

Forschung, Wissenschaft

Computerunterstützte Spektreninterpretation für die Strukturaufklärung organischer Verbindungen *

J. T. Clerc **

Laboratorium für Organische Chemie der Eidgenössischen Technischen Hochschule

Abstract

Three different methodical approaches to the computer aided interpretation of spectra are reviewed, namely artificial intelligence, pattern recognition, and library search. Design considerations for a generalised library search system are given.

1. Einleitung

Spektroskopische Methoden sind heute zu einem wichtigen Werkzeug in der Hand des Analytikers geworden, das er insbesondere für die Identifikation und Strukturaufklärung organischer Verbindungen routinemässig einsetzt. Dies trifft insbesondere für die Infrarot-Spektroskopie, die Spektroskopie im sichtbaren und ultraviolett Spektralbereich, für die Kernresonanzspektroskopie sowie für die Massenspektroskopie zu, auf die sich die folgenden Betrachtungen im wesentlichen beschränken. Bei der Registrierung und Aufarbeitung spektroskopischer Daten ist der Einsatz von Computern in manchen Anwendungsfällen bereits zur Selbstverständlichkeit geworden [1–3]. Im Gegensatz dazu haben computerunterstützte Interpretationsmethoden noch nicht jenen Reifegrad erreicht, der für einen allgemeinen Einsatz in der Praxis Voraussetzung wäre. In der vorliegenden Arbeit soll die grundsätzliche Problematik der computerunterstützten Spektreninterpretation kurz umrissen und über einen Lösungsweg berichtet werden, der seine Feuerprobe im praktischen Einsatz bereits einigermassen erfolgreich bestanden hat.

2. Systemanalyse

Jede Automatisierung muss notwendigerweise mit einer Analyse des zu bearbeitenden Systems beginnen. Dementsprechend soll auch hier zum mindesten im Ansatz eine Systemanalyse des Gesamtprozesses durchgeführt werden. Bei der Aufnahme eines Spektrums einer chemischen Verbindung wird die in der Struktur der Probe enthaltene Information durch das Spektrometer in die spektroskopischen Daten umgesetzt. Wir können also die spektroskopischen Daten sozusagen als «verschlüsselte Struktur» einer Verbindung auffassen. Um aus den spektroskopischen Daten

einer Verbindung ihre Struktur zurückzugewinnen, müssen wir demnach lediglich die Codierungsregeln in umgekehrter Richtung anwenden, die Spektren also decodieren. Der Umsetzung dieser Betrachtungsweise in die Praxis stehen nun allerdings einige wesentliche Hindernisse entgegen. Einmal ist die Umformung der Struktur in spektroskopische Daten immer mit einem gewissen Informationsverlust verbunden, so dass die spektroskopischen Daten die ursprüngliche Struktur nicht mehr eindeutig und in allen Einzelheiten beschreiben. Im weiteren basieren unsere Kenntnisse über die Zusammenhänge zwischen Struktur und spektroskopischen Daten auf stark vereinfachten Modellvorstellungen, die die Realität meist nur unvollkommen beschreiben. Sie sind in erster Linie dazu geeignet, aus Teilstrukturen Teilspektren vorauszuberechnen und liefern dabei im Regelfall eindeutige Resultate mit einer für die Praxis ausreichenden Genauigkeit. In der umgekehrten Richtung angewendet sind die Resultate jedoch oft nicht eindeutig, indem mit einem gegebenen Satz spektroskopischer Daten im allgemeinen mehrere unterschiedliche Teilstrukturen vereinbar sind. Dementsprechend geht der Analytiker bei der semiempirischen Interpretation von Spektren zur Identifikation und Strukturbestimmung organischer Verbindungen [4] im allgemeinen so vor, dass er vorerst aufgrund recht allgemeiner Korrelationsregeln einen Satz von solchen Strukturelementen zusammenstellt, deren Anwesenheit durch die spektroskopischen Daten weitgehend gesichert ist. Aus diesen Teilstrukturen baut er sich dann einen plausiblen Vorschlag für die Struktur der Unbekannten auf. Für diese Struktur werden dann die spektroskopischen Daten vorausgesagt, sei es durch Anwendung exakter Berechnungen, aufgrund von vereinfachten Modellvorstellungen oder mit Hilfe von Referenzdaten. Die vorausgesagten Daten werden nun mit den gemessenen verglichen. Ist die Übereinstimmung gut, so wird angenommen, dass die vorgeschlagene Struktur der tatsächlichen entspricht. Andernfalls wird der Strukturvorschlag in geeigneter Weise modifiziert. Dann werden die Erwartungswerte für den neuen Strukturvorschlag bestimmt und erneut mit den Messwerten verglichen. Dieser iterative Prozess läuft nun so lange ab, bis keine weiteren brauchbaren Strukturvorschläge mehr gefunden werden. Selbstverständlich wird sich der praktische Analytiker dabei nicht auf die

* Nach einem anlässlich der Herbsttagung der Schweizerischen Chemischen Gesellschaft, Genf, 8. bis 9. Oktober 1976, gehaltenen Vortrag.

** PD Dr. J. T. Clerc, Laboratorium für Organische Chemie der ETH, Universitätsstrasse 16, CH-8006 Zürich

ausschliessliche Verwendung spektroskopischer Kriterien beschränken, sondern alle ihm zur Verfügung stehende Information über die unbekannte Probe zur Problemlösung beiziehen.

3. Algorithmische Methoden

Der naheliegendste Ansatz zur Automatisierung der Spektreninterpretation besteht in der vollen Automatisierung aller in der Systemanalyse genannten Teilschritte. In entsprechenden Computerprogrammen werden die Denk- und Entscheidungsprozesse eines menschlichen Analytikers nachvollzogen, der Analytiker also sozusagen auf dem Computer simuliert. Solche Programme werden dementsprechend oft mit der Etikette «Artificial Intelligence» versehen.

Es sind einige Programme bekannt geworden, die die automatische Spektreninterpretation in dieser Weise zu realisieren versuchen, wobei den verschiedenen Teilschritten oft recht unterschiedliches Gewicht beigemessen wird [5–8].

Die bei diesen Verfahren auftretenden Probleme sind jedoch teilweise so komplex, dass für die Praxis brauchbare Systeme in der überschaubaren Zukunft nicht zu erwarten sind. Insbesondere sind die heutigen Verfahren zur Speicherung und Manipulation chemischer Strukturen [9] noch viel zu schwerfällig und zu teuer für einen sinnvollen Einsatz in diesem Problemkreis. Ebenso bereitet die streng algorithmische Formulierung spektroskopischer Interpretationsregeln einigermaßen allgemeiner Gültigkeit heute noch unüberwindliche Schwierigkeiten. Solche Programme zur vollautomatischen Spektreninterpretation sind daher im wesentlichen auf Stoffklassen beschränkt geblieben, die ein besonders einfaches und leicht genau voraussagbares spektroskopisches Verhalten zeigen. Dies sind aber genau jene Stoffklassen, bei denen die Strukturanalyse mit spektroskopischen Methoden einfach, oft sogar trivial wird. Dementsprechend bieten die heute verfügbaren algorithmischen Programme zur automatischen Spektreninterpretation keinen nennenswerten Nutzen für die Praxis. Es ist aber mit Sicherheit falsch, sie als intellektuelle Spielerei abzutun. Die Entwicklung solcher Programme hat nicht nur auf dem Gebiet der semiempirischen Spektreninterpretation viel zum besseren Verständnis der Probleme beigetragen, sondern hat auch die Entwicklung auf anderen Gebieten wesentlich und positiv beeinflusst.

4. «Pattern Recognition» und verwandte Methoden

Unter den (oft unpräzise verwendeten) Überbegriff «Pattern Recognition» werden gemeinhin solche computerunterstützten Interpretationsmethoden zusammengefasst, bei denen eine unbekannte Verbindung aufgrund ihrer spektroskopischen Daten mit Hilfe eines Algorithmus vorgegebener Struktur einer von zwei disjunkten Klassen zugeordnet wird. Die Umformung der spektroskopischen Daten zur Verarbeitung

durch den Klassierungs-Algorithmus wie auch die eigentliche Klassierung erfolgen dabei weitgehend ohne Einsicht in die physikalisch-chemischen Vorgänge bei der Spektrenregistrierung, aber auch unbelastet durch Spekulationen und/oder übervereinfachende Modellvorstellungen. Vielmehr werden statistische Kriterien zum Aufbau einer Entscheidungsregel herangezogen. Man geht dabei von einem Satz von Spektren bekannter Struktur aus, der aufgrund des zu unterscheidenden Merkmals (z.B. An- bzw. Abwesenheit einer vorgegebenen Teilstruktur) in zwei disjunkte Teilmengen aufgeteilt wird. Mit Hilfe dieser Spektren werden nun die Parameter des (seiner Form nach willkürlich festgelegten) Entscheidungs-Algorithmus für beste Wiedererkennung optimiert. Neben klassischen Optimierungsverfahren [10] werden dabei oft iterative Programme eingesetzt, bei denen die Optimierung strukturell einem Lernprozess entspricht [11]. Die Implementation eines solchen Programms wird dementsprechend als «Lernmaschine» [12] bezeichnet. Den Satz von Spektren bekannter Klassenzugehörigkeit, der dem Optimierungsprozess zugrunde liegt, nennt man dann den Trainings-Satz. Die Funktionsweise einer einfachen Lernmaschine soll im folgenden kurz dargestellt werden.

Die Trainings-Spektren müssen vorerst in eine für die Bearbeitung durch die Lernmaschine geeignete Form gebracht werden. Dazu werden sie durch eine Serie von Merkmalen (Messgrößen) beschrieben. Es kann sich dabei zum Beispiel um Peakpositionen, Peakintensitäten oder Peakbreiten handeln. Auch komplexere Merkmale wie Anzahl Peaks, Verteilung der Peaks innerhalb des betrachteten Spektralbereichs u.a. sind ohne weiteres denkbar. Wesentlich ist einzig, dass sich die gewählten Merkmale als einfache Zahl angeben lassen. Als nächstes muss eine Entscheidungsregel formuliert werden. Diese Entscheidungsregel ist nichts anderes als eine Rechenvorschrift, die auf die gewählten Merkmale angewendet wird. Das Resultat dieser Rechnung soll anzeigen, zu welcher der beiden Klassen das bearbeitete Spektrum gehört. Eine mögliche Entscheidungsregel könnte beispielsweise die gewichtete Summe aller Merkmale berechnen und das Vorzeichen dieser Summe mit den beiden Klassen korrelieren. Die bei der Summierung verwendeten Gewichte der einzelnen Merkmale bilden dabei den zu optimierenden Parameter-Satz.

Die gewählte Entscheidungsregel wird nun, beginnend mit einem frei gewählten Satz von Werten für die Parameter, der Reihe nach auf alle Spektren des Trainings-Satzes angewendet. Immer dann, wenn ein Spektrum falsch klassiert wird, werden die Parameter der Entscheidungsregel in geeigneter Weise so verändert, dass eine erneute Klassierung nun das korrekte Resultat liefert. Dieser «Lernprozess mit negativer Rückkopplung» läuft nun so lange ab, bis alle Trainings-Spektren richtig klassiert werden oder wird nach einer bestimmten Zeitspanne abgebrochen. Ob und gegebenenfalls

wie schnell dieses Verfahren konvergiert, hängt in hohem Masse von der Art des Korrekturverfahrens für den Parametersatz ab. Betrachtungen darüber würden den Rahmen dieser Arbeit sprengen.

Von der resultierenden Entscheidungsregel erwartet man nun, dass sie Spektren unbekannter Verbindungen mit einer Sicherheit klassiert, die der Wiedererkennungsrate bei den Trainings-Spektren nahe kommt. In der Praxis werden solche Erwartungen in den meisten Fällen allerdings nicht erfüllt. Dafür ist in erster Linie die im folgenden gegebene Ursache massgebend.

Der verwendete Satz von Trainings-Spektren muss für den Bereich der späteren praktischen Anwendung repräsentativ sein. Dementsprechend muss er eine repräsentative Stichprobe solcher Verbindungen enthalten, die das interessierende Strukturmerkmal aufweisen, also zur Ja-Klasse gehören. Dies kann mit entsprechendem Aufwand meist realisiert werden. Andererseits muss aber auch die Nein-Klasse, also der Satz jener Trainings-Verbindungen, die das interessierende Strukturmerkmal nicht enthalten, repräsentativ sein. In dieser Teilmenge muss also der gesamte Bereich der organischen Chemie angemessen vertreten sein. Eine einigermaßen repräsentative Stichprobe bedingt daher eine sehr grosse Zahl von Nein-Spektren, so dass der Trainings-Satz im Regelfall durch die Nein-Spektren dominiert wird. Der Lernvorgang wird sich dann vorwiegend an diesen Spektren orientieren, so dass die resultierende Entscheidungsregel zwar die Nein-Spektren mit einigermaßen brauchbarer Sicherheit erkennen, für die Ja-Spektren aber weitgehend erratische Resultate liefern wird. Um das zu vermeiden, muss entweder die Zahl der Ja-Spektren drastisch erhöht oder die Zahl der Nein-Spektren limitiert werden. Wenn die erstgenannte Lösung nicht schon daran scheitert, dass nicht genügend geeignete Ja-Spektren zugänglich sind, so resultiert ein so umfangreicher Trainings-Satz, dass beim heutigen Stand der Technik eine Verarbeitung durch iterative Lernmaschinen des beschriebenen Typs aus wirtschaftlichen Gründen nicht vertretbar ist. Im anderen Fall ist es äusserst schwierig wenn nicht gar unmöglich, die Forderung nach einer repräsentativen Stichprobe zu erfüllen. Die mit einem nicht repräsentativen Trainings-Satz erhaltenen Resultate sind für die Praxis aber meist wenig relevant. Es ist zwar üblich, vor Beginn der eigentlichen Lernphase einen Teil der Spektren aus dem Trainings-Satz auszuschneiden und die gefundene Entscheidungsregel an diesen Spektren zu überprüfen. Die dabei erhaltenen Resultate sagen aber wenig aus über die echte Zuverlässigkeit der Entscheidungsregel, da der Satz der Test-Spektren offensichtlich in genau gleicher Weise voreingenommen und unausgeglichen ist wie der Trainings-Satz.

Aus den genannten Gründen sind Entwicklung und praktischer Einsatz von Entscheidungsmaschinen des beschriebenen Typs heute eigentlich nur dann sinnvoll, wenn die Vielfalt der Typen der zu klassierenden Objekte nicht zu gross ist. Das Arbeitsgebiet muss also

klar abgegrenzt und in seiner inneren Struktur gut bekannt sein.

Einblick in die innere Struktur einer Menge von Objekten kann man mit den Methoden der «Cluster Analysis» erhalten [13, 14]. Im Gegensatz zu «Pattern Recognition», wo Gruppen von Objekten mit bestimmten vorgegebenen Eigenschaften gegeneinander abgegrenzt werden, sucht man hier nach Gruppen von Objekten, die sich leicht gegeneinander abgrenzen lassen und versucht nachher die einer solchen Gruppe gemeinsamen Eigenschaften zu bestimmen. Diese Arbeitsweise gewährleistet zwar leistungsfähige und zuverlässige Entscheidungsregeln für «natürliche» Klassen, die so gefundene «natürliche» Klasseneinteilung ist aber nur in seltenen Glücksfällen problembezogen und damit für praktische Anwendungen brauchbar.

Die im grossen und ganzen recht unbefriedigenden Resultate, die im Bereich der analytischen Chemie mit den genannten Methoden erhalten wurden, dürfen aber nicht darüber hinwegtäuschen, dass diesen Methoden ein enormes Entwicklungspotential innewohnt. Die zur Identifikation und Strukturaufklärung organischer Verbindungen notwendige Information ist in den spektroskopischen Daten zweifellos vorhanden, das beweist schon die Interpretierbarkeit der Spektren durch den geübten Analytiker. Was hingegen fehlt sind Kenntnisse darüber, wo diese Information in den Spektren zu finden ist, wie man sie extrahiert und in numerische Form bringt, und auf welche Weise man sie im Computer darstellt und manipuliert. Für Untersuchungen in dieser Richtung sind umfangreiche Sammlungen von spektroskopischen Daten in computerlesbarer Form eine unabdingbare Voraussetzung. Dabei ist es wesentlich, dass diese Sammlungen die gesamte Information der Spektren enthalten, dass also nicht schon bei der Registrierung eine Vorauswahl getroffen wird, bei der ein wesentlicher Teil der Information unwiederbringlich verloren geht. Der Aufbau und vor allem auch der Unterhalt solcher Sammlungen ist äusserst zeit- und arbeitsintensiv und stellt eine Reihe nicht trivialer Probleme, die heute noch nicht zufriedenstellend gelöst sind [9, 15, 16, 20]. Dementsprechend fehlt heute die Basis für seriöse, grundlegende Untersuchungen über die Anwendbarkeit von «Pattern Recognition» und ähnlichen Methoden im Bereich der computerunterstützten Spektreninterpretation noch weitgehend. Mit den enormen Fortschritten im Bereich der Laborautomation sind aber in den letzten Jahren zum mindesten die technischen Voraussetzungen für den Aufbau geeigneter Datensammlungen erfüllt worden, so dass in der überschaubaren Zukunft mit wesentlichen Fortschritten gerechnet werden darf.

5. Bibliotheks-Such-Verfahren

Dieses altbewährte und in der Praxis gut eingeführte Verfahren stützt sich auf den direkten Vergleich der Spektren einer unbekannten Verbindung mit Referenz-

spektren. Man hofft dabei, in der Referenzspektrensammlung auf einen Spektrensatz zu stossen, der mit den Spektren der Unbekannten identisch ist. Dann wird man vermuten, dass die Struktur der Probe mit der Struktur der Referenzverbindung identisch ist. Der Aufwand für die manuelle Verarbeitung grösserer Sammlungen von Referenzspektren wird sehr bald unzumutbar. Wenn aber sowohl die Referenzdaten wie auch die Daten der Unbekannten in maschinenlesbarer Form vorliegen, kann das Durchsuchen der Sammlung und der Vergleich der Spektren einem Computer überbunden werden. Die Maschine wird dann alle Referenzspektren der Reihe nach durchgehen und mit den Daten der Unbekannten vergleichen. Falls sie dabei für einen Spektrensatz perfekte Übereinstimmung findet, wird sie die entsprechende Struktur ausgeben und dem Benutzer mitteilen, dass das Problem gelöst sei.

In der Praxis sind die Verhältnisse allerdings wesentlich komplizierter. Einmal muss man auch bei identischer Struktur zweier Proben mit manchmal beträchtlichen Unterschieden in den spektroskopischen Daten rechnen, insbesondere wenn bei der Registrierung der Spektren nicht sämtliche apparativen Parameter peinlich genau konstant gehalten wurden und wenn die Probenvorbereitung nicht für beide Proben in exakt genau gleicher Weise erfolgte. Diese beiden Vorbedingungen lassen sich in der Praxis auch bei Spektrensammlungen bescheidenen Umfangs nicht erfüllen, so dass immer Abweichungen auftreten werden. Eine Spektrenvergleichsmaschine muss dementsprechend beim Vergleich spektroskopischer Daten eine gewisse Toleranz walten lassen. Sie darf sich nicht stur auf totale Übereinstimmung sämtlicher Einzeldaten festlegen, sondern muss durch technische Artefakte bedingte Abweichungen unberücksichtigt lassen.

Ein weiteres grundsätzliches Problem ergibt sich daraus, dass die Annahme, in der Sammlung seien die Spektren einer mit der Probe identischen Referenzverbindung vorhanden, offensichtlich oft nicht zutrifft. Ein primitives Bibliotheks-Such-System kann dann natürlich keine brauchbare Antwort liefern, ausser der eher trivialen Feststellung, dass die Spektren der unbekannten Probe in der Sammlung nicht registriert seien. Man wird dementsprechend danach trachten, den Vergleich der Spektren so durchzuführen, dass auch solche Referenzverbindungen erkannt werden, die in ihrer Struktur wenigstens teilweise mit der Probe übereinstimmen. Die Vergleichsmaschine soll also auch strukturelle Ähnlichkeiten erkennen. Das Resultat einer Suche wird dann aus einer Liste jener Referenzverbindungen bestehen, die eine der Unbekannten ähnliche Struktur aufweisen, wobei die Eintragungen in dieser Liste zweckmässig nach absteigendem Ähnlichkeitsgrad geordnet werden.

Ein verfeinertes Bibliotheks-Such-System muss also drei Voraussetzungen erfüllen. Es soll gegenüber technischen Artefakten unempfindlich sein, es soll aufgrund

spektroskopischer Daten strukturelle Ähnlichkeiten erkennen, und es soll den Grad dieser Ähnlichkeit quantitativ bewerten. Dazu kommt die selbstverständliche Forderung, dass sich der Aufwand für Aufbau, Unterhalt und Betrieb in einem wirtschaftlich sinnvollen Rahmen halten lässt.

Die beiden erstgenannten Forderungen legen fest, was für Einzeldaten zum Vergleich herangezogen werden sollen. Es muss sich um solche Grössen handeln, die in Verbindungen identischer Struktur konstant bleiben, also durch die Art der Probenvorbereitung und durch instrumentelle Parameter weitgehend unbeeinflusst bleiben. Solche Grössen sind relativ einfach zu finden. Im weiteren sollen sie bei veränderter Struktur wesentlich andere Werte annehmen, also strukturempfindlich sein. Auch diese Forderung kann üblicherweise leicht erfüllt werden. Schliesslich sollen die ausgewählten Grössen aber nur geringe Unterschiede aufweisen, wenn zwei Verbindungen ähnlicher Struktur verglichen werden. Diese dritte Forderung nun ist keineswegs trivial. Sie lässt sich im Regelfall nur damit erfüllen, dass man zur Beschreibung der Spektren solche Grössen auswählt, die weniger den exakten Wert eines spektroskopischen Parameters wiedergeben, sondern eher den allgemeinen Habitus des Spektrums beschreiben. So sind beispielsweise in der Massenspektrometrie genaue Intensitätsangaben für Peaks bei einzelnen ausgewählten Massenzahlen in dieser Hinsicht recht ungünstig. Vorzuziehen sind eher Grössen, die beispielsweise den allgemeinen Verlauf des Totalionenstroms in Funktion des Massenbereichs oder die relativen Intensitäten von Ionenserien beschreiben [17, 18]. Auch in der ^{13}C -Kernresonanzspektroskopie wird man sich weniger an exakten Werten für die chemische Verschiebung orientieren, sondern ein Spektrum viel eher durch die relative Anzahl von Signalen in verschiedenen Spektralbereichen oder durch das Protonen/Kohlenstoff-Verhältnis beschreiben, wie es sich aus dem «off resonance»-entkoppelten Spektrum leicht erhalten lässt [19].

Spektroskopische Parameter dieser Art sind zwar wenig selektiv, wenn es darum geht, eine chemische Verbindung möglichst eindeutig zu identifizieren, hingegen sind sie hervorragend geeignet, um Verbindungen ähnlicher Struktur (Homologe, Verbindungen mit gleichem Gerüst, Derivate, etc.) in einer Spektrensammlung aufzufinden. Beste Resultate sind dann zu erwarten, wenn sich der Spektrenvergleich sowohl auf Merkmale der beschriebenen Art wie auch auf spezifische, mehr selektive Merkmale abstützt, wobei den letztgenannten Merkmalen aber weniger Gewicht zukommt. Von einer ausgewogenen bewerteten Auswahl von Merkmalen beider Typen darf man erwarten, dass die unspezifischen, relativ hoch bewerteten Merkmale die Referenzverbindungen unähnlicher Struktur stark diskriminieren, so dass mehr oder weniger zufällige Übereinstimmungen in spezifischen Merkmalen nicht zum Tragen kommen. Innerhalb einer Gruppe von Referenz-

renzverbindungen ähnlichen Typs sind jedoch die Korrelationen zwischen Strukturelementen und klassischen spektroskopischen Parametern recht zuverlässig. Zur Festlegung der endgültigen Rangfolge der Referenzspektren sind demnach auf Einzeldaten basierende Merkmale gut geeignet.

Ein Bibliotheks-Such-System, das den Spektrenvergleich in der oben skizzierten Art durchführt, wird demzufolge auch dann brauchbare Resultate liefern, wenn keine mit der Unbekannten identische Referenzverbindung in der Sammlung vorhanden ist. In diesem Fall werden strukturell ähnliche Referenzverbindungen ausgegeben. Zudem weisen Systeme dieser Art ein erstaunlich hohes Mass an Toleranz gegenüber menschlichen Fehlern und Irrtümern auf, da sich diese im Regelfall kaum auf den allgemeinen Spektrenhabitus auswirken und damit die hoch bewerteten Vergleichsmerkmale wenig stören [20].

Eine grundsätzliche Beschränkung der Leistungsfähigkeit jedes Bibliotheks-Such-Systems liegt darin, dass auch das raffinierteste Vergleichsverfahren keine strukturell der Unbekannten ähnliche Referenz auffinden und ausgeben kann, wenn keine solche Referenzverbindung in der Spektrensammlung vorhanden ist. Umfangreiche und vor allem auch reichhaltige Sammlungen spektroskopischer Daten in computerlesbarer Form sind demnach auch hier das Fundament jeder weiteren Entwicklung.

6. Das OCETH Bibliotheks-Such-System

Das am Laboratorium für Organische Chemie der ETH Zürich entwickelte Bibliotheks-Such-System berücksichtigt die im vorhergehenden Abschnitt dargelegten Grundsätze vollumfänglich. Sein Aufbau, seine Arbeitsweise sowie die Stärken und Schwächen der gewählten Lösung sollen als Illustration zum Vorausgegangenen hier kurz dargestellt werden.

Als Basis für das OCETH-System dienen uns frei zugängliche maschinenlesbare Spektrensammlungen, die die spektroskopischen Daten in unverkürzter Form enthalten. Es sind dies einmal eine kommerziell erhältliche Sammlung von Massenspektren [21], die in der bei uns vorliegenden Version knapp 10000 Spektren umfasst. Weiter verfügen wir über eine selbst aufgebaute Sammlung von ^{13}C -Kernresonanzspektren mit gegenwärtig etwa 2000 Eintragungen. Computerlesbare Sammlungen von Infrarotspektren [22] und von Spektren im ultravioletten und sichtbaren Spektralbereich befinden sich im Aufbau. Auf die Benutzung der kommerziell in maschinenlesbarer Form erhältlichen, von der American Society for Testing and Materials herausgegebenen Sammlung von etwa 130000 Infrarotspektren wurde bewusst verzichtet, da diese Sammlung die spektroskopischen Daten nur in arg verstümmelter Form enthält.

Für den Vergleich werden die Spektren durch solche Grössen beschrieben, die den im Abschnitt 5 genannten

Bedingungen genügen. Detaillierte Beschreibungen der verwendeten Merkmale finden sich in [17–19]. Die Umformung der spektroskopischen Daten kann bei kompliziert definierten Vergleichsmerkmalen zu einem relativ aufwendigen Prozess werden. Für den Betriebsaufwand eines Suchsystems fällt er aber trotzdem kaum ins Gewicht, wie dies die folgenden Überlegungen zeigen. Die Verschlüsselung der Spektrenbibliothek ist zwar aufwendig, braucht aber für eine gegebene Sammlung nur ein einziges Mal durchgeführt zu werden. Im laufenden Betrieb müssen die spektroskopischen Daten der jeweiligen Unbekannten in gleicher Weise codiert werden, es ergibt sich also eine Codierungsoperation pro Anfrage. Hingegen müssen für jede Anfrage sämtliche Referenzspektren mit dem Spektrum der jeweiligen Unbekannten verglichen werden. Es sind also pro Anfrage so viele Vergleichsoperationen notwendig, als Referenzspektren in der Bibliothek vorhanden sind. Schon bei einer Referenzspektrensammlung bescheidener Grösse wird also der Betriebsaufwand durch die Spektrenvergleiche dominiert werden. Es ist demzufolge lohnend, ohne Rücksicht auf den Codierungsaufwand die aus den spektroskopischen Rohdaten herausgezogenen Merkmale kompromisslos auf raschesten Vergleich zu optimieren. Einen weiteren signifikanten Beitrag zum Gesamtaufwand bilden die Kosten für die Speicherung der Spektrenbibliothek. Um diese zu minimalisieren, wird man für die codierten Spektren eine möglichst kompakte computerinterne Darstellung wählen. Um diese beiden Forderungen in optimaler Weise zu erfüllen, werden im OCETH-System grundsätzlich binäre Merkmale verwendet, also solche, die nur zwei einander ausschliessende Werte annehmen können. Solche Merkmale können computerintern in einem Bit verschlüsselt und mit einfachen logischen Operationen verarbeitet werden, sie sind also der logischen Struktur heutiger Digitalcomputer optimal angepasst.

Beim Aufbau des Systems werden die Spektren sämtlicher Referenzverbindungen verschlüsselt. Das Resultat der Codierung eines Spektrums bezeichnen wir als die Signatur der entsprechenden Verbindung. Sie besteht aus einer geordneten Kette von Nullen und Einsen. Jeder Stelle in dieser Kette ist ein Merkmal fest zugeordnet, wobei der Zahlenwert einer Stelle den Zustand des betreffenden Merkmals in den Spektren der Verbindung angibt. Die Länge der Signatur entspricht der Anzahl der zur Beschreibung der Spektren verwendeten Merkmale. Sie liegt beim beschriebenen System in der Grössenordnung von 100 pro spektroskopische Methode.

Die Berechnung der Signaturen erfolgt durch ein Computerprogramm aufgrund der vorgegebenen Definitionen der Merkmale, wobei als Eingangsgrössen die vollständigen spektroskopischen Daten der Referenzverbindungen verwendet werden. Die Signaturen werden dann in einer Signaturenbibliothek gesammelt. Der Aufbau dieser Signaturenbibliothek wird nur ein einziges Mal beim Aufbau des Systems durchgeführt. Bei der

Bearbeitung einer Anfrage laufen folgende Schritte ab. Als erstes werden dem System die vollständigen spektroskopischen Daten der zu bearbeitenden Unbekannten eingegeben, die in die Signatur der Unbekannten umgewandelt werden. Dazu dient das gleiche Programm, mit dem die Signaturenbibliothek aufgebaut wurde. Dann wird die Signatur der Unbekannten der Reihe nach mit den Signaturen aller Referenzverbindungen verglichen, wobei für jedes Signaturenpaar der Grad der Übereinstimmung berechnet wird. Dazu ist jedem Signaturelement und damit dem entsprechenden Merkmal ein Satz von Gewichten zugeordnet, der den relativen Wert dieses Merkmals für die Bestimmung der Ähnlichkeit zweier Strukturen angibt. Da der Vergleich für ein einzelnes Signaturelement zu vier verschiedenen Resultaten führen kann, sind pro Merkmal jeweils vier verschiedene Gewichte vorhanden. Die Bewertung der Ähnlichkeit zweier Signaturen erfolgt nun in der Weise, dass jeweils entsprechende Signaturelemente aus der Signatur der Unbekannten und aus der Signatur der Referenzverbindung miteinander verglichen werden, wobei die dem jeweiligen Resultat eines solchen Elementarvergleichs entsprechenden Gewichte aufsummiert werden. Die erhaltene Summe wird dann bezüglich jenes Wertes normiert, der beim Vergleich der Signatur der Unbekannten mit sich selbst erhalten würde. Dieser normierte Wert wird direkt als Ähnlichkeitsmass verwendet. Das System führt dabei eine Liste der Referenzverbindungen mit den höchsten Werten für das Ähnlichkeitsmass. Nach jedem Vergleich eines Signaturenpaars wird geprüft, ob das Resultat besser ist als das an letzter Stelle in der Liste verzeichnete. Ist dies nicht der Fall, so wird die gegenwärtig bearbeitete Referenz sofort verworfen. Andernfalls wird sie an der ihr zukommenden Stelle in die Liste eingefügt und verdrängt die an letzter Stelle stehende Referenz. Nach Verarbeitung der gesamten Signaturenbibliothek enthält diese Liste die der Unbekannten ähnlichsten Referenzverbindungen in geordneter Sequenz.

Im Regelfall sind für eine gegebene Unbekannte offensichtlich nur wenige potentiell brauchbare Referenzverbindungen in der Sammlung vorhanden. Der Hauptteil der Rechenzeit wird also dazu verwendet, um unbrauchbare Referenzverbindungen zu identifizieren. Es ist aber völlig uninteressant zu wissen, wie schlecht eine ungeeignete Referenzverbindung ist. Wenn die Ähnlichkeit ein gewisses Mass unterschreitet, so ist die betreffende Referenzverbindung mit Sicherheit für das gegebene Problem uninteressant. Man kann daher den Vergleich einer Referenzsignatur mit der Signatur der Unbekannten jeweils dann vorzeitig abbrechen, wenn diese Schwelle überschritten wird. Damit wird ein beträchtlicher Teil der Rechenzeit eingespart. Im beschriebenen System wird dazu folgendermassen vorgegangen. Vorerst wird für die gegebene Unbekannte der maximal erreichbare Wert des Ähnlichkeitsmasses errechnet, indem man die Signatur der Unbekannten mit sich selbst vergleicht. Ein vorbestimmter Teil dieses

Wertes, im Regelfall in der Grössenordnung von einigen Prozenten, dient nun als Schwellenwert für die maximal zulässigen Verluste, die eine Referenzverbindung beim Vergleich erleiden darf, ohne dass sie vorzeitig verworfen wird. Der Entscheid darüber, ob der Vergleich zu Ende geführt wird, wird nun sehr bald nach Beginn der Vergleichsoperation getroffen, üblicherweise nachdem etwa 10 bis 20% der Signaturelemente einer Referenzverbindung verglichen wurden. Wenn also eine Referenzverbindung nach dem Vergleich der ersten paar Merkmale schon mehr Verluste hinnehmen musste, als dem Schwellenwert entspricht, so darf man vermuten, dass sie kaum einen brauchbaren Beitrag zur Lösung des Problems liefern wird. Der Signaturvergleich wird dementsprechend sofort abgebrochen und die betreffende Referenzverbindung ohne weitere Prüfung verworfen. Man nimmt dabei einerseits in Kauf, dass in seltenen Fällen auch eine solche Referenzverbindung vorzeitig ausgeschieden wird, die im weiteren Verlauf des Signaturvergleichs keine zusätzlichen Verluste mehr erlitten und so ein vielleicht noch einigermaßen brauchbares Resultat erreicht hätte. Es kann aber keine Referenzverbindung an diesem Test scheitern, die im Endresultat weniger Fehler aufweist, als dem Schwellenwert entspricht. Demzufolge ist mit absoluter Sicherheit gewährleistet, dass keine Referenz mit hoher Ähnlichkeit verloren geht. Andererseits ist es auch möglich, dass solche Referenzen den Test bestehen, die im weiteren Verlauf kaum mehr übereinstimmende Signaturelemente aufweisen, also völlig unbrauchbar sind. Um ihre Zahl möglichst gering zu halten ist es notwendig, jene Signaturelemente vor dem Test zu vergleichen, die bei Nichtübereinstimmung zu den absolut grössten Verlusten führen. Dementsprechend werden die Signaturelemente nicht in der Reihenfolge verglichen, wie sie in der Signatur aufgeführt sind, sondern in einer aufgrund einer Prioritätsordnung gegebenen Sequenz. Die Priorität eines Signaturelementes wird dabei auf folgende Weise festgelegt. An sich können beim Vergleich zweier Merkmale vier verschiedene Resultate erhalten werden. Das Signaturelement in der Referenzverbindung kann mit dem entsprechenden Signaturelement der Unbekannten übereinstimmen oder nicht übereinstimmen, und das Signaturelement der Unbekannten kann den Wert Null oder Eins aufweisen. Für eine gegebene Unbekannte sind aber jeweils nur zwei dieser vier Resultate möglich, da die Werte seiner Signaturelemente gegeben sind. Die Differenz zwischen den diesen beiden Resultaten zugeordneten Gewichten entspricht dem Verlust, den eine Referenzverbindung bei Nichtübereinstimmen des entsprechenden Signaturelementes erleidet. Dementsprechend ergibt sich die Prioritätsordnung aus der Grösse dieser Differenzen. Sie hängt bei einem gegebenen Satz von Gewichten von der Signatur der Unbekannten ab, das System passt sich also automatisch den spektroskopischen Eigenschaften der jeweiligen Unbekannten an. Die Einsparungen an Rechenzeit, die sich mit die-

sem Verfahren realisieren lassen, sind ganz beträchtlich, indem bei optimaler Wahl der Parameter die Vergleichszeit bis auf gegen 10% zurückgehen kann [17]. Für die Qualität der Resultate eines Suchsystems der beschriebenen Art sind im wesentlichen drei Faktoren massgebend. In erster Linie ist es die Reichhaltigkeit der Spektrensammlung, die die Leistungsfähigkeit einschränkt, da nur in der Referenzsammlung vertretene Strukturen ausgegeben werden können. Ist keine brauchbare Referenzstruktur vorhanden, so ist das Resultat unabhängig von allen anderen Einflussgrössen wertlos. Innerhalb der durch diese Tatsache gegebenen Grenzen sind die Art der zum Vergleich herangezogenen spektroskopischen Merkmale und ihre Bewertung entscheidend. Beim heutigen, noch recht beschränkten Stand des Wissens erfolgt ihre Festlegung notwendigerweise empirisch. Eine theoretisch wohl fundierte Basis für die getroffene Auswahl kann dementsprechend nicht gegeben werden, ihre einzige Rechtfertigung sind praktisch brauchbare Resultate. Immerhin sollen hier einige der bei der Auswahl relevanten Gedankengänge kurz dargestellt werden. Was die Auswahl der spektroskopischen Merkmale anbetrifft, sei hier auf die zu Beginn dieses Abschnittes dargelegten Grundsätze verwiesen. Bei der Festsetzung der den Merkmalen zugeordneten Gewichte sind zwei Gesichtspunkte zu berücksichtigen. Die Gewichte sollen einmal ein der Denkweise des Chemikers angepasstes Ähnlichkeitsmass für Strukturen ermöglichen und andererseits für eine optimale Suchstrategie sorgen. Diese beiden Effekte gehen nun aber nicht unbedingt parallel. Hinsichtlich des Ähnlichkeitsmasses sind einzig und allein chemische Gesichtspunkte von Bedeutung, unabhängig davon wie die Spektrenbibliothek zusammengesetzt ist. Für die Optimierung der Suchstrategie dagegen spielt die Zusammensetzung der Referenzsammlung eine entscheidende Rolle. Wichtigstes Element einer optimalen Suchstrategie ist der rasche und möglichst vollständige Ausschluss ungeeigneter Referenzverbindung und nicht die differenzierte Bewertung der besten Referenzen. Mit dem hier beschriebenen Verfahren hat man es nun in der Hand, beide Aspekte einigermaßen unabhängig voneinander zu beeinflussen. Die Prioritätsordnung der Merkmale, die die Suchstrategie im wesentlichen festlegt, ergibt sich aus der Differenz von Gewichtspaaren. Der entsprechende Beitrag zum Ähnlichkeitsmass dagegen wird vorwiegend durch den Quotienten dieser beiden Gewichte bestimmt.

Das hier beschriebene Verfahren zum automatischen Vergleich von Spektren eignet sich hervorragend zur Implementation auf Computern bescheidener Grösse, da die notwendigen Programme sehr leicht in Segmenten verarbeitet werden können. Ein erstes Segment dient dazu, die spektroskopischen Daten der Unbekannten einzulesen und in die Signatur umzuwandeln. Im nächsten Segment, das den Platz des nun nicht mehr benötigten Signatur-Erzeugungsprogramms ein-

nimmt, werden die Signaturen der Referenzverbindungen sequentiell eingelesen und mit den Signaturen der Unbekannten verglichen. Das letzte Segment, das an Stelle des Vergleichsprogramms geladen wird, sorgt für eine saubere und übersichtlich dargestellte Ausgabe der Resultate. Diese Segmentierung erlaubt es, mit recht bescheidenem Kernspeicherbedarf auszukommen. Für die Speicherung der Signaturenbibliothek muss ein Massenspeicher vorhanden sein. Da die gespeicherten Signaturen aber strikte sequentiell gelesen und verarbeitet werden, sind auch hier von der Hardware-Seite her nur bescheidene Voraussetzungen zu erfüllen.

Ein weiteres wesentliches Element des beschriebenen Algorithmus besteht darin, dass alle auf eine bestimmte spektroskopische Methode bezogenen Teile im ersten Segment zusammengezogen sind, in dem die spektroskopischen Daten in Signaturen umgewandelt werden. Das Vergleichs-Segment hingegen ist in keiner Weise davon beeinflusst, ob Kernresonanz-, Massen- oder Infrarotspektren verglichen werden. Dementsprechend ist für den Einschluss einer anderen spektroskopischen Methode nur die Anpassung des ersten Segmentes notwendig. Ebenso ist es vom Programmablauf her problemlos, die Daten mehrerer spektroskopischer Methoden gleichzeitig zu verarbeiten. Dazu werden einfach die Teilsignaturen der einzelnen Methoden zu einer Totalsignatur vereinigt.

Eine objektive Bewertung des Suchsystems hinsichtlich der Qualität der erhaltenen Resultate ist sehr schwierig, wenn nicht gar unmöglich. Es wäre dazu ein objektives Mass zur Bewertung der strukturellen Ähnlichkeit zweier Verbindungen notwendig. Die Bedeutung der Relation «ähnlich» kann aber bestenfalls nur im Hinblick auf eine gegebene konkrete Fragestellung angegeben werden, so dass sie notwendigerweise in hohem Mass problembezogen ist. Die Beurteilung der Qualität eines Suchsystems muss sich demzufolge heute darauf abstützen, ob es im praktischen Einsatz bei wirtschaftlich vertretbarem Aufwand brauchbare und nützliche Resultate erbringt. Zuverlässige Angaben sind also erst nach einer längeren Einsatzperiode erhältlich. Für eine rasche und überschlagsmässige Vorevaluation geht man in der Regel so vor, dass man aus der Referenzsammlung einzelne Verbindungen auswählt und als «Unbekannte» behandelt, wobei die spektroskopischen Daten gegebenenfalls künstlich vertauscht werden, um die nichtideale Reproduzierbarkeit zu simulieren. Für die Bewertung der Resultate einer solchen Voruntersuchung ist es nun wesentlich, dass nicht nur darauf geachtet wird, dass die «Unbekannte» wieder aufgefunden wird und welchen Rang sie erreicht. Von unvergleichlich grösserer Bedeutung ist die Frage, ob diejenigen ausgegebenen Referenzverbindungen, die nicht mit der «Unbekannten» identisch sind, leicht zu erfassende Hinweise auf die Struktur und den Typus der «Unbekannten» erkennen lassen. Bei einem eventuell wenig befriedigenden Einzelresultat ist ausserdem

abzuklären, ob überhaupt sinnvolle Referenzverbindungen in der Sammlung vorhanden sind. Ist dies nicht der Fall, so beweist das Resultat lediglich das Unge-nügen der Spektrenbibliothek, sagt aber über das ver-wendete Vergleichs- und Bewertungsverfahren nichts aus. Die heute frei zugänglichen computerlesbaren Spektrensammlungen sind in ihrem Umfang recht be-scheiden und bezüglich der in ihnen vertretenen Ver-bindungstypen stark unausgewogen. Die Grenzen der Leistungsfähigkeit von Systemen zur computerunter-stützten Spektreninterpretation durch Vergleich mit Referenzdaten sind daher heute vorwiegend durch die ungenügende Datenbasis bestimmt.

Bei der Strukturaufklärung organischer Verbindungen ist ein auch nur einigermaßen erfahrener Analytiker der Maschine weit überlegen, wenn es um die Lösung relativ komplexer Fragestellungen geht. Die Maschine hat dem Menschen aber voraus, dass sie einerseits extrem schnell und zuverlässig und andererseits völlig unvoreingenommen arbeitet. Dementsprechend er-geben sich für die computerunterstützte Spektreninter-pretation durch Vergleich mit Referenzdaten heute zwei Haupteinsatzgebiete. Einmal ist die Maschine auf-grund ihrer Schnelligkeit hervorragend geeignet, um aus einer grossen Zahl von Spektren diejenigen heraus-zusortieren, die bekannten, in einer Referenzsammlung vorhandenen Substanzen entsprechen. Problemstellun-gen dieser Art ergeben sich beispielsweise beim Nach-weis von Pestizidrückständen mit Hilfe von GC/MS-Methoden. Auch in anderen Bereichen der Umwelt-analytik, in der klinischen Chemie und gelegentlich auch in der Forensik sind analoge Situationen recht häufig. Der zweite, weniger deutlich ausgeprägte Schwerpunkt des Anwendungsbereichs liegt bei der Bearbeitung von solchen Proben, für die neben den spektroskopischen Daten widersprüchliche und ver-wirrende oder auch überhaupt keine zusätzliche Infor-mation vorliegt. Hier liegt der Vorteil der Maschine gegenüber dem Menschen darin, dass sie jede, auch die abwegigste Möglichkeit mit der gleichen Sorgfalt über-prüft, während der Mensch im allgemeinen Probleme dieser Art mit einer im Unbewussten bereits vorgefas-sen Meinung angeht. Zwar wird die Maschine nur in seltenen Glücksfällen direkt eine brauchbare Lösung erbringen. Ihre Vorschläge sind aber meist vorzüglich geeignet, den Analytiker auf von ihm nicht in Betracht gezogene Nebenlösungen hinzuweisen und so seine Arbeit wesentlich zu erleichtern und zu beschleunigen. Eine technische Begrenzung des Einsatzes computer-unterstützter Methoden zur Spektreninterpretation liegt darin, dass sowohl die Referenzdaten wie auch die spektroskopischen Daten für die zu untersuchende Probe in maschinenlesbarer Form vorliegen müssen. Die nachträgliche Digitalisierung analog registrierter Kurvenzüge ist beim Aufbau von Referenzspektren-sammlungen unter Umständen vertretbar, da sich we-gen der wiederholten Benützung dieser Datensätze ein gewisser Aufwand durchaus lohnen kann. Die nach-

trägliche Digitalisierung der Spektren der unbekannten Probe ist hingegen wirtschaftlich kaum je tragbar. Dementsprechend ist der Einsatz automatischer Inter-pretationsmethoden heute im allgemeinen erst dann sinnvoll, wenn die spektroskopischen Rohdaten ohne zusätzlichen Aufwand direkt in digitaler Form erhalten werden können. Diese Möglichkeit ist zum gegenwärtigen Zeitpunkt in analytischen Laboratorien noch nicht allgemein gegeben. Die rasante technische Entwicklung im Bereich der Elektronik hat nun aber insbesondere bei den Mikroprozessoren nicht nur zu wesentlich ge-steigelter Leistungsfähigkeit, sondern auch zu einem dramatischen Preiszusammenbruch geführt. Damit dürften die wichtigsten technischen und wirtschaftli-chen Barrieren für die Konstruktion voll digitalisierter analytischer Instrumente zur Verwendung im Routine-labor gefallen sein. Entsprechende Geräte werden seit kurzem bereits auf dem Markt angeboten. In dem Mass, wie sich solche dem modernsten Stand der Tech-nik entsprechend voll digitalisierte Instrumente die Routine-Analytik erobern, wird auch der Bedarf an raffinierten Systemen zur computerunterstützten Spek-treninterpretation ansteigen. Es ist dementsprechend eine der Aufgaben der Forschung im Bereich der Struk-turanalytik organischer Verbindungen, diejenigen Grundlagen zu erarbeiten und dem Praktiker zur Ver-fügung zu stellen, die zu diesem nicht mehr fernen Zeit-punkt benötigt werden.

Verdankung

Der Autor möchte dem Schweizerischen Nationalfonds zur För-derung der wissenschaftlichen Forschung sowie der Kommission der Europäischen Gemeinschaften, Aktion COST 64b, für die grosszügige Unterstützung der vorliegenden Arbeit danken.

Literaturverzeichnis

- 1 F.A. Mellon: in Mass Spectrometry, Specialist Periodical Report. R.A.W. Johnston, Ed. The Chemical Society, London 1975, p.117.
- 2 E. Ziegler: Computer in der instrumentellen Analytik. Akad. Verlagsges. Frankfurt/Main 1973.
- 3 D. Ziessow: On-line Rechner in der Chemie. W.deGruyter, Berlin 1973.
- 4 R.M. Silverstein, G.C. Bassler und T.C. Morill: Spectro-metric Identification of Organic Compounds. John Wiley & Sons, New York 1974.
- 5 B.G. Buchanan, D.M. Smith, W.C. White, R.J. Gritter, E.A. Feigenbaum, J.Lederberg und C.Djerassi: J.Amer.Chem. Soc. 98 (1976) 6168.
- 6 V.A. Dement'ev: Zh. Prikl. Spektrosk. 23 (1975) 479.
- 7 R. Carhart und C.Djerassi: J.Chem.Soc.Perkin II (1973) 1753.
- 8 L.A. Gribov, V.A. Dement'ev, M.E. Elyashberg und E.Z. Ya-kupov: J. Molecular Structure 22 (1974) 161.
- 9 W.T. Wipke, S.R. Heller, R.J. Feldmann und E. Hyde, Ed. Computer Representation and Manipulation of Chemical Information. John Wiley & Sons, New York 1974.
- 10 T.F. Lam, C.L. Wilkins, T.R. Brunner, L.J. Soltzberg und S.L. Kaberline: Anal.Chem. 48 (1976) 1768.
- 11 P.C. Jurs und T.L. Isenhour: Chemical Applications of Pattern Recognition. John Wiley & Sons, New York 1975.

- 12 *N.J. Nilsson*: Learning Machines. McGraw Hill, New York 1965.
- 13 *B.S. Duran* und *P.L. Odell*: Cluster Analysis. Springer, Heidelberg 1974.
- 14 *H.H. Bock*: Automatische Klassifikation. Vandenhoeck & Ruprecht, Göttingen 1974.
- 15 *J. Meili*: Diss. ETH Zürich Nr. 5521 (1975).
- 16 *J.T. Clerc* und *J. Meili*: Chromatographia 7 (1974) 515.
- 17 *P.R. Nägeli* und *J.T. Clerc*: Anal. Chem. 46 (1974) 739 A.
- 18 *P. Nägeli*: Diss. ETH Zürich Nr. 5509 (1975).
- 19 *R. Schwarzenbach*, *J. Meili*, *H. Köntzer* und *J.T. Clerc*: Org. Mag. Res. 8 (1976) 11.
- 20 *J.T. Clerc*: Computerunterstützte Erfassung, Dokumentation und Interpretation spektroskopischer Daten für die Strukturaufklärung organischer Verbindungen. Habilitationsschrift ETH Zürich, 1975.
- 21 Mass Spectral Data Collection, UKAEA, Aldermaston, Reading, UK.
- 22 *J.T. Clerc*, *R. Knutti*, *H. Köntzer* und *J. Zupan*: Z. Anal. Chem. 283 (1977) 177.