mit beliebigen spektralen Merkmalen neu erstellt werden. Es können also die zum Aufbau der Signaturen verwendeten Merkmale mit geringem Aufwand (200 s Rechenzeit für 9000 Massenspektren) dem jeweils neuesten Stand der Erkenntnis angepasst werden.

Ein zweites Programm dient zum Aufsuchen von Referenzen in der Datenbank. Es können damit bis zu 20 unbekannte Verbindungen gleichzeitig bearbeitet werden. Deren Spektren werden wie die Rohdaten der Datenbank codiert. Das Programm erzeugt selbsttätig die binären Signaturen, analysiert sie und legt für jede der 20 Verbindungen individuell die beste Reihenfolge der Elementarvergleiche und das Abbruchkriterium fest. Hierauf werden die Signaturen der unbekannten Verbindungen mit jenen aller Referenzen verglichen. Für jede der unbekannten Verbindungen werden am Ende jene 20 Referenzverbindungen ausgegeben, deren Signaturen die grösste Ähnlichkeit (höchster Wert für S) aufweisen (vgl. Tab. 1).

Tabelle 1. *Computer-Output einer Recherche an der kombinierten Datenbank*
Gesuchte Verbindung: 1VR BO1

Name	Rel. S	Summenformel	*Wiswesser*-Notationen
VAR 234	86,30	C 9 H10 O 3	1VR DQ CO1
VAR 192	86,17	C 8 H 8 O 1	1VR
S 144	86,02	C 8 H 9 N 1 O 2	ZR DVO1
S 250	85,10	C 9 H 9 CL1 O 2	GVYOR
S 196	84,49	C10 H10 O 3	QV1U1R DO1 -T
S 91	84,45	C17 H20 N 2 O 1	1N1+R D- 2V
S 176	84,43	C 8 H 8 O 3	QR BVO1
S 189	84,35	C 8 H 7 F 1	FR C1U1 -V
VAR 230	84,28	C 9 H 8 O 2	QV1U1R -T
S 9	84,19	C14 H14 N 2	1R C DNUNR -T
VAR 500	84,15	C 8 H 7 CL1 O 3	QV1OR DG
S 88	84,04	C15 H10 O 2	T66 BO EVJ CR
VAR 628	83,97	C14 H12 O 5	T 0566 DO JV MOJ BO1 HO1 E
S 30	83,96	C11 H12 O 2	L66 BVT+J HO1
S 119	83,82	C 8 H 8 O 2	T5OJ B1U1V1 -U
S 70	83,78	C 8 H 6 BR2 O 1	E1VR DE
S 210	83,57	C 9 H10 O 3	QV1OR C
S 104	83,39	C15 H16 N 2 O 2	1OR DNUNR B DO1 -T
VAR 340	83,36	C20 H20 O 1	L66J B2R DQ C E
S 68	83,07	C 9 H10 O 3	VHR DQ CO2

Beide Programme sind modular aufgebaut, so dass sie einerseits beliebig erweiterbar sind und andererseits mit «Overlay»-Strukturierung in verhältnismässig kleinen Kernspeichern laufen können.

6. Resultate. – 6.1. *Einfluss der Schwellenhöhe* (*Abbruchkriterium*). Die Schwellenhöhe, die den vorzeitigen Abbruch der Vergleiche kontrolliert, soll so gewählt werden, dass möglichst viel Rechenzeit eingespart wird, ohne Qualitätsverminderung der aufgefundenen Referenzen. Nach unseren Erfahrungen erreicht man dies folgendermassen: Ein Vergleich wird dann als erfolglos abgebrochen, wenn nach n oder mehr Elementarvergleichen die Übereinstimmungssumme p% des jeweiligen Maximums unterschreitet. Für die Wahl des Parameters n (Anzahl der in jedem Fall durchgeführten Elementarvergleiche) ist die Gesamtzahl der Signaturelemente bestimmend; ein Wert in der Gegend von 10% der Signaturlänge hat sich gut bewährt. Der Einfluss des Parameters p (Schwellenhöhe in % des Maximalwerts) ist in Fig. 6 dargestellt, in der die Rechenzeit sowie Anzahl der wegen vorzeitigen Abbruchs des Vergleichsprozesses verlorengegangenen Referenzen gegen die Schwellenhöhe p aufgetragen sind. Diese Darstellung zeigt, dass bei einer Schwellenhöhe von etwa 60% die Rechenzeit schon beträchtlich abgenommen hat, ohne dass sich im Resultat Änderungen zeigen. Die Länge der Signaturen sowie der Umfang der Datenbank sind hier kaum von Bedeutung.

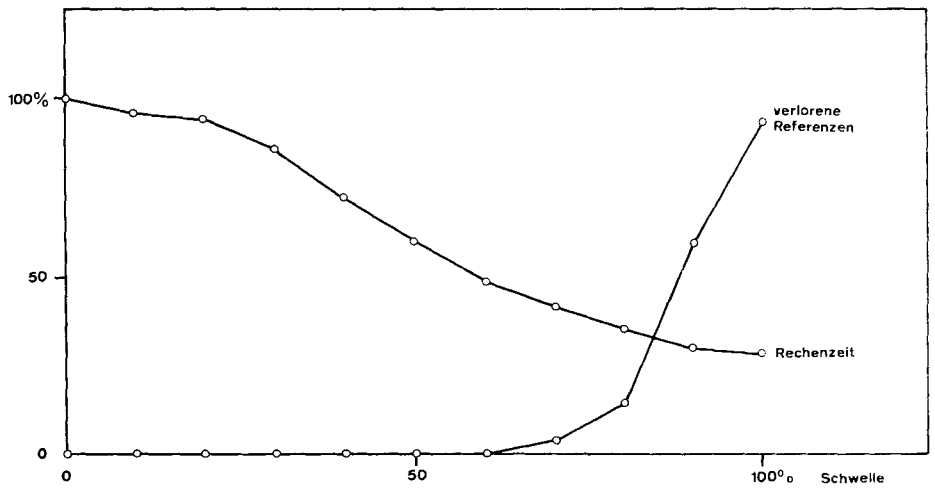


Fig. 6. *Einfluss der Schwellenhöhe auf Rechenzeit und Verluste an Referenzen*

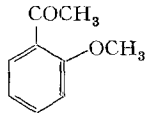
6.2. *Individuelle Anpassung der Vergleichsstrategie.* Die individuelle Gewichtung der spektroskopischen Merkmale und die daraus resultierende für jede Verbindung individuelle Abstimmung der Vergleichssequenz führt dazu, dass die informationsreichen Merkmale zuerst verglichen werden und einen höheren Beitrag zur Übereinstimmungssumme liefern. Dementsprechend werden bei einer Verbindung mit interessantem IR.-Spektrum und unspezifischem NMR.-Spektrum zuerst IR.-Merkmale verglichen, und der Gesamtbeitrag des IR.-Spektrums zur Übereinstimmungssumme S wird grösser sein als jener des NMR.-Spektrums und umgekehrt. Dieses Verhalten wird gut illustriert, z.B. durch Tetrachlorphthalsäure und 2-Bromacetaldehyd-diäthylacetal. In Fig. 7 sind die jeweiligen Anteile der IR.- bzw. NMR.-Spektren an der Übereinstimmungssumme S aufgetragen. Wie erwartet dominiert bei der Tetrachlorphthalsäure das IR.-Spektrum über das nur ein einziges unstrukturiertes Signal aufweisende NMR.-Spektrum, während beim 2-Bromacetaldehyd-diäthylacetal das wenig charakteristische IR.-Spektrum gegenüber dem sehr informativen Kernresonanzspektrum zurücktritt.

6.3. *Recherchen an einer kombinierten Datenbank.* Die spektralen Daten der oben beschriebenen Datenbank wurden in binäre Signaturen von 240 Bit Länge umgewandelt. Dabei lieferte das NMR.-Spektrum 48 Bit, das IR.-Spektrum 72 Bit und das Massenspektrum 120 Bit.

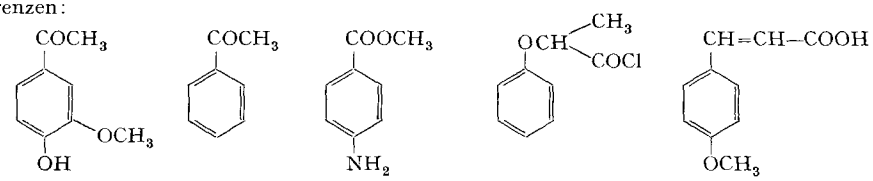
Zur Codierung der NMR.-Spektren wurde der Bereich der chemischen Verschiebung in 16 sich zum Teil überlappende Unterbereiche eingeteilt, denen je 3 Bit zugeordnet sind. Jeweils ein Bit gibt an, ob im betreffenden Bereich ein Signal auftritt; die beiden anderen Bits enthalten Angaben über die Signalform. Bei den IR.-Spektren wurden 24 Bereiche zu je drei Bit verwendet. Wiederum zeigt das erste Bit an, ob im betreffenden Bereich eine Bande auftritt, während in den anderen beiden Bits die Intensität codiert ist. Im Massenspektrum schliesslich wurden einmal die Intensitäten der *modulo* 14 reduzierten Spektren [5] in je drei Bit verschlüsselt. Ausserdem

wurde 78 ausgewählten *m/e*-Werten je ein Bit zugeordnet, das beim Auftreten eines Piks an der betreffenden Stelle gesetzt wurde.

Ein repräsentatives Resultat einer Recherche ist in folgenden Formeln zusammengestellt.

gesucht: 

Referenzen:



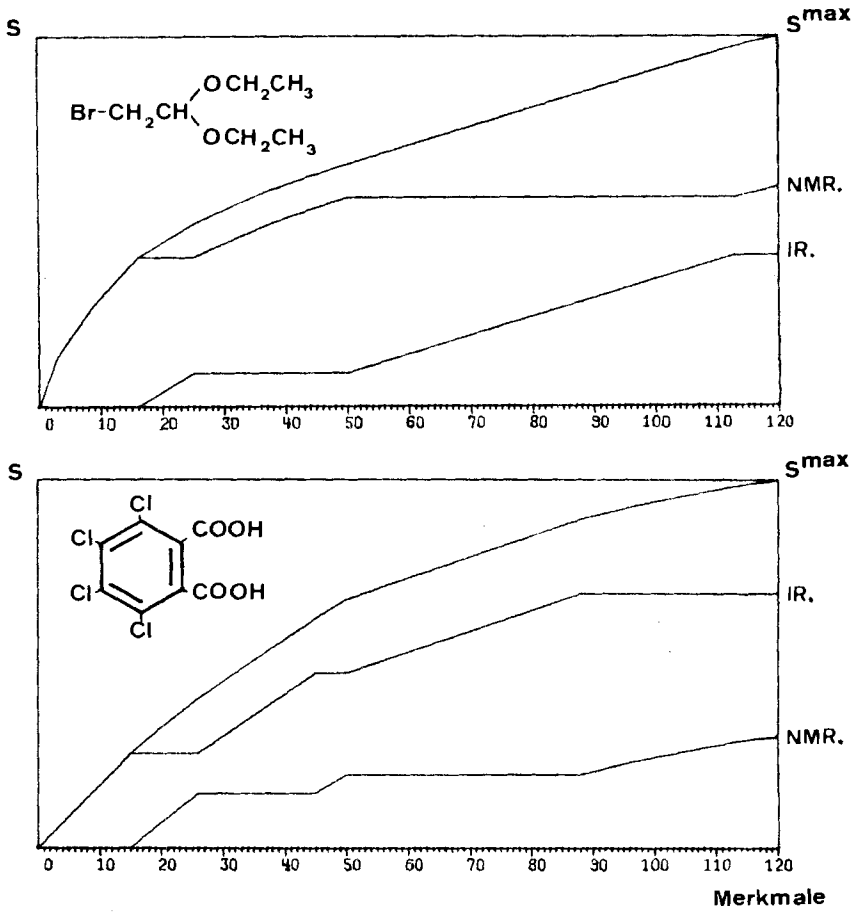


Fig. 7. *Einfluss der individuellen Gewichtung*
Anteil der IR.- bzw. NMR.-Spektren an der Übereinstimmungssumme *S*

In Tab. 1 ist der dazugehörige Computer-Output wiedergegeben. Bei der Würdigung der Resultate muss angemessen berücksichtigt werden, dass das System nur die etwa 900 in der kombinierten Datenbank vorhandenen Verbindungen als Referenzen vorschlagen kann und dementsprechend in seiner Auswahl recht beschränkt ist.

6.4. *Recherchen an einer MS.-Datenbank.* Die Signaturen entsprechen genau dem MS.-Teil der Signaturen aus der kombinierten Datenbank (120 Bit), hingegen enthält die Datenbank hier über 9000 Spektren. Resultate einer Recherche sind in Tab. 2 zusammengestellt.

Tabelle 2. *Resultate einer Recherche an einer MS.-Datenbank*
Gesuchte Verbindung: Isobutyl-isobutanoat

Referenzen: Übereinstimmung (in % von S_{max})	Summenformel	Trivialname
98,75	$C_8H_{16}O_2$	Isobutyl-isobutanoat
94,49	$C_8H_{16}O_2$	2-Butyl-butanoat
94,12	$C_8H_{16}O_2$	Butyl-butanoat
93,35	$C_8H_{16}O_2$	Butyl-butanoat
92,68	$C_8H_{16}O_2$	Butyl-isobutanoat
92,24	$C_8H_{16}O_2$	Isobutyl-isobutanoat
90,44	$C_{10}H_{20}O_2$	Hexyl-butanoat
90,43	$C_7H_{14}O_2$	Isopropyl-butanoat
90,10	$C_6H_{12}OS$	S-Äthyl-isobutanthioat
89,94	$C_7H_{14}O_2$	Propyl-butanoat
89,94	$C_7H_{14}O_2$	Propyl-butanoat
89,74	$C_{10}H_{20}O_2$	Hexyl-butanoat

Wie schon gesagt, können in einem Durchgang bis zu 20 unbekannte Verbindungen eingegeben werden. Auf dem ETH-Rechnersystem (CDC 6400/6500) benötigt der Vergleich von 100000 Massenspektren etwa 170 Sek. Die durchschnittliche Bearbeitungszeit von nur 1,7 Millisek. pro Spektrum zeigt am besten die enorme Leistungsfähigkeit des beschriebenen Vergleichsverfahrens.

7. Ausblick. – Das beschriebene Verfahren der computerunterstützten Interpretation von Spektren durch Vergleich mit Referenzspektren wird im organisch-chemischen Laboratorium der ETH Zürich seit kurzem eingesetzt und hat sich bis jetzt recht gut bewährt. Die noch zu lösenden Probleme betreffen heute noch vor allem die Auswahl der spektroskopischen Merkmale und die Festsetzung ausgewogener Gewichte. Das Problem der Behandlung nicht vollständig dokumentierter Verbindungen (unvollständige Signatur) scheint hingegen weitgehend gelöst.

Die gegenwärtig verwendeten Merkmale wie auch die ihnen zugeordneten Gewichte bedürfen noch weiterer Verbesserung. Dementsprechend ist auf eine detaillierte Dokumentation des gegenwärtigen Zustandes verzichtet worden. Auf Anfrage hin stellen wir jedoch Interessenten die Programme sowie die gegenwärtig aktuellen Merkmalkataloge und Gewichte gerne zur Verfügung.

Die vorliegende Arbeit wurde von *Schweizerischen Nationalfonds zur Förderung der wissenschaftlichen Forschung* unterstützt.

LITERATURVERZEICHNIS

[1] *W. Simon & J. T. Clerc*, Pure Appl. Chemistry *25*, 35 (1971).

[2] *W. W. Raznikov & W. L. Talroze*, Dokl. Akad. Nauk SSSR *170*, 379 (1966); *M. Barbes, P. Powers, M. J. Wallington & W. A. Wolstenholme*, Nature *212*, 784 (1966); *K. Biemann, C. Cone & B. R. Webster*, J. Amer. chem. Soc. *88*, 2597 (1966); *K. Biemann, C. Cone, B. R. Webster & G. P. Arsenault, ibid. 88*, 5598 (1966); *M. Senn, R. Venkataraghavan & F. W. McLafferty, ibid. 88*, 5593 (1966); *K. Biemann & P. V. Fennessey*, Chimia *21*, 226 (1967); *B. Petterson & R. Ryhage*, Analyt. Chemistry *39*, 790 (1967); *Shin-Ichi Sasaki, Hidetsugu Abe, Tatsumi Ouki, Masayoshi Sakamoto & Shukichi Ochiai, ibid. 40*, 2220 (1968); *J. Lederberg & E. A. Feigenbaum*, «Mechanization of Inductive Inference in Organic Chemistry» in «Formal Representation of Human Judgement», B. Kleinmutz Ed., John Wiley, New York 1968; *R. Venkataraghavan, F. W. McLafferty & G. E. Van Lear*, Organic Mass Spectrometry *2*, 1 (1969); *L. R. Crawford & J. D. Morrison*, Analyt. Chemistry *41*, 994 (1969); *J. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, V. A. Robertson, A. M. Duffield & C. Djerassi*, J. Amer. chem. Soc. *91*, 2977 (1969); *G. Schroll, A. M. Duffield, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum & J. Lederberg, ibid. 91*, 7440 (1969); *A. Buchs, A. B. Delfino, A. M. Duffield, C. Djerassi, B. G. Buchanan, E. A. Feigenbaum & J. Lederberg*, Helv. *53*, 1394 (1970); *B. Sheldrik*, Quart. Rev. *24*, 454 (1970).

[3] *B. Petterson & R. Ryhage*, Ark. Kemi *26*, 293 (1967).

[4] *P. C. Jurs, B. R. Kowalski, T. L. Isenhour & C. N. Reilley*, Analyt. Chemistry *41*, 695, 1949 (1969); 1387 (1970); *T. L. Isenhour & P. C. Jurs, ibid. 43*, 20 A (1971).

[5] *L. R. Crawford & J. D. Morrison*, Analyt. Chemistry *40*, 1469 (1968).

[6] *F. Erni & J. T. Clerc*, Chimia *24*, 388 (1970).

[7] *S. Abrahamsson, S. Ställberg-Stenhagen & E. Stenhagen*, Biochem. J. *92*, 2P (1964); *W. L. Talroze, W. W. Raznikov & G. D. Tantsyrev*, Dokl. Akad. Nauk SSSR *159*, 182 (1964); *E. Stenhagen*, Chimia *20*, 354 (1966); *D. H. Anderson & G. L. Covet*, Analyt. Chemistry *39*, 1288 (1967); *D. S. Erley, ibid. 40*, 895 (1968); *L. R. Crawford & J. D. Morrison, ibid. 40*, 1464 (1968); *R. A. Hites & K. Biemann* in «Advances in Mass Spectrometry», Vol. 4, S. 37, E. Kendrick Ed., The Institute of Petroleum, London 1968; *E. S. Schwartz*, J. chem. Doc. *9*, 39 (1969); *F. E. Lytle*, Analyt. Chemistry *42*, 355 (1970); *F. E. Lytle & T. L. Brazie*, Analyt. Chemistry *42*, 1532 (1970); *R. A. Hites & K. Biemann, ibid. 42*, 855 (1970); *R. S. Nigmatullin, W. I. Lobanov, I. K. Korobeinicheva, W. C. Botchlarev & W. A. Koptiug*, Vestnik Akad. Nauk SSSR *8*, 75 (1970); *D. S. Erley*, Appl. Spectros. *25*, 200 (1971); *P. C. Jurs*, Analyt. Chemistry *43*, 364 (1971); *H. S. Hertz, R. A. Hites & K. Biemann, ibid. 43*, 364 (1971).

[8] *S. L. Grotch, ibid. 42*, 1214 (1970); *L. E. Wangen, W. S. Woodward & T. L. Isenhour, ibid. 43*, 1605 (1971).

[9] *B. A. Knock, I. C. Smith, D. E. Wright & R. G. Ridley, ibid. 42*, 1516 (1970).

[10] *P. C. Jurs, B. R. Kowalski & T. L. Isenhour*, Analyt. Chemistry *41*, 21 (1969); *P. C. Jurs, B. R. Kowalski, T. L. Isenhour & C. N. Reilley, ibid. 41*, 690 (1969); *P. C. Jurs, ibid. 42*, 1633 (1970); *43*, 22 (1971).

[11] *E. Schultze*, «Einführung in die mathematischen Grundlagen der Informationstheorie», S. 4, Springer-Verlag, Berlin 1959.

[12] *E. G. Smith*, «The Wiswesser Line-Formula Notation», McGraw-Hill, New York, San Francisco, Toronto, London, Sydney 1968.

[13] UKAEA, AWRE, Mass Spectrometry Data Centre, Aldermaston, Reading RG7 4PR, United Kingdom.