

Klassifikationschromatogramme — eine vorbereitende Interpretation von GC/MS-Analysen mit Methoden der automatischen Mustererkennung

Von

Kurt Varmuza

Aus dem Institut für Allgemeine Chemie der Technischen Universität Wien,
Österreich

Mit 4 Abbildungen

(Eingegangen am 16. Juli 1975)

Classification Chromatograms—a Preliminary Interpretation of GC/MS-Analysis by Pattern Recognition Methods

Mass spectra series from GC/MS-analysis or from fractionated evaporation can be interpreted easily and quickly by spectra classifiers like binary, linear decision vectors ("learning machines"). Graphical representation of such classification results corresponds to a gas chromatogram registered with a detector sensitive only to a certain chemical structure. By analogy with mass chromatograms this graph is called classification chromatogram.

Einleitung

Die Kopplung Gaschromatograph — Massenspektrometer (GC/MS) ist eine sehr leistungsfähige qualitative Analysenmethode für komplexe Gemische organischer Substanzen¹. Die große Datenmenge, die bei einer GC/MS-Analyse anfällt, kann aber nur mit Hilfe einer elektronischen Datenverarbeitungsanlage zufriedenstellend bewältigt werden. Üblicherweise werden während der chromatographischen Trennung automatisch alle 2 bis 4 Sekunden Massenspektren registriert und auf Magnetband oder Magnetplatte gespeichert. Die anschließende Auswertung der erhaltenen Serie von bis zu einigen hundert Massenspektren ist aber trotz Computer mühsam. Angaben über die Identität oder über chemische Partialstrukturen der Komponenten des untersuchten Gemisches sind in den Massenspektren nur in stark verschlüsselter Form enthalten.

Die einfachste Auswertemethode besteht darin, daß man einzelne Massenspektren auf einem Plotter oder einem Bildschirmgerät zeichnen

läßt und versucht die Spektren zu interpretieren. Die meisten automatisierten Auswerteverfahren vergleichen mit Hilfe des Computers die unbekanntes Spektren mit den Spektren einer Bibliothek; die jeweils ähnlichsten Spektren werden in einer Tabelle ausgedruckt¹⁻⁴.

Da jedes unbekanntes Spektrum im allgemeinen mit mehreren Tausend Bibliotheksspektren verglichen werden muß, sind diese Verfahren sehr zeitaufwendig; es werden daher nur wenige Spektren der *GC/MS*-Analyse in einer stark reduzierten Form herangezogen. Nur wenn die Zahl der in Frage kommenden Substanzen klein ist, können alle unbekanntes Massenspektren mit allen Bibliotheksspektren verglichen werden. Für Vergiftungsanalysen wurde von *Biemann*^{5, 6} eine besonders übersichtliche Darstellung des Ergebnisses einer solchen erschöpfenden Bibliothekssuche angegeben: Die gefundenen Substanznamen werden entlang der Kontur des Chromatogrammes ausgedruckt. Alle Bibliothekssuchverfahren versagen jedoch weitgehend, wenn das Massenspektrum der betreffenden Substanz in der Bibliothek nicht vorhanden ist. Eines der Bibliotheksspektren ist zwar auch dann dem unbekanntes Spektrum am ähnlichsten, aber über die Art der chemischen Ähnlichkeit der Substanzen ist meist wenig bekannt. Zur Vorinterpretation von *GC/MS*-Analysen wurde weiters ein einfaches Vergleichsverfahren vorgeschlagen, das die Substanzklassen mit Hilfe von Ionenserienspektren (modulo 14 summierte Spektren) bestimmt⁷.

Eine andere weit verbreitete Auswertemethode von Massenspektrenserien wird durch Darstellung der Meßresultate in Form von Massenchromatogrammen (auch Massenfragmentogramme) durchgeführt^{1, 8, 9}. In derartigen Diagrammen wird die Peakhöhe einer bestimmten Massenzahl in Abhängigkeit von der Spektrennummer dargestellt. Massenchromatogramme sind insbesondere wertvoll für die Erkennung nichtaufgetrennter *GC*-Peaks und für den Nachweis von Spurenkomponenten. Voraussetzung für diese Auswertemethode sind aber charakteristische Massenzahlen der Komponenten und eine gewisse Vorkenntnis über die auftretenden Substanzgruppen.

In der vorliegenden Abhandlung wird eine neue Methode zur Vorinterpretation von Massenspektrenserien vorgeschlagen, die eine einfache Methode der automatischen Mustererkennung (pattern recognition) benützt. Mit Hilfe von Klassifikatoren werden sämtliche Spektren einer *GC/MS*-Analyse hinsichtlich einer bestimmten chemischen Partialstruktur klassifiziert.

Die graphische Darstellung des Ergebnisses entspricht einem Chromatogramm, das mit einem Detektor registriert wurde, der nur für die betreffende Partialstruktur empfindlich ist.

Automatische Klassifikation von Massenspektren

Die Interpretation von Massenspektren mit verschiedenen Methoden der automatischen Mustererkennung wurde bereits in mehreren Übersichtsreferaten beschrieben¹⁰⁻¹².

Ein Klassifikator ist ganz allgemein eine Rechenvorschrift, nach der beispielsweise die Daten eines Massenspektrums (Peakhöhen) mathematisch verknüpft werden — mit dem Ziel, daß das Ergebnis

dieser Rechnung (Klassifikationsergebnis) eine Antwort auf eine bestimmte chemische Fragestellung erlaubt. Ein binärer Klassifikator entscheidet zum Beispiel auf Grund des Massenspektrums über das Vorhandensein einer bestimmten Partialstruktur. Oft erlaubt das Klassifikationsergebnis auch eine Angabe der Sicherheit der Klassifikation. In der vorliegenden Arbeit wurden binäre, lineare, adaptive Entscheidungsvektoren („learning machines“) als Klassifikatoren verwendet, die in früheren Arbeiten berechnet und getestet wurden¹³⁻¹⁵. Ein derartiger Klassifikator besteht aus einer Tabelle, die für jede Massenzahl einen Faktor (positiv oder negativ) enthält. Zur Klassifikation eines Spektrums wird jede Peakhöhe mit dem entsprechenden Faktor multipliziert und die Produkte summiert. Das Vorzeichen dieses Klassifikationsergebnisses S (es ist das Skalarprodukt aus Entscheidungsvektor und Massenspektrenvektor) zeigt an, ob in der Substanz eine bestimmte Partialstruktur vorhanden ist ($S > 0$) oder nicht ($S < 0$). Vorteile dieser Methode sind der geringe Speicherbedarf im Computer und die kurze Rechenzeit; nachteilig ist eine gewisse Unsicherheit der Klassifikation (üblicherweise liegt die „Erfolgsquote“¹⁶ zwischen 80 und 95%) und die ja/nein-Entscheidung über jeweils nur eine Partialstruktur.

Klassifikationschromatogramme

Infolge der einfachen und raschen Anwendbarkeit von Spektrenklassifikatoren ist es ohne großen Aufwand möglich, jeweils alle Spektren einer *GC/MS*-Analyse zu klassifizieren. Wird das Klassifikationsergebnis in Abhängigkeit von der Spektrennummer (Zeit) graphisch dargestellt, so kann man durch Vergleich mit dem Chromatogramm den einzelnen *GC*-Peaks bestimmte chemische Partialstrukturen oder Substanzklassen zuordnen.

Im einfachsten Fall wird in der Ordinatenrichtung die binäre Entscheidung aufgetragen (z. B. + 1 für „Partialstruktur vorhanden“ und - 1 für „Partialstruktur nicht vorhanden“). Ist der absolute Zahlenwert des Klassifikationsergebnisses von Bedeutung (z. B. als Maß für die Verlässlichkeit der Klassifikation¹⁷), kann er in der Ordinate aufgetragen werden. Es kann aber in der graphischen Darstellung auch berücksichtigt werden, daß jene Massenspektren, die mit geringer Intensität registriert wurden, weniger Aussagekraft haben. Im folgenden Beispiel wurde daher für jedes Spektrum das Klassifikationsergebnis S mit dem Total-ionenstrom des Spektrums multipliziert und dieses Produkt als Ordinate verwendet (als Total-ionenstrom wurde die Summe der Peakhöhen verwendet). Diese graphische Darstellung ähnelt einem Gaschromatogramm und wird in Analogie zum Massenchromatogramm als *Klassifikationschromatogramm* bezeichnet. Positive Werte im Klassifi-

kationschromatogramm bedeuten, daß die betreffende Partialstruktur vorhanden ist, negative Werte zeigen die Abwesenheit an.

Die Vorinterpretation von Massenspektrenserien mit Hilfe von Klassifikationschromatogrammen wird an Hand einer „künstlichen

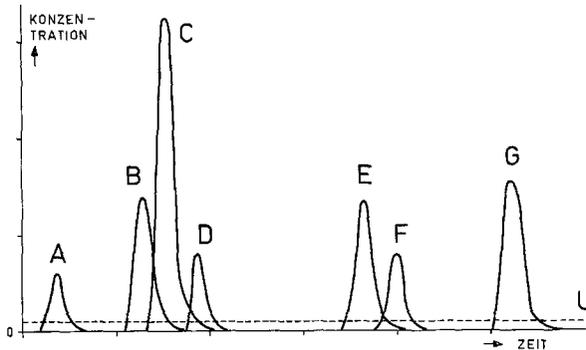


Abb. 1. Künstliches Gaschromatogramm eines Gemisches aus 7 Komponenten. Zeit und Konzentration in willkürlichen, linearen Einheiten. *A* Isopropylbenzol, *B* n-Butylbenzol, *C* Cycloheptanon, *D* Methylbenzoat, *E* Chinolin, *F* 1-Methylnaphthalin, *G* 3-Methylindol, *U* Untergrund (Luft)

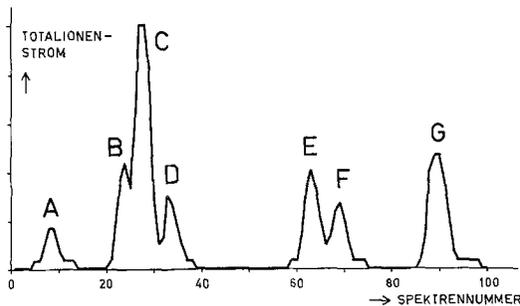


Abb. 2. Rekonstruiertes Gaschromatogramm aus 100 Massenspektren einer *GC/MS*-Analyse nach Abb. 1. Als Ordinate ist der Total-ionenstrom (Summe aller Peakhöhen im Spektrum) in willkürlichen, linearen Einheiten aufgetragen

GC/MS-Analyse“ gezeigt (ein *GC/MS*-System stand nicht zur Verfügung). Die angenommenen Komponenten und Konzentrationen im Gaschromatogramm sind in Abb. 1 angegeben. Mit Absicht wurde eine schlechte Trennung der Komponenten vorausgesetzt. Für die *GC/MS*-Analyse wurde angenommen, daß in gleichen Zeitabständen insgesamt 100 Massenspektren registriert wurden. Diese Massenspektren wurden aus den Spektren der Reinsubstanzen¹⁸ und den Konzentrationsangaben aus Abb. 1 errechnet. Zusätzlich wurde ein konstantes Untergrundspek-

trum (Luft) addiert. Abb. 2 zeigt das aus dem Total-ionenstrom rekonstruierte Gaschromatogramm*.

Jedes der 100 Massenspektren wurde mit einem Entscheidungsvektor klassifiziert, der angibt, ob in der Substanz ein (beliebig substituierter) Benzolring vorhanden ist oder nicht. Das Klassifikationschromatogramm in Abb. 3 zeigt auf einen Blick, daß die Komponente *C* keinen Benzolring enthält, alle übrigen Komponenten enthalten diese Partial-

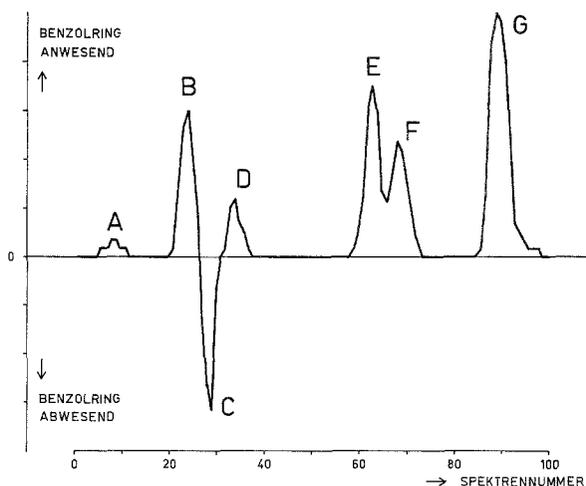


Abb. 3. Klassifikationschromatogramm für „Anwesenheit eines beliebig substituierten Benzolringes“. Ordinate: Klassifikationsergebnis \times Totalionenstrom in beliebigen, linearen Einheiten. Die negativen Werte für Komponente *C* zeigen die Abwesenheit eines Benzolringes an; für alle übrigen Komponenten wurde ein Benzolring gefunden. Alle Komponenten (siehe Abb. 1) wurden richtig klassifiziert

struktur. Die richtige Klassifikation aller Spektren wurde weder durch das Untergrundspektrum, noch durch die teilweise große Überlappung der *GC*-Peaks gestört. Abb. 4 zeigt ein Klassifikationschromatogramm für „Anwesenheit eines oder mehrerer Sauerstoffatome im Molekül“. Infolge der schlechten Trennung der Komponenten *B* und *C* und der Störung durch das Untergrundspektrum ist in diesem Fall die Klassifikation der Komponente *B* unsicher. Ähnlich gute Ergebnisse wurden mit Spektren erzielt, deren Peakhöhen statistisch verrauscht waren (Standardabweichung bis 30%). Aufnahmebedingte und statistische Unregelmäßigkeiten werden in Klassifikationschromatogrammen ausgeglichen, weil für jede *GC*-Komponente mehrere Massenspektren klassi-

* Dieses Chromatogramm ähnelt einer von *Biemann*² angegebenen *GC/MS*-Analyse, mit der Spektrenvergleichsverfahren getestet wurden.

fiziert werden. Darüber hinaus könnte die Sicherheit der Entscheidung durch Anwendung verschiedener Klassifikatoren für die gleiche Problemstellung geprüft werden.

Zusammenfassung

Massenspektrenserien, die bei *GC/MS*-Analysen oder bei fraktionierten Verdampfungen im Direkteinlaßsystem des Massenspektro-

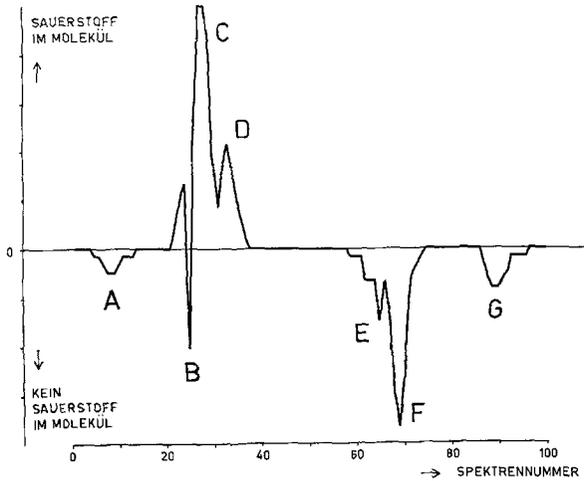


Abb. 4. Klassifikationschromatogramm für „Anwesenheit eines oder mehrerer Sauerstoffatome im Molekül“. Ordinate: Klassifikationsergebnis \times Total-ionenstrom in beliebigen, linearen Einheiten. Alle Komponenten (siehe Abb. 1) wurden richtig klassifiziert; das Ergebnis für *B* ist jedoch unsicher

meters mit Hilfe eines Datensystems aufgenommen wurden, können ohne großen Aufwand mit Spektrenklassifikatoren vorinterpretiert werden. Die graphische Darstellung des Ergebnisses wird als Klassifikationschromatogramm (bzw. Klassifikationsfragmentogramm) bezeichnet und entspricht einem Chromatogramm, das mit einem Detektor registriert wurde, der nur für eine bestimmte chemische Partialstruktur empfindlich ist. Der gegenüber einem üblichen Massenchromatogramm vermehrte Rechenaufwand fällt nicht ins Gewicht. Die Klassifikationen erfolgen so rasch, daß sie auch in Echtzeit zwischen zwei registrierten Spektren durchführbar sind (wie es für eine andere Klassifikationsmethode bereits vorgeschlagen wurde¹⁹). Durch die Klassifizierung einer möglichst großen Zahl von Massenspektren einer *GC/MS*-Analyse werden statistische oder methodische Unsicherheiten beseitigt oder zumindest erkannt. Die erforderlichen Klassifikatoren können nach

bekannten Methoden der automatischen Mustererkennung im Prinzip für alle jene chemischen Fragestellungen entwickelt werden, die durch Massenspektren lösbar sind.

Daten und Programme

Die als Klassifikatoren verwendeten Entscheidungsvektoren wurden mit Hilfe von 500 niedrig aufgelösten Massenspektren¹⁸ berechnet¹⁴; die Spektren stammen von Substanzen mit Molekulargewichten zwischen 16 und 150 und Summenformeln $C_{1-11}H_{1-22}O_{0-2}N_{0-2}$. Die Massenspektren aller sieben reinen *GC*-Komponenten wurden richtig klassifiziert. Alle Programme wurden in Fortran geschrieben und liefen auf einer CDC Cyber 74 im Rechenzentrum der Technischen Universität Wien. Die Rechenzeit für 100 Massenspektren und 4 Klassifikatoren betrug etwa 3 Sekunden — ohne Optimierung des Programmes oder der Datenspeicherung. Die Ergebnisse wurden als Tabelle und als graphische Darstellung am Schnelldrucker ausgegeben. Abb. 2, 3 und 4 entstanden durch einfaches Umzeichnen der Computerausdrucke. Eine Implementierung des Programmes auf einem in der Massenspektrometrie üblichen Kleincomputer erscheint problemlos.

Dank

Herrn Prof. Dr. *A. Maschka* danke ich für die freundliche Unterstützung dieser Arbeit. Herrn *H. Urban* danke ich für seine tatkräftige Mithilfe.

Literatur

- ¹ *W. H. McFadden*, Techniques of Combined Gas Chromatography/Mass Spectrometry. New York: Wiley. 1973.
- ² *H. S. Hertz, R. A. Hites* und *K. Biemann*, Anal. Chem. **43**, 681 (1971).
- ³ *R. Reimendal* und *J. B. Sjövall*, Anal. Chem. **45**, 1083 (1973).
- ⁴ *E. Jellum, O. Stokke* und *L. Eldjarn*, Anal. Chem. **45**, 1099 (1973).
- ⁵ *K. Biemann*, in: The Applications of Computer Techniques in Chemical Research (Hrsg. *P. Hepple*), S. 18. London: Institute of Petroleum. 1972.
- ⁶ *C. E. Costello, T. Sakai* und *K. Biemann*, in Lit. ¹, S. 377.
- ⁷ *D. H. Smith* und *G. Eglinton*, Nature **235**, 325 (1972).
- ⁸ *R. A. Hites* und *K. Biemann*, Anal. Chem. **42**, 855 (1970).
- ⁹ *D. Henneberg, B. Weimann* und *E. Ziegler*, Chromatographia **7**, 483 (1974).
- ¹⁰ *P. C. Jurs* und *T. L. Isenhour*, Chemical Applications of Pattern Recognition. New York: Wiley. 1975.
- ¹¹ *T. L. Isenhour* und *P. C. Jurs*, Anal. Chem. **43**, 20 A (1972).
- ¹² *P. Krenmayr* und *K. Varmuza*, Allgem. Prakt. Chem. **23**, 289 (1972).
- ¹³ *K. Varmuza* und *P. Krenmayr*, Z. Anal. Chem. **266**, 274 (1973).
- ¹⁴ *K. Varmuza* und *P. Krenmayr*, Z. Anal. Chem. **271**, 22 (1974).
- ¹⁵ *K. Varmuza*, Mh. Chem. **105**, 1 (1974).

- ¹⁶ *H. Rotter* und *K. Varmuza*, *Org. Mass Spectrom.* **10**, 874 (1975).
¹⁷ *L. E. Wangen*, *N. M. Frew* und *T. L. Isenhour*, *Anal. Chem.* **43**, 845 (1971).
¹⁸ *E. Stenhagen*, *S. Abrahamsson* und *F. W. McLafferty*, *Atlas of Mass Spectral Data*. New York: Interscience. 1969.
¹⁹ *J. Franzen*, *Chromatographia* **7**, 518 (1974).

Korrespondenz und Sonderdrucke:

Dr. K. Varmuza
Institut für Allgemeine Chemie
Technische Universität Wien
Lehár-gasse 4
A-1060 Wien
Österreich