communications in mathematical
match
and in computer chemistry

no. 37, March 1998

ISSN 0340-6253

MATCDY (37) 163-177 (1998)

BESTIMMUNG VON SUMMENFORMELN AUS MASSENSPEKTREN DURCH ERKENNUNG ÜBERLAGERTER ISOTOPENMUSTER

T. Grüner[1], A. Kerber, R. Laue, M. Liepelt[1], M. Meringer[1]

Universität Bayreuth, Mathematisches Institut,
D-95440 Bayreuth

K. Varmuza, W. Werther

Technische Universität Wien, Abteilung Chemometrie,
Getreidemarkt 9/152, A-1060 Wien

Zusammenfassung

In der vorliegenden Arbeit wird eine Methode zur automatischen Erkennung von Isotopenpeakmustern in niedrig aufgelösten EI-Massenspektren beschrieben. Insbesondere wird auf Überlagerung von Peakmustern verschiedener Fragmente eingegangen. Eine Verallgemeinerung auf das gesamte Spektrum wird vorgestellt und anhand eines Beispiels wird gezeigt, wie Kandidaten für die Summenformel der zu bestimmenden Reinsubstanz aufgrund der Erklärbarkeit aller Peaks des Spektrums bewertet werden. Schließlich wird beschrieben, wie die so gewonnenen Vergleichswerte herangezogen werden, um bei vorgegebener Zuverlässigkeit relevante Kandidaten auszuwählen. Die Anwendbarkeit der Methode wird anhand eines Testsatzes von über 5000 Massenspektren organischer Verbindungen demonstriert.

[1]unterstützt von der Deutschen Forschungsgemeinschaft

Element	Masse	%	Masse	%	Masse	%
H	1	100				
C	12	100	13	1.1		
N	14	100	15	0.37		
O	16	100	17	0.04	18	0.2
F	19	100				
Si	28	100	29	5.1	30	3.4
P	31	100				
S	32	100	33	0.79	34	4.4
Cl	35	100			37	32.0
Br	79	100			81	97.3
I	127	100				

Tabelle 1: Isotopenverteilungen der häufigsten Elemente in der organischen Chemie.

1 Einführung

1.1 Theoretische Isotopenmuster chemischer Verbindungen

Bei den meisten natürlich vorkommenden Elementen gibt es Atome mit unterschiedlicher Atommasse, sogenannte Isotope. Tabelle 1 (vgl. [3]) gibt eine Übersicht wichtiger Isotope der am häufigsten in der organischen Chemie auftretenden Elemente. Basierend auf der bekannten natürlichen Isotopenverteilung der chemischen Elemente kann zu jeder Elementkombination ihr theoretisches Isotopenmuster exakt berechnet werden. Zunächst betrachten wir zwei beliebige Elemente X und Y mit Häufigkeit $\mathrm{I}_X(m)$, bzw. $\mathrm{I}_Y(m)$ für ein Isotop der Masse m. Als nominale Masse eines Elements bezeichnen wir die ganzzahlige Masse des häufigsten Isotops. Bei der Verbindung XY treten Teilchen der Masse m mit Häufigkeit

$$\mathrm{I}_{XY}(m) = \sum_{(i,j):i+j=m} \mathrm{I}_X(i) \cdot \mathrm{I}_Y(j)$$

auf. Diese Art der Verknüpfung von I_X und I_Y heißt Faltung. Das theoretische Isotopenmuster I_{XY} von XY erhält man also durch Faltung von I_X und I_Y. Die theoretischen Isotopenmuster von Verbindungen mit mehreren Atomen lassen sich durch Komposition von Faltungen bekannter Isotopenmuster berechnen.
Im folgenden schreiben wir ein Isotopenmuster als Tripel $\mathcal{P} = (l, h, v)$, wobei l die kleinste, h die größte Masse und v einen Vektor der Länge $h-l+1$ bezeichnet, der die Intensitäten zwischen l und h angibt: $v = (I(l), ..., I(h))$. Weiterhin vereinbaren wir die Konvention, Intensitäten so zu normieren, daß die höchste den Wert 100 erhält. Für CH_4O ist $\mathcal{P} = (32, 34, (100, 1.14, 0.20))$ das theoretische Isotopenmuster (vgl. Abb. 1). Weitere Beispiele für theoretische Isotopenmuster findet man in Tabelle 3.

1.2 Überlagerte Isotopenmuster

Die im Massenspektrometer entstandenen Ionen hinterlassen im Massenspektrum nur Informationen über ihre Massen, die Isotopenzusammensetzung und die relative Häufigkeit. Unterscheiden sich die auftretenden Ionen jeweils um mehrere Masseneinheiten, so könnte allein durch Vergleich möglicher theoretischer Isotopenmuster mit dem gemessenen Spektrum die elementare Zusammensetzung des Analyten bestimmt werden. In der Praxis kommt es aber fast immer zu einer Überlagerung mehrerer verschiedener Isotopenmuster.
Wir nennen zwei Isotopenmuster $\mathcal{P}_1 = (l_1, h_1, v_1)$ und $\mathcal{P}_2 = (l_2, h_2, v_2)$ überlagert, wenn es mindestens eine Masse gibt, bei der sowohl $\mathcal{P}_1$ als auch $\mathcal{P}_2$ Intensitäten aufweisen, d.h.

$$\exists m : l_1 \leq m \leq h_1 \wedge l_2 \leq m \leq h_2 \wedge v_{1,m-l_1+1} > 0 \wedge v_{2,m-l_2+1} > 0.$$

Treten im Massenspektum zwei Ionen F_1 und F_2 mit Häufigkeiten f_1 bzw. f_2 auf und überlagern sich ihre Isotopenmuster, so erhält man für die resultierenden relativen Intensitäten bei m/z= $m \in \{l_1, ..., h_1\} \cap \{l_2, ..., h_2\}$

$$I(m) = f_1 \cdot I_{F_1}(m) + f_2 \cdot I_{F_2}(m),$$

wobei $I_{F_1}(m)$ und $I_{F_2}(m)$ die theoretischen Intensitäten von F_1 und F_2 bei m bezeichnen. Überlagern sich mehr als zwei Isotopenmuster, müssen entsprechend mehr Summanden berücksichtigt werden. Abbildung 1 zeigt ein Massenspektrum von Methanol (CH_4O). Auffällig dabei ist, daß der höchste Peak nicht bei m/z=32 liegt, was aufgrund der Masse des Molekülions M^+ zu erwarten wäre. Tatsächlich kommt es hier zur Überlagerung der Isotopenmuster von M^+, $[M-H]^+$, $[M-2H]^+$, $[M-3H]^+$ und $[M-4H]^+$. Dabei tritt $[M-H]^+$ häufiger auf als M^+ und erklärt den Basispeak bei m/z=31.
Da die relativen Häufigkeiten der verschiedenen Ionen unbekannt sind, können die theoretischen Intensitäten der so entstandenen Peakgruppe nicht ad hoc berechnet werden. Allerdings gibt es eine Möglichkeit zu entscheiden, wie gut die gemessene Peakgruppe durch einen vorgegebenen Satz theoretischer Isotopenmuster von Fragmenten dargestellt werden kann. Diese Methode wird in Abschnitt 2 vorgestellt und bedeutet eine wesentliche Verbesserung im Vergleich zu früheren Ansätzen zu dieser Problematik.

1.3 Peakgruppen

Zur Interpretation eines Massenspektrums ist es sinnvoll, das Spektrum in Gruppen von Peaks zu unterteilen, die jeweils verschiedenen Fragmenten zugeordnet werden können. Als Peakgruppe bezeichnen wir Folgen von Peaks, die sich um eine oder zwei Einheiten von m/z unterscheiden. Verschiedene Peakgruppen unterscheiden sich um mindestens drei Einheiten. Wir schreiben sie ebenfalls als Tripel $\mathcal{P} = (l, h, v)$. Die Einträge entsprechen denen der Isotopenmuster: l ist die niedrigste, h die höchste Masse der Peakgruppe und v ein Vektor der Länge $h-l+1$ mit $v_j = I_S(j+l-1)$, $j = 1, ..., h-l+1$. $I_S(m)$

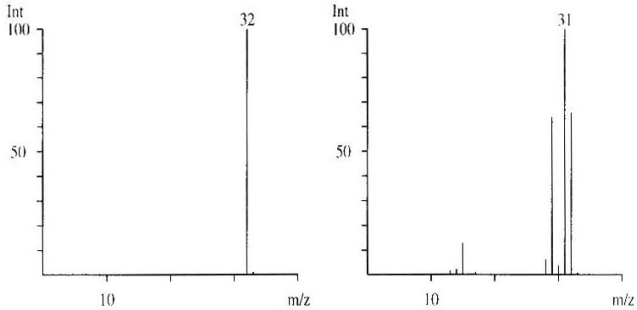

Abbildung 1: Theoretisches Isotopenmuster von CH_4O (links) und Massenspektrum von Methanol (rechts).

bezeichnet die Intensität bei m/z=m im Massenspektrum. Demnach ist $\mathcal{P}$ Peakgruppe genau dann, wenn

$$I_S(l) \neq 0 \wedge I_S(h) \neq 0 \wedge I_S(l-2) = I_S(l-1) = I_S(h+1) = I_S(h+2) = 0$$
$$\wedge\ \forall m \in \{l, ..., h\} \text{ mit } I(m) = 0 : I_S(m-1) \neq 0 \wedge I_S(m+1) = 0$$

Demnach besteht das Spektrum von Methanol (Abb. 1) aus zwei Peakgruppen: $(12, 17, (0.3, 1.7, 2.4, 13, 0.2, 1.0))$ und $(28, 34, (6.3, 64, 3.8, 100, 66, 1.0, 0.1))$.
Für die praktische Arbeit am Massenspektrum muß diese Definition weiter gefaßt werden, da sonst eine Separierung des Spektrums in Peakgruppen oft nicht erreicht werden kann. Vielfach bereitet gerade bei der automatischen Interpretation von Massenspektren schon die sinnvolle Einteilung in Peakgruppen große Probleme. Das diesem Artikel zugrundeliegende Computerprogramm ist dazu mit einer Routine versehen, die ermöglicht, auch Spektren mit „Rauschen" zu verarbeiten. An dieser Stelle soll aber nicht auf Details der Implementation eingegangen werden. Für unsere theoretischen Betrachtungen ist die obige Definition gut geeignet.
Von besonderem Interesse bei der Interpretation von Massenspektren ist die zum Molekülion gehörende Peakgruppe. Die Fragmente, deren Isotopenmuster zu dieser Peakgruppe beitragen, können nur durch Abspaltung eines oder mehrerer Wasserstoffatome vom Molekülion entstanden sein. Abschnitt 3 behandelt diesen Spezialfall. Sehr selten in der Elektronenstoß-Massenspektrometrie gibt es Spektren, bei denen $[M+H]^+$ Ionen zur Peakgruppe des Molekülions beitragen. Dieser Effekt wird bei den vorliegenden Betrachtungen nicht berücksichtigt. Der Leser wird leicht feststellen, daß das Modell problemlos erweitert werden kann, um auch diesem Phänomen gerecht zu werden.
Abschnitt 4 stellt einen Algorithmus vor, der zu gegebenem Massenspektrum und Summenformel Kandidaten auf kanonische Weise einen Vergleichswert ermittelt, der die

Erklärbarkeit aller Peaks des Spektrums widerspiegelt. Die Kandidaten werden dann nach aufsteigenden Vergleichswerten sortiert in eine Hitliste eingefügt. Besonders interessant ist dieser Ansatz deshalb, weil zwei der großen Probleme bei der MS Interpretation gelöst, bzw umgangen werden:

- alle möglichen Überlagerungen von Isotopenmustern werden berücksichtigt und
- prinzipiell bedarf es keiner Einteilung des Spektrums in Peakgruppen.

2 Überlagerung der Isotopenmuster von n Fragmenten

Wir betrachten eine gemessene Peakgruppe $\mathcal{P}_0 = (l_0, h_0, v_0)$, sowie theoretische Isotopenmuster $\mathcal{P}_i = (l_i, h_i, v_i)$, $i = 1, ..., n$ von n Fragmenten, die zu der Peakgruppe beitragen können. Sei $l := \min(l_0, ..., l_n)$, $h := \max(h_0, ..., h_n)$ und $p = h - l + 1$. Dann liegen sowohl die Peaks der theoretischen Isotopenmuster als auch die der gemessenen Peakgruppe bei m/z-Werten von l bis $l+p-1$. Weiterhin erklären wir Vektoren $c, d_i \in I\!R^p$ für $i = 1, ..., n$ durch

$$
\begin{aligned}
c_j &= \begin{cases} v_{0,l-l_0+j}, & \text{falls } l_0 - l < j \leq h_0 - l + 1, \\ 0, & \text{falls } j \leq l_0 - l \ \vee \ h_0 - l + 1 < j \leq p. \end{cases} \\
d_{ij} &= \begin{cases} v_{i,l-l_i+j}, & \text{falls } l_i - l < j \leq h_i - l + 1, \\ 0, & \text{falls } j \leq l_i - l \ \vee \ h_i - l + 1 < j \leq p. \end{cases}
\end{aligned}
$$

In Worten ausgedrückt: Die Intensitätsvektoren v_i werden nach links und rechts um Komponenten mit Einträgen 0 erweitert. Die erweiterten Intensitätsvektoren c und d_i sind Vektoren gleicher Länge und haben in entsprechenden Komponenten jeweils Intensitäten zu gleichen Massen als Einträge.
Unter der Annahme einer exakten Messung existiert dann $(x_1, ..., x_n) \in I\!R^n$ mit

$$(c_1, ..., c_p) = (d_{11}, ..., d_{1p}) \cdot x_1 + ... + (d_{n1}, ..., d_{np}) \cdot x_n, \qquad x_i \geq 0.$$

Da in der Realität die Messungen mit Fehlern behaftet sind, gilt i. a.

$$(c_1, ..., c_p) - (d_{11}, ..., d_{1p}) \cdot x_1 - ... - (d_{n1}, ..., d_{np}) \cdot x_n \neq 0.$$

Berücksichtigt man alle Fragmente, die zur betrachteten Peakgruppe c beitragen, so sollte $(x_1, ..., x_n)$ zu finden sein, so daß nur eine „sehr geringe" Abweichung resultiert. Wir suchen $(x_1, ..., x_n)$ für den kleinsten Meßfehler. Dazu minimieren wir die Summe der Fehlerquadrate in den einzelnen Komponenten. Dabei ist zu berücksichtigen, daß

$x_1, \ldots, x_n$ als Häufigkeiten nicht negativ sein dürfen. Das zugehörige Optimierungsproblem (Constrained Least Squares) läßt sich wie folgt formulieren:

$$\begin{aligned} \min \quad & \sum_{j=1}^{p} (c_j - \sum_{i=1}^{n} x_i d_{ij})^2, \\ \text{NB:} \quad & x_i \geq 0 \qquad i = 1, \ldots, n. \end{aligned}$$

Dieser Ansatz trägt der Tatsache Rechnung, daß große Fehler in wenigen Komponenten unwahrscheinlicher sind als geringfügige Abweichungen in mehreren Komponenten. Durch Quadrierung der Einzelfehler erfolgt eine entsprechende Gewichtung.

3 Die Peakgruppe des Molekülions

Zur Peakgruppe des Molekülions M^+ können neben dem Molekülion selbst nur Fragmente, die durch H-Abspaltung aus dem Molekülion entstanden sind, beitragen. Alle anderen Fragmente weisen im Bezug zum Molekülion einen zu großen Massenunterschied (≥ 13) auf. Für obiges Optimierungsproblem ergibt sich daraus folgender Spezialfall:
Sei $\mathcal{P}_0 = (l_0, h_0, v_0)$ die Peakgruppe des Molekülions, $\mathcal{P}_1 = (l_1, h_1, v_1)$ das theoretische Isotopenmuster des Molekülions. Dann sind höchstens $l_1 - l_0$ Wasserstoffabspaltungen zu beachten, d.h. $n = l_1 - l_0 + 1$ Fragmente werden berücksichtigt: $F_1 := M^+$, $F_2 := [M-H]^+$, $F_3 := [M-2H]^+, \ldots, F_n := [M-(n-1)H]^+$
Wie in Abschnitt 2 erklären wir $l := \min(l_0, l_1)$, $h := \max(h_0, h_1)$, $p := h - l + 1$ und die Vektoren $c, d_1 \in I\!R^p$ mit

$$\begin{aligned} c_j &= \begin{cases} v_{0,l-h_0+j}, & \text{falls } l_0 - l < j \leq h_0 - l + 1, \\ 0, & \text{falls } j \leq l_0 - l \ \vee \ h_0 - l + 1 < j \leq p. \end{cases} \\ d_{1j} &= \begin{cases} v_{1,l-l_1+j}, & \text{falls } l_1 - l < j \leq h_1 - l + 1, \\ 0, & \text{falls } j \leq l_1 - l \ \vee \ h_1 - l + 1 < j \leq p. \end{cases} \end{aligned}$$

Die erweiterten Intensitätsvektoren von $F_1, \ldots, F_n$ haben dann folgende Form:

$$\begin{aligned} d_1 &= (\underbrace{0, \ldots, 0}_{l-l_1}, v_{11}, \ldots, v_{1,h_1-l_1+1}, \underbrace{0, \ldots, 0}_{h-h_1}), \\ d_2 &= (\underbrace{0, \ldots, 0}_{l-l_1-1}, v_{11}, \ldots, v_{1,h_1-l_1+1}, \underbrace{0, \ldots, 0}_{h-h_1+1}), \\ &\vdots \\ d_n &= (v_{11}, \ldots, v_{1,h_1-l_1+1}, \underbrace{0, \ldots, 0}_{h-h_1+n-1}). \end{aligned}$$

Ist die Peakgruppe des Molekülions vorhanden, so kann auf folgende Weise ein Ranking für die Vorschläge zur Molekülion Summenformel vorgenommen werden: Für jeden Kandidaten ist sein theoretische Isotopenmuster $\mathcal{P}_1$ zu berechnen und das constrained

Least-Squares Problem zu lösen. Das Minimum der Fehlerquadratsumme wird dabei als Vergleichswert herangezogen. Für die korrekte Summenformel ist dieser Vergleichswert also höchstens gleich der Summe aus den Quadraten der tatsächlichen Meßfehler. Bei falschen Kandidaten tragen die unpassenden theoretischen Isotopenmuster zur Vergleichsgröße bei. Deshalb sind hinreichend genaue Messungen Voraussetzung für die Anwendbarkeit des Verfahrens.

Bei Spektren mit gut ausgeprägter Peakgruppe des Molekülions und Elementkombinationen mit auffälligem Isotopenmuster kann oft schon auf diese Weise die Summenformel bestimmt werden. Um Fehlinterpretationen leichter ausschließen zu können, ist es wichtig, auch Informationen über Fragmentionen zu berücksichtigen (vgl. [2], [5]). Entscheidend dabei ist, daß alle Fragment-Summenformeln die Teilmengenrelation zur Molekül-Summenformel erfüllen: *Für jedes Element X ist die Anzahl der Atome vom Element X in der Fragment-Summenformel nicht größer als die Anzahl der Atome vom Element X in der Molekül-Summenformel.* Diese Regel dient als Grundlage für folgende

4 Verallgemeinerung

Gegeben sei ein Kandidat F_1 für die Molekül-Summenformel und ein Massenspektrum. Wir schreiben auch das Spektrum als Tripel $\mathcal{P}_0 = (l_0, h_0, v_0)$, wobei $l_0 = 1$, $h_0 = \max\{m : I_S(m) > 0\}$ und $v_{0j} = I_S(j)$. Wir betrachten nun analog das gesamte Spektrum als Überlagerung von Isotopenmustern von Fragmenten, deren Summenformeln der Teilmengenrelation zu F_1 genügen. Wie in Abschnitt 2 stellen wir das Spektrum als Linearkombination der theoretischen Isotopenmuster der möglichen Fragmente dar, so daß die Summe der Quadrate der Abweichungen in den einzelnen Komponenten minimiert wird. Ziel ist die Bewertung des Kandidaten aufgrund der Erklärbarkeit aller Peaks des Spektrums:

4.1 Algorithmus

1. Erzeuge für alle j mit $c_{0j} \neq 0$ alle Elementkombinationen mit Masse j, die der Teilmengenrelation zu F_1 genügen. Man erhält die Menge der Summenformeln aller möglichen Fragmentionen.

2. Berechne die theoretischen Isotopenmuster $\mathcal{P}_1, \ldots, \mathcal{P}_n$ zu den Summenformeln der möglichen Fragmentionen.

3. Führe die Optimierung durch wie in Abschnitt 2 und nehme die Wurzel der Least-Squares-Summe als Vergleichswert.

4.2 Bemerkung

Wendet man diesen Algorithmus auf alle Kandidaten für die Summenformel an, so erhält man ein Ranking der Kandidaten aufgrund ihrer minimalen Fehlerquadratsummen. Diejenige Summenformel, zu der das gemessene Spektrum am besten als Linearkombination der theoretischen Isotopenmuster möglicher Fragmente dargestellt werden kann, ist führend in unserer Hitliste. Unter der theoretischen Annahme einer *exakten* Messung ist der Vergleichswert der korrekten Summenformel null. Bei einer guten Messung ist ein kleiner Vergleichswert notwendiges Kriterium für die korrekte Summenformel. Natürlich gibt es Spektren, bei denen auch falsche Summenformeln gute Werte erzielen. Diesem Effekt könnte schon beim Generieren der Summenformeln durch Konsistenztests mit Ergebnissen anderer Methoden [6] entgegengewirkt werden. Doch werden meist schon bei mangelnder Erklärbarkeit einzelner Peaks signifikant höhere Vergleichswerte berechnet, wie auch Beispiel 5.1 deutlich macht.

5 Ergebnisse

5.1 Beispiel

Im folgenden führen wir ein einfaches Beispiel an, welches die Arbeitsweise des Algorithmus demonstriert. Tabelle 2 enthält gemessene Intensitäten (Int) für Benzol ([3], Unknown 2.5.), sowie die optimierte Darstellung (Opt) aus den theoretischen Isotopenmustern der möglichen Fragmente von C_6H_6. Tabelle 3 enthält die Ein- und Ausgaben für die Optimierung: Es sind $n = 28$ Summenformeln zu berücksichtigen. Die erste Spalte gibt m/z-Werte an, bei denen Intensitäten ungleich Null auftreten. Dabei sind nur solche Werte relevant, die höchstens gleich der nominalen Masse des Kandidaten sind[2]. In der zweiten Spalte findet man die möglichen Fragment Summenformeln zu diesen Massen, gefolgt von den theoretischen Isotopenmustern und schließlich den Faktoren x_i, die das Optimierungsproblem lösen.
Zu dem Kandidaten C_6H_6 für die Summenformel erhält man 0,28 als Vergleichswert. Die Vergleichswerte der nächstplazierten Kandidaten sind in Tabelle 4 zusammengestellt. Abbildung 2 zeigt das Spektrum von Benzol, Abbildungen 3–7 zeigen die absoluten Abweichungen zwischen gemessenem Spektrum und dem jeweiligen optimalem theoretischen Spektrum für die fünf besten Summenformel-Kandidaten.

[2] In dem angegebenen Beispiel gibt es für jede Masse nur eine Summenformel. Im allgemeinen trifft dies nicht zu.

m/z	Int	Opt	Diff	m/z	Int	Opt	Diff	m/z	Int	Opt	Diff
12.00	0.20	0.20	0.00	38.00	0.00	0.12	0.12	64.00	0.20	0.20	0.00
13.00	0.40	0.40	0.00	39.00	13.00	13.00	0.00	72.00	0.40	0.40	0.00
14.00	0.40	0.40	0.00	40.00	0.40	0.43	0.03	73.00	1.00	1.00	0.00
15.00	1.00	1.00	0.00	50.00	16.00	16.00	0.00	74.00	3.90	3.90	0.00
24.00	0.40	0.40	0.00	51.00	19.00	19.00	0.00	75.00	2.20	2.20	0.00
25.00	0.80	0.80	0.00	52.00	20.00	20.00	0.00	76.00	7.00	7.00	0.00
26.00	3.20	3.20	0.00	53.00	0.80	0.86	0.06	77.00	15.00	15.00	0.00
27.00	2.60	2.60	0.00	60.00	0.20	0.20	0.00	78.00	100.00	100.02	0.02
28.00	0.00	0.06	0.06	61.00	0.40	0.40	0.00	79.00	6.80	6.56	0.24
36.00	0.90	0.90	0.00	62.00	0.80	0.80	0.00	80.00	0.20	0.18	0.02
37.00	3.80	3.80	0.00	63.00	2.90	2.90	0.00				

Tabelle 2: Gemessenes und optimales theoretisches Spektrum von Benzol.

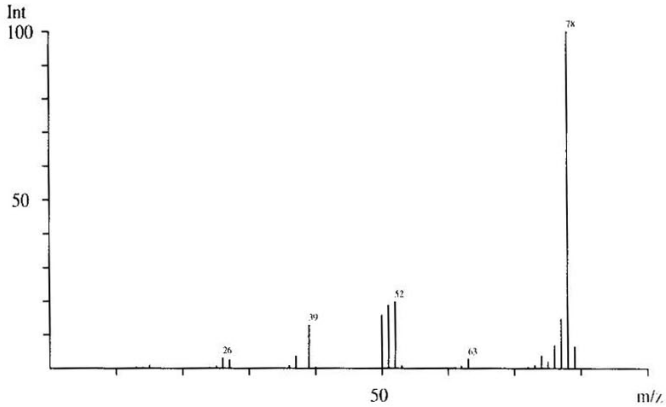

Abbildung 2: Experimentelles Massenspektrumpektrum von Benzol.

m/z	Fragment Summenformel	theoretisches Isotopenmuster	Faktor
78	C_6H_6	(78, 80, (100, 6.6, 0.18))	0.9901
77	C_6H_5	(77, 79, (100, 6.6, 0.18))	0.1454
76	C_6H_4	(76, 78, (100, 6.6, 0.18))	0.0686
75	C_6H_3	(75, 77, (100, 6.6, 0.18))	0.0195
74	C_6H_2	(74, 76, (100, 6.6, 0.18))	0.0384
73	C_6H	(73, 75, (100, 6.6, 0.18))	0.0097
72	C_6	(72, 74, (100, 6.6, 0.18))	0.0040
64	C_5H_4	(64, 66, (100, 5.5, 0.12))	0.0001
63	C_5H_3	(63, 65, (100, 5.5, 0.12))	0.0286
62	C_5H_2	(62, 64, (100, 5.5, 0.12))	0.0078
61	C_5H	(61, 63, (100, 5.5, 0.12))	0.0039
60	C_5	(60, 62, (100, 5.5, 0.12))	0.0020
53	C_4H_5	(53, 55, (100, 4.4, 0.07))	0.0000
52	C_4H_4	(52, 54, (100, 4.4, 0.07))	0.1918
51	C_4H_3	(51, 53, (100, 4.4, 0.07))	0.1830
50	C_4H_2	(50, 52, (100, 4.4, 0.07))	0.1600
40	C_3H_4	(40, 42, (100, 3.3, 0.04))	0.0000
39	C_3H_3	(39, 41, (100, 3.3, 0.04))	0.1300
37	C_3H	(38, 40, (100, 3.3, 0.04))	0.0377
36	C_3	(37, 39, (100, 3.3, 0.04))	0.0090
27	C_2H_3	(27, 29, (100, 2.2, 0.01))	0.0253
26	C_2H_2	(26, 28, (100, 2.2, 0.01))	0.0318
25	C_2H	(25, 27, (100, 2.2, 0.01))	0.0079
24	C_2	(24, 26, (100, 2.2, 0.01))	0.0040
15	CH_3	(15, 16, (100, 1.1))	0.0100
14	CH_2	(14, 15, (100, 1.1))	0.0040
13	CH	(13, 14, (100, 1.1))	0.0040
12	C	(12, 13, (100, 1.1))	0.0020

Tabelle 3: Mögliche Fragment Summenformeln von C_6H_6, deren Isotopenmuster und Anteile im optimalen theoretischen Spektrum von Benzol.

Kandidat	Vergleichswert
C_6H_6	0.28
$C_3F_2H_4$	3.75
$C_4H_2N_2$	4.93
C_2FH_4P	14.41
$C_2FH_3O_2$	14.84

Tabelle 4: Kandidaten für die Summenformel und deren Vergleichswerte für das betrachtete Spektrum von Benzol.

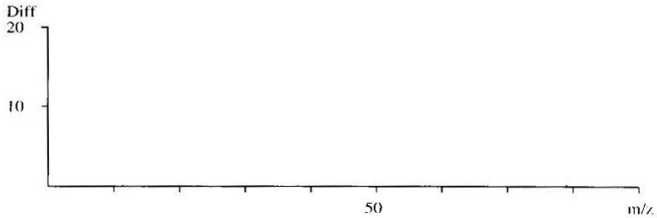

Abbildung 3: Absolute Abweichungen zwischen experimentellem Spektrum von Benzol und optimalem theoretischen Spektrum für Kandidat C_6H_6.

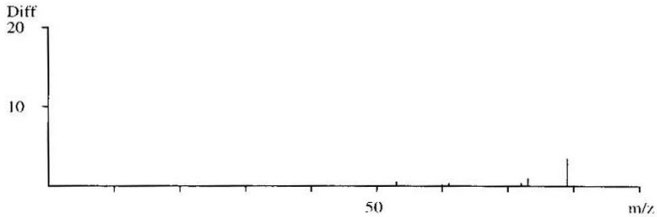

Abbildung 4: Absolute Abweichungen zwischen experimentellem Spektrum von Benzol und optimalem theoretischen Spektrum für Kandidat $C_3F_2H_4$.

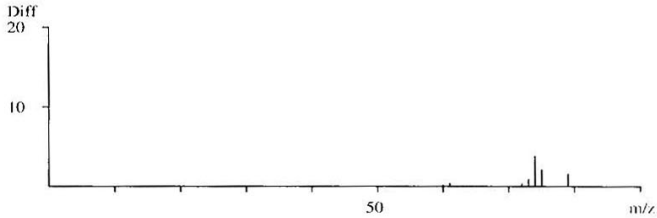

Abbildung 5: Absolute Abweichungen zwischen experimentellem Spektrum von Benzol und optimalem theoretischen Spektrum für Kandidat $C_4H_2N_2$.

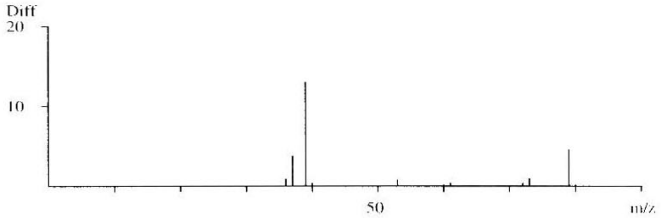

Abbildung 6: Absolute Abweichungen zwischen experimentellem Spektrum von Benzol und optimalem theoretischen Spektrum für Kandidat C_2FH_4P.

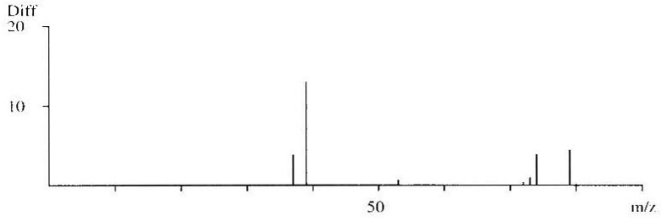

Abbildung 7: Absolute Abweichungen zwischen experimentellem Spektrum von Benzol und optimalem theoretischen Spektrum für Kandidat $C_2FH_3O_2$.

Abbildung 8: Vorgehensweise zum Testen des Verfahrens.

5.2 Statistik

Zur Beurteilung der Anwendbarkeit eines solchen Verfahrens genügt es nicht, sich auf ausgewählte Beispiele zu beschränken. Um aussagekräftige statistische Nachweise für die Qualität der Vergleichswerte darzulegen, wurden Spektren von Substanzen mit den oben angegebenen 11 Elementen und Molekülmasse bis 150 amu aus einer Spektrendatenbank [3] extrahiert. Unser Anliegen besteht darin, bei gegebenem Spektrum nicht nur ein Ranking auf den Summenformel-Kandidaten vorzunehmen, sondern eine Menge von Kandidaten anzugeben, die die korrekte Summenformel mit vorgegebener Verlässlichkeit enthält.
Für die 5254 extrahierten Spektren wurden zunächst die Vergleichswerte bzgl. der korrekten Summenformeln gemäß Algorithmus 4.1 berechnet, aufsteigend sortiert und Grenzen ermittelt, so daß $p = 50, \ldots, 99$ Prozent der Werte nicht über dieser Grenze $b(p)$ liegen. Diese Grenzen dienen im folgenden dazu, für einen beliebigen Summenformel-Kandidaten zu entscheiden, ob er bei vorgegebener Zuverlässigkeit akzeptiert werden soll, oder nicht. Generiert man dann zu einem der 5254 Spektren *alle möglichen* Summenformeln, gibt eine Zuverlässigkeit p vor und wählt alle diejenigen aus, deren Vergleichswert nicht über der Grenze $b(p)$ liegen, so gehört mit p Prozent Wahrscheinlichkeit die richtige Summenformel zu den akzeptierten. Dabei sind *„alle möglichen"* Summenformeln so zu wählen, daß sich die korrekte unter ihnen befindet.

In unserer Versuchsanordnung (vgl. Abb. 8) wird die korrekte Molekülmasse eingelesen. Um die aufwendige Optimierung nicht mit allen Summenformeln zur vorgegebenen Masse durchführen zu müssen (bei Masse 150 wären dies 764, wenn alle 11 Elemente aus Tabelle 1 berücksichtigt werden), wird zunächst ein erstes, auf Isotopenmustertests und Überprüfung von Teilmengenbeziehungen basierendes Ranking vorgenommen. Diese Berechnungen sind wesentlich einfacher und schneller durchführbar als die Ermittlung der kleinsten Fehlerquadrate, aber genau genug, um sie als Filter zu verwenden. Für den aufwendigeren Least-Squares-Fit werden dann nur noch die besten 50 (bzw. 100) Kandidaten des ersten Rankings herangezogen. Tabelle 5 faßt die Ergebnisse der Untersuchungen zusammen. Dabei steht in Spalte

Vorgabe: die vorgegebene Zuverlässigkeit p (in Prozent);

Grenze: die zugehörige Grenze $b(p)$ für Vergleichswerte, bis zu der Kandidaten akzeptiert werden;

Top 50: die Ergebnisse, falls man die besten 50 Kandidaten des ersten Rankings zum Least-Squares-Fit heranzieht;

Top 100: die Ergebnisse, falls man die besten 100 Kandidaten des ersten Rankings zum Least-Squares-Fit heranzieht;

[3] Chemical Concepts Quality Collection, Chemical Concepts GmbH, Postfach 10 02 02, D 69412 Weinheim

		Top 50		Top100	
Vorgabe	Grenze	Kand	Quote	Kand	Quote
50	0.58	10.8	46.27	15.5	48.61
75	1.19	17.1	69.03	24.8	72.63
80	1.39	18.5	73.58	27.0	77.43
90	2.33	23.0	82.45	34.3	86.77
95	3.89	28.2	86.87	43.1	91.55
99	8.81	33.7	90.27	54.6	95.36

Tabelle 5: Ergebnisse des Interpretationsalgorithmus.

Kand: jeweils die durchschnittliche Anzahl der akzeptierten Kandidaten,

Quote: jeweils das Verhältnis der Anzahl korrekter Summenformeln unter den akzeptierten Kandidaten zur Anzahl untersuchter Spektren (in Prozent);

Tabelle 5 zeigt, daß bei Betrachtung der 100 besten Kandidaten des ersten Rankings im Vergleich zur Top 50 Variante die Anzahl der durchschnittlich vorgeschlagenen Kandidaten ca. um Faktor $1\frac{1}{2}$ wächst. Dem Benutzer bleibt überlassen, ob der Zugewinn an Sicherheit die größere Anzahl akzeptierter Kandidaten rechtfertigt.
Die Interpretation der 5254 Spektren benötigte ca. 2 Tage (Top 50) bzw. 4 Tage (Top 100) auf einem PC Pentium Pro mit 200 MHz. Möchte man auf das erste Ranking verzichten, so müssen wesentlich längere Rechenzeiten in Kauf genommen werden. Nach Wahl der Grenzen $b(p)$ würde man dann gerade die Vorgabe als Quote erhalten.

5.3 Implementation

In das C++ Programm zur Ermittlung von guten Summenformel Kandidaten aus niedrig aufgelösten Massenspektren wurde ein in Fortran implementierter Algorithmus [4] zur Bearbeitung des Optimierungsproblems eingebunden. Dabei handelt es sich um ein sequentielles quadratisches Programmierungsverfahren (SQP Verfahren). Dieses beruht auf der sukzessiven Lösung quadratischer Teilprobleme, die durch eine quadratische Approximation der Lagrange Funktion und eine Linearisierung möglicher Restriktionen entstehen.
Die Interpretation von Spektren mit Massen unter 300 amu kann in wenigen Minuten durchgeführt werden. In der derzeitigen Version muß als Voraussetzung für eine erfolgreiche Interpretation der Molekülionpeak im Spektrum vorhanden sein. Der entscheidende laufzeitkritische Faktor ist die exponentiell wachsende Anzahl von Summenformeln zu gegebener Molekülmasse. Dies kann umgangen werden, indem man die Menge der möglichen Elemente (Defaulteinstellung: H, C, N, O, F, Si, P, S, Cl, Br und I) verkleinert.

5.4 Abschließende Bemerkung

Ziel zukünftiger Entwicklungen ist es, die vorgestellte Methode zu verfeinern und durch Kopplung mit der MS-Klassifizierungssoftware MSclass [6] und dem Strukturgenerator MOLGEN [1] ein Expertensystem zur Bestimmung der Strukturformel aus niedrig aufgelösten Massenspektren verfügbar zu machen. Das erfolgreiche Zusammenspiel von MSclass und MOLGEN wurde bereits mehrfach dokumentiert (vgl. [7]). Ein Summenformelgenerator versehen mit dem vorgestellten Bewertungsalgorithmus schließt eine weitere Lücke auf dem Weg zur vollautomatisierten MS-Interpretation.

Literatur

[1] C. Benecke, T. Grüner, A. Kerber, R. Laue, T. Wieland: *Molecular Structure Generation with MOLGEN, new Features and Future Developments.* Fresenius J. Anal. Chem. Vol. 358, 23-32 (1997).

[2] K. Kumar, A.G. Menon: *Computer-assisted Determination of Elemental Composition of Fragments in Mass Spectra.* Rapid Communications in Mass Spectrometry, Vol. 6, 585-591 (1992).

[3] F.W. McLafferty, F. Turecek: *Interpretation of Mass Spectra.* Fourth Edition, University Science Books, Mill Valley, California (1993).

[4] K. Schittkowsky: *NLPQL: A FORTRAN Subroutine Solving Constrained Nonlinear Programming Problems.* Annals of Operations Research, Vol. 5, 485-500 (1985).

[5] A. Tenhosaari: *Computer-assisted Composition Analysis of Unknown Compounds by Simultaneous Analysis of the Intensity Ratios of Isotope Patterns of the Molecular Ion and Daughter Ions in Low-resolution Mass Spectra.* Organic Mass Spectrometry, Vol. 23, 236-239 (1988).

[6] K. Varmuza, W. Werther: *Mass Spectral Classifiers for Supporting Systematic Structure Elucidation.* Journal of Chemical Information and Computer Sciences, Vol. 36, 323-333 (1996).

[7] K. Varmuza, W. Werther, F. Stancl, A. Kerber, R. Laue: *Computer-assisted Structure Elucidation of Organic Compounds, based on Mass Spectra Classification and Exhaustive Isomer Generation.* Software Development in Chemistry, Vol. 10, 303-314 (1996).