TU Bergakademie Freiberg
Fachbereich Chemie
Institut für Analytische Chemie

Methoden zur Bibliothekssuche im IR-Bereich

Diplomarbeit
von
Hans-Martin Klötzer

Betreuer: Prof. Dr. rer. nat. habil. M. Otto
eingereicht am 11.7.1992

An erster Stelle möchte ich mich bei meinem Betreuer, Herrn Prof. Dr. rer. nat. habil. Otto für die interessante Aufgabenstellung und seine wertvollen Hinweise bedanken.
Mein Dank gilt weiterhin Herrn Dathe für die Unterstützung beim Einsatz Neuronaler Netze. Auch möchte ich Frau Wemme und Frau Tesch für die Aufnahme der Spektren danken.

Erklärung

Hiermit erkläre ich, daß ich die vorliegende Arbeit selbständig, ohne fremde Hilfe und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Freiberg, den 11.7.1992

Inhalt

Einführung und Aufgabenstellung

Die Aufklärung der Struktur chemischer Verbindungen ist eine der Hauptaufgaben der instrumentellen Analytik. Diese Meßverfahren haben durch die Entwicklung der Computertechnik entscheidende Impulse zu ihrer Perfektionierung erhalten. Im besonderen Maße trifft dies auf die spektroskopischen Methoden zu. Heute ist es Stand der Technik, Spektrometer mit Rechnern zu koppeln und somit aufwendige mathematische Verfahren, z.B. FFT[1], zur Auswertung nutzen zu können.

Es gab bereits sehr zeitig erste Versuche, die Vorzüge der Rechentechnik mit denen der Analysentechnik zu verbinden. So wurden bereits 1951 durch Kuenzel [1] Versuche beschrieben, eine Spektrendatenbank aufzubauen. Während in dieser Zeit oft die Aufgabe stand, die Meßwerte den Möglichkeiten der Hardware anzupassen, sind heute von dieser Seite kaum noch Schranken gesetzt. Die zum Betreiben einer solchen Datenbank notwendigen Werkzeuge müssen diesem Umstand Rechnung tragen. Waren in der Anfangszeit der Datenbankentwicklung Tools zur Datenreduktion, z.B. Generierung von Linientabellen aus Vollspektren, erforderlich, sind es zum heutigen Zeitpunkt Algorithmen zum Wiederfinden bereits gespeicherter Daten. Daraus leitet sich eine Gruppe von Suchalgorithmen ab. Es sind dies die Methoden zur Identitätssuche. Das Ziel einer Datenbankabfrage kann jedoch auch darin bestehen, alle ähnlichen Informationen, z.B. Spektren oder Strukturen, zu finden.

Seit einiger Zeit erfreuen sich Methoden, bei denen hochleistungsfähige Trennverfahren mit ebensolchen Detektionsverfahren (GC-IR, LC-IR) gekoppelt werden, einer steigenden Verbreitung. Die dabei anfallenden Datenmengen sind enorm. Es kommt daher darauf an, die bereits einmal identifizierten Stoffe nicht ein zweites Mal einer detaillierten Strukturaufkärung zu unterwerfen. Die IR-Spektroskopie ist sehr gut geeignet, die Identität einer Verbindung zu ermitteln. Deshalb bezeichnet man sie auch gern als eine "Fingerprint-Methode". Es ist möglich, über die IR-Spektren bereits in einer Datenbank gespeicherte Informationen zu dieser Substanz abzufragen. Dadurch kann der Einsatz weiterer strukturaufklärender Methoden, z.B. MS oder NMR, auf die nicht identifizierbaren Substanzen begrenzt werden. Da die IR-Spektroskopie eine schnell und preiswert durchzuführende Analysenmethode ist, ist dies ein nicht zu unterschätzender Vorteil.

Ziel meiner Arbeit ist es, verschiedene Suchverfahren zur Ähnlichkeitssuche miteinander zu vergleichen und Methoden zur Klassifizierung von Spektren zu erproben. Hierbei sollten verschieden kodierte Spektren (Voll-, Linien- und

---

1 Ein Abkürzungsverzeichnis befindet sich am Ende der Arbeit

Strichspektren) zu vergleichbaren Resultaten führen. Weiterhin galt es, Aussagen zur Sicherheit und Reproduzierbarkeit solcher Datenbankabfragen zu treffen. Es sollte versucht werden, die Probleme bei der Indizierung einer Datenbank für spektroskopische Daten mit Mitteln der Mustererkennung zu verringern. Dabei ist an den Einsatz spezieller Neuronaler Netze gedacht worden. Durch diese Methoden könnten sehr zeitintensive Suchvorgänge auf einen bereits eingegrenzten Datenbereich beschränkt werden.

1. Grundlagen

1.1. IR-Spektroskopie

1.1.1. Physikalische Grundlagen

Von den vielfältigen Wechselwirkungen elektromagnetischer Strahlung mit Materie sind für den Chemiker diejenigen von besonderem Interesse, welche Aussagen über den Aufbau einer chemischen Verbindung zulassen. Mit der Infrarotspektroskopie kann man besonders gut die Struktur organischer Substanzen aufklären. In einem solchen IR-Spektrum sind jedoch so viele Informationen enthalten, daß eine rein theoretische Auswertung zum gegenwärtigen Zeitpunkt noch nicht möglich ist. Dennoch sind für das Verständnis einige Grundlagen [2,3] unerläßlich.
Trifft Strahlung auf eine zu untersuchende Substanz, so treten Reflexion (R), Streuung (S), Transmission und Absorption (A) auf. Es gilt somit:

$$I_0 = I_R + I_S + I + I_A \,. \qquad (1)$$

Der Parameter I_R wird in Größe und Richtung allein durch die makroskopischen Eigenschaften bestimmt. I_S hängt vom Verhältnis der Teilchendurchmesser zur Wellenlänge des Lichtes ab. Der durch die chemische Verbindung absorbierte Strahlungsanteil spiegelt sich in I_A wieder.
Die Energie des absorbierten Lichtes führt zur Anregung spezifischer Energiezustände der Substanz. Die Größe der Anregung unterliegt den Gesetzen der Quantenmechanik.

$$E = h * \tau \qquad (2)$$

Diese besagen, daß ein Molekül nicht beliebige, sondern nur ganz bestimmte Energiezustände einnehmen kann. Dadurch können nur diskrete Energiewerte aufgenommen bzw. abgegeben werden.
Die durch IR übertragbare Energie führt bevorzugt zu Molekülschwingungen (Bild 1). Ein Molekül aus N Atomen besitzt 3N Freiheitsgrade

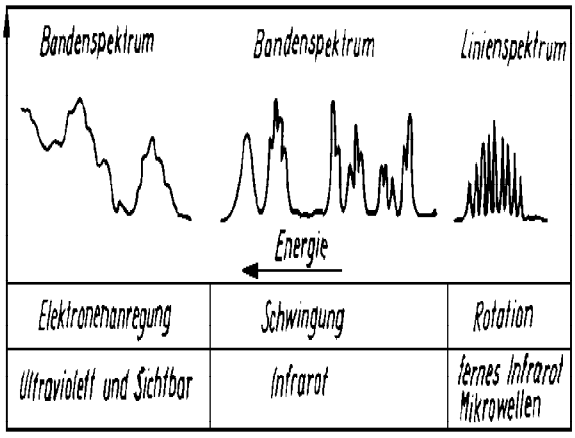


Bild 1 Zusammenhang Energie Spektrenart [2]

der Bewegung. Davon entfallen drei auf die Translation, die restlichen 3N-3 verteilen sich auf Molekülrotation und -schwingung. Die Zahl der Rotations- und Schwingungsfreiheitsgrade ergibt sich aus dem räumlichen Aufbau der Moleküle. Lineare Moleküle besitzen 3N-5, nicht lineare 3N-6 Freiheitsgrade. Als Normalschwingungen bezeichnet man diejenigen,die diesen Formeln folgen. Je nachdem, ob sich die Kernabstände oder die Bindungswinkel ändern, werden sie als Valenzschwingungen (τ) bzw. als Deformationsschwingungen (δ) bezeichnet. Erstere können sowohl symmetrisch, (μ_s) als auch antisymmetrisch (μ_{as}) erfolgen.
Neben den Eigenschwingungen treten auch noch Oberschwingungen und Kombinationsschwingungen in den Spektren auf. Die Intensität dieser Banden ist sehr klein im Vergleich zu den Normalschwingungen.
Durch die Energie der infraroten Strahlung werden auch die energetisch tiefer liegenden Rotationsniveaus mit angeregt. In einem so entstandenen Spektrum sind folglich neben den Schwingungs- auch die Rotationsübergänge enthalten. Jeder Schwingungsübergang besteht demnach aus einer Folge dicht beieinander liegender Linien, die in der Regel nicht mehr aufgelöst werden können und zu Absorptionsbanden zusammenfallen.
Durch die Dehnung und Beugung von Bindungen verschieben sich die Ladungszentren im Molekül in ihrer Lage. Je nach Art der Bewegung der Molekülbausteine zueinander, kann sich daraus eine Dipolmomentsänderung ergeben. Jene ist eine weitere Voraussetzung für die Strahlungsabsorption: Das absorbierende Molekül muß ein elektrisches Dipolmoment μ besitzen, das durch Wechselwirkung mit elektromagnetischer Strahlung zur Änderung seiner Größe und/oder Richtung angeregt werden kann. Dies ist der Grund, weshalb nicht alle Eigenschwingungen im IR-Spektrum sichtbar sind.

$$\frac{\partial \mu}{\partial r} \neq 0 \quad \text{IR} - \text{aktiv}$$
$$\frac{\partial \mu}{\partial r} = 0 \quad \text{IR} - \text{inaktiv} \qquad (3)$$

1.1.2. Charakteristische Bereiche und Molekülaufbau

Die Vielzahl der organischen Verbindungen besteht aus wenigen Grundbausteinen. Dies spiegelt sich auch in ihren Infrarotspektren wieder. So entstehen Bereiche, in denen sowohl sehr viele Linien liegen, als auch sehr schwach besetzte Gebiete. Man erkennt es deutlich, wenn man sich ein Histogramm der Linien über viele Spektren ansieht. Daraus wird ersichtlich, daß es für den Vergleich Bereiche unterschiedlicher Wertigkeit in einem Spektrum gibt.

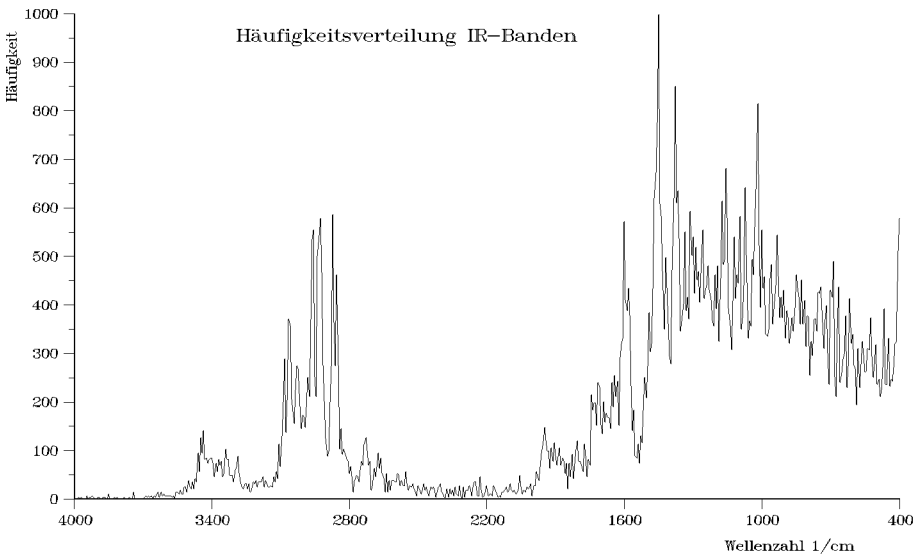


Bild 2 Häufigkeitsverteilung der IR-Banden ermittelt für 1000 Spektren

Wie man dem Bild 2 entnehmen kann, ist das Gebiet unterhalb etwa 1500 cm^{-1} besonders linienreich. Hier sind die Gerüstschwingungen, zahlreiche Deformationsschwingungen, jedoch auch Valenzschwingungen schwerer Atome zu finden. Für Aussagen über Teilstrukturen ist dieser als "Fingerprint" bezeichnete Bereich schwer zu nutzen, es können nur die intensivsten Signale (Ester, Ether, Alkohole) interpretiert werden. Er eignet sich jedoch hervorragend zur Identifikation einer chemischen Substanz.
Zwischen etwa 1500 und 1900 cm^{-1} sind die Valenzschwingungen der Doppelbindungen zu finden. Ihnen folgt der Valenzschwingungsbereich der Dreifachbindungen ca. 1900 - 2400 cm^{-1}. Im kurzwelligen Gebiet um 2500 bis 3700 cm^{-1} absorbieren die Valenzschwingungen des Wasserstoffs gegen Kohlenstoff.

500	1500	2000 2500	3000 3500 cm^{-1}
Deformations-schwingungen Gerüst-schwingungen Valenz-schwingungen schwerer Atome	Valenz-schwingungen A=B	Valenz-schwingungen A≡B	Valenz-schwingungen A-H
C-Cl ≈ 700 C-Br ≈ 550 C-I ≈ 500 C-Si ≈ 800 C-Pb ≈ 450	C=C ≈ 1650 C=N ≈ 1650 C=O ≈ 1700 P=O ≈ 1300 Cr=O ≈ 850	C≡C ≈ 2100 C≡N ≈ 2250	C-H ≈ 3000 O-H ≈ 3600 N-H ≈ 3500 P-H ≈ 2400 Sn-H ≈ 1800

Tabelle 1 Schlüsselfrequenzen in cm^{-1} für wichtige Strukturgruppen

Neben diesen Banden treten auch solche auf, die aus der Kopplung von Schwingungen benachbarter Strukturgruppen entstehen. Diese sind nicht mehr für eine Substruktur spezifisch, sondern für die Stellung der Gruppen zueinander. Am deutlichsten ist dies an verschieden substituierten Phenylverbindungen zu erkennen. Es ergibt sich für jeden Substitutionstyp ein charakteristisches Muster (Bild 3).

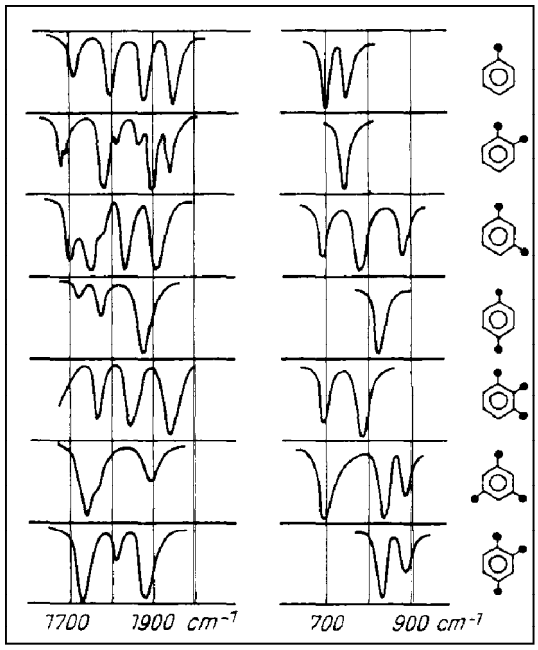


Bild 3 Typische Muster für verschieden substituierte Aromaten [3]

1.1.3. Spektreninterpretation

In diesem Abschnitt soll kurz auf die allgemeine Vorgehensweise bei der Interpretation eines Spektrums eingegangen werden. Das erscheint notwendig, da in einem solchen Prozeß die Bibliotheksdaten nach verschiedensten Gesichtspunkten mit denen der Probe verknüpft werden, was sowohl bei rechnergestützten Interpretationssystemen, als auch bei der manuellen Auswertung zu wiederholten Bibliothekssuchen führt, dem Inhalt dieser Arbeit.
Als erstes sollte man möglichst alle verfügbaren Daten aus der Vorgeschichte und die über die "menschlichen Sensoren" (Aussehen, Geruch) zugänglichen erfassen. Das kann sich bei der rechnergestützten Interpretation als sehr hilfreich erweisen, da der Computer alle Möglichkeiten prüft, auch völlig unsinnige, die ein erfahrener Spektroskopiker bereits über die Vorwerte ausschließen kann. Danach erfolgt die Überprüfung des Spektrums auf Störungen, wie sie durch CO_2, Wasser und Lösungsmittel hervorgerufen werden können. Eine nicht korrekt arbeitende Untergrundkompensation kann besonders bei elektronischen Suchalgorithmen zu fatalen Ergebnissen führen. Die anschließende Zuordnung von Banden zu Substrukturen liefert eine Menge von potentiell im Molekül enthaltenen Gruppen. Eine weitere wichtige Informationsquelle ist der Ausschluß von Linien. Damit kann eine Korrektur der im vorhergehenden Schritt erhaltenen Liste vorgenommen werden. Aus der so erhaltenen Menge von Molekülfragmenten müssen nun Strukturvorschläge erzeugt werden. Dabei fließen das chemische Wissen, das Wissen um die Vorgeschichte und begründete Vermutungen zusammen. Jetzt wird versucht, den Kandidaten die im Spektrum

enthaltene substanzspezifischen Informationen zuzuordnen. Bei Aromaten wären es unter anderem Aussagen zum Substitutionstyp. Möglicherweise kann man aus den charakteristischen Verschiebungen einzelner Signale die Verknüpfung von funktionellen Gruppen untereinander erkennen. Die nun noch vorhandenen Vorschläge für die Substanzen können in einem Spektrenatlas gesucht werden. Sollte dies zu einem sehr wahrscheinlichen Ergebnis führen, muß es z.B. durch die Zugabe der reinen Substanz zur Probe überprüft werden. Meistens muß eine zweite oder eine weitere Methode zu Rate gezogen werden.
In diesem ganzen Prozeß wird das vorhandene Datenmaterial, ob in Buchform oder als elektronisches Medium, mehrfach nach verschiedenen Gesichtspunkten wie Identität, Substituenten, ähnliches Gerüst, Substitutionstyp u.a. durchsucht.

1.1.4. Fehler bei der Ausbildung von IR-Spektren

Wenn man Betrachtungen zu Bibliothekssuchen anstellt, sollte man sich auch über die Beeinflussungen der Daten Klarheit verschaffen.

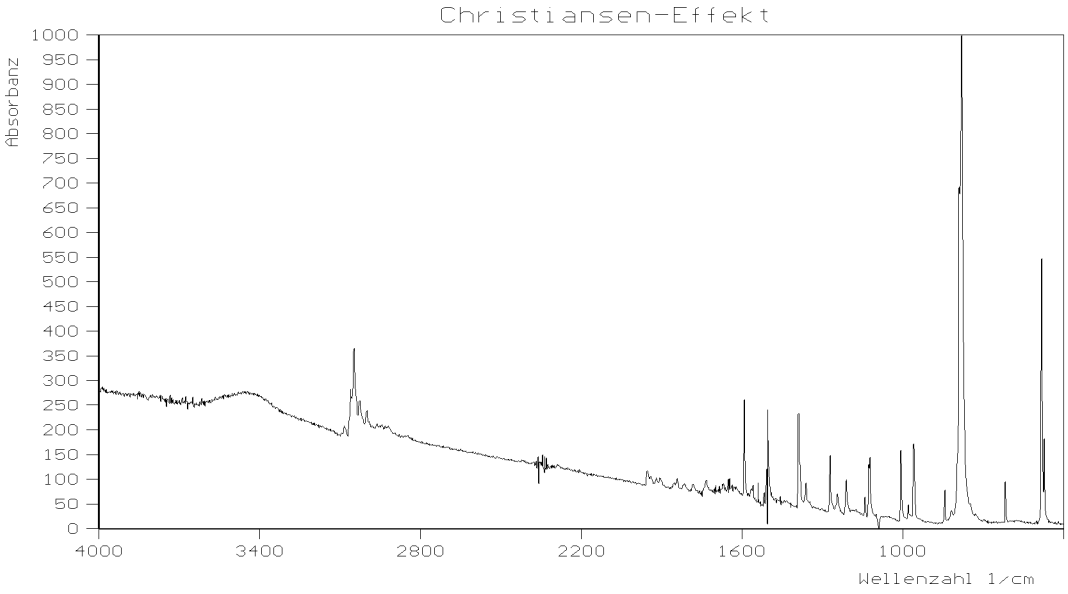


Bild 4 Durch Christiansen-Effekt gestörtes Spektrum von Naphthalin

Der Parameter I_S in Gleichung (1) beschreibt den Anteil der Strahlung, welche gestreut wird. Die Streuung des infraroten Lichtes hängt von der Wellenlänge und dem Teilchendurchmesser ab. Sie wird umso stärker, je größer die Partikel und je kürzer die Wellenlängen sind. Bei sehr harten, schlagzähen oder sehr weichen Substanzen kann es vorkommen, daß die Zerkleinerung innerhalb der üblichen Zeit unzureichend ausfällt. Dies manifestiert sich in einer zu hohen Wellenzahlen hin ansteigenden Grundlinie. Der Anteil I_S vergrößert sich mit

Zunahme der Wellenzahl. Eine große Differenz der Brechungsindices zwischen Probe und Einbettungsmittel führt auf Grund stark reflektierender innerer Grenzflächen zu dem im Bild 4 gut zu erkennenden Phänomen (Christiansen-Effekt).

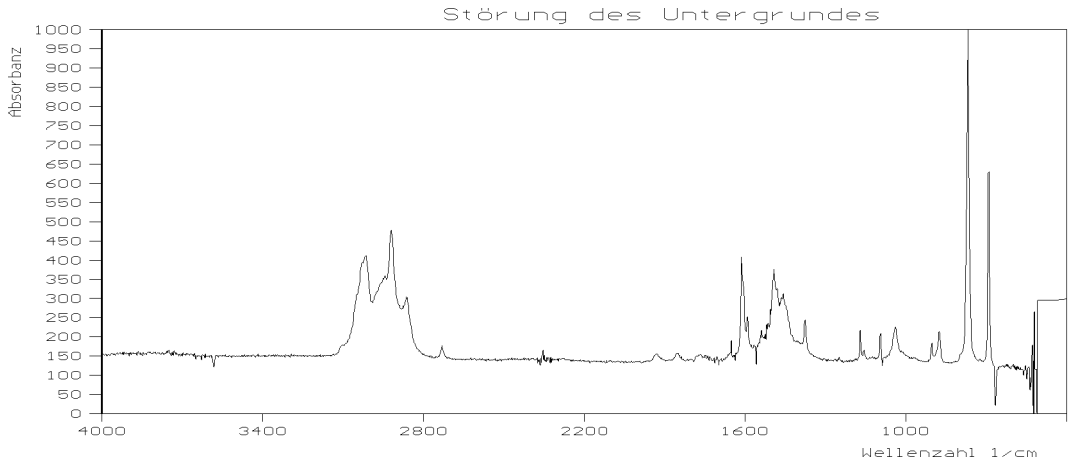


Bild 5 Störung der Grundlinie von meta-Xylen durch nicht kompensierbare Absorptionen des Küvettenmaterials

Durch die Vermischung der Proben mit einem Einbettungsmittel ergeben sich weitere Fehlerquellen. So sind Reaktionen der Probe mit z.B. KBr nicht immer auszuschließen. Weiterhin können Hydratationen durch die dem KBr anhaftende Feuchtigkeit auftreten. Verschiedene Einbettungsmaterialien weisen selbst Absorptionen im IR-Bereich auf. Diese können zwar bei Zweistrahlmessungen kompensiert werden, aber die Resultate lassen manchmal etwas zu wünschen übrig, wie Bild 5 zeigt. Gelangt ein solches Spektrum unkorrigiert in ein automatisches Auswertesystem, wird dies wahrscheinlich zu Fehlern führen.

Für flüssige Substanzen eignet sich die Einbettung nicht. Solche Stoffe müssen in Küvetten in den Strahlungsgang gebracht werden. Das üblicherweise genutzte Küvettenmaterial ist NaCl. Dadurch ergibt sich eine Einschränkung des langwelligen Bereiches des Spektrums im Vergleich zu KBr (Bild 6). Auf die ebenfalls verfügbaren KBr-Küvetten weicht man nur dann aus, wenn es der Spektralbereich unbedingt erforderlich macht.

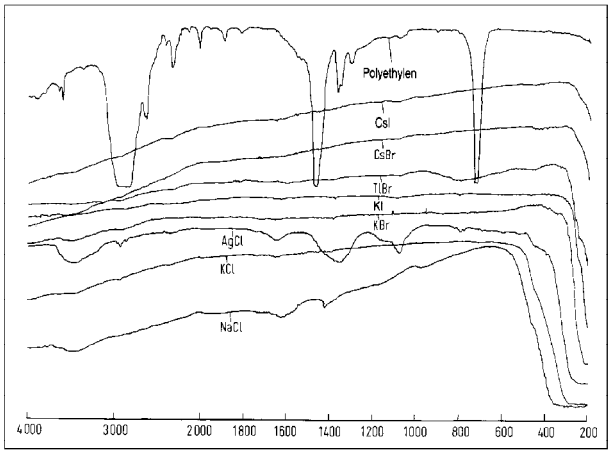


Bild 6 Spektren üblicher Einbettungsmittel [18]

Nicht immer ist es möglich, die Probe als Feststoff oder reine Flüssigkeit zu

präparieren. Wenn sich der Einsatz eines Lösungsmittels nicht vermeiden läßt, muß man die Beeinflussung der Probe und die störenden Absorptionen durch das Lösungsmittel berücksichtigen. Auf Grund der Wechselwirkungen können Banden in ihrer Lage verändert werden.

Einen sich zwar verkleinernden, aber immer noch merklichen Einfluß auf das Spektrum hat die Gerätetechnik. Die wellenlängendispersiven Geräte müssen mit Monochromatoren ausgerüstet werden. Sie benötigen einen scharf gebündelten parallelen Lichtstrahl. Um diesen zu erzielen, setzt man eine Spaltblende ein. Die Backen dieser Blende können gegeneinander verschoben werden. Je kleiner man die Öffnung wählt, umso besser ist die Auflösung des Gerätes.

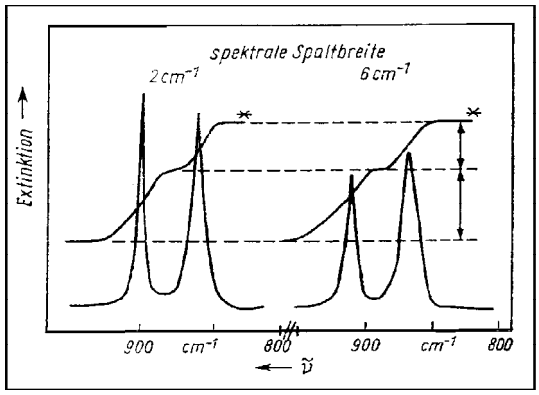


Bild 7 Einfluß der Spaltbreite auf die Bandenform [21]

Es gibt jedoch einen Zusammenhang zwischen Spaltbreite und Bandenform. Dieser wird in Abbildung 7 deutlich. Ihn gilt es zu berücksichtigen, wenn man Spektren unterschiedlicher Herkunft vergleichen will. Mit der zunehmenden Verbreitung von FT-IR-Geräten verringert sich diese Problematik, da in solchen Instrumenten keine optischen Systeme zur Strahlungszerlegung eingesetzt werden müssen (Bild 8).

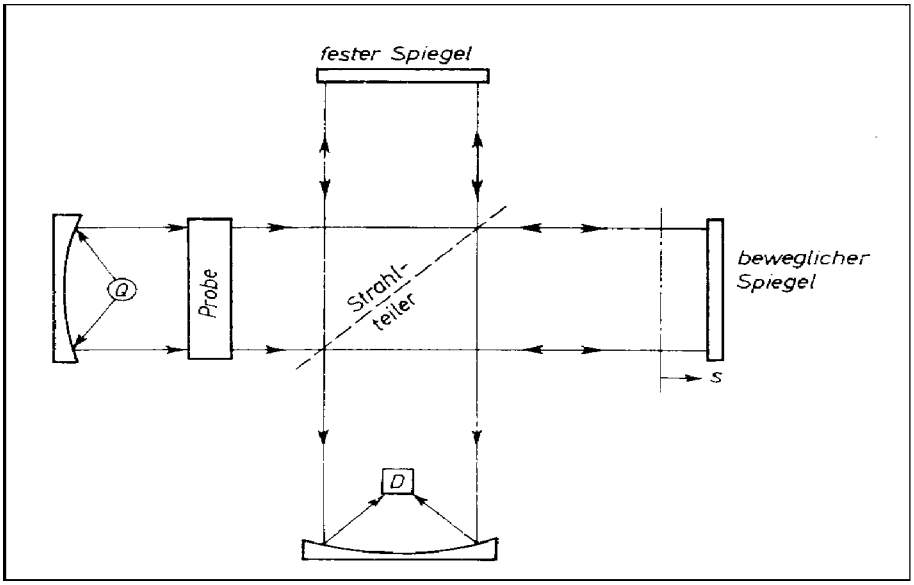


Bild 8 Prinzip eines FT-IR-Spektrometers in Michelson-Aufstellung [22]

Der verschiedenen hier aufgezeigten Fehlermöglichkeiten sollte man sich bei der Bewertung einer Suche in einer umfangreichen Bibliothek immer bewußt sein. Da solche Datensammlungen historisch entstanden sind, muß damit gerechnet werden, Daten aus verschiedenen gerätetechnischen Epochen und mit unterschiedlichen Qualitätsstandards anzutreffen.

1.2. Spektrendatenbanken

1.2.1 Datenbankeinsatz in der Chemie

Datenbanken sind Programmsysteme zur geordneten Ablage und Einsichtnahme beliebiger Sachverhalte in einem elektronischem Medium. Sie umfassen Werkzeuge zum Im- und Export von Informationen, zur Plausibilitätsprüfung und Verwaltung der Daten. Der Inhalt solcher Speicher ist sehr vielfältig. Neben Sammlungen rein numerischer Daten, dem ursprünglichen Einsatzgebiet, existieren Datenbanken, die hauptsächlich Texte enthalten. In einem neueren Ansatz bei der Entwicklung von Expertensystemen, den sog. "Fall-basierenden" Expertensystemen, benötigt man eine Datenbank für die bereits bearbeiteten Fälle. Es ließe sich hier nahezu jedes mit Rechnern in Verbindung stehende Gebiet aufzählen.
Die Chemie ist von dieser Vielfalt nicht ausgenommen. Der Einsatz reicht von Informationssystemen wie CAS-Online über Reaktionsdatenbanken bis zu Wissensbasen für Expertensysteme. In dieser Arbeit soll speziell auf die elektronischen Datensammlungen infraroter Spektren eingegangen werden.
Die Datensätze einer solchen Bibliothek entsprechen immer nachfolgendem Prinzip:

Headerblock | Datenblock

Dabei beinhaltet der Headerblock die Steuerinformationen und der Datenblock die eigentlichen Informationen. Genauer auf die Struktur der Daten einzugehen erscheint wenig sinnvoll, da jeder Anbieter von elektronischen Spektren-bibliotheken ein eigenes Format benutzt. Für den Austausch von Daten hat sich das JCAMP-DX Format [4] durchgesetzt. Dadurch können Daten verschiedener Systeme problemlos untereinander ausgetauscht werden.

1.2.2. Speicherung und Kodierung von Spektren

Durch die Senkung der Kosten pro Speicherplatz kommt diesem Punkt nicht mehr die gleiche Bedeutung zu, wie noch vor wenigen Jahren. Ein kleines Rechenexempel soll dies untermauern. Geht man von 10 Mio. bekannten

Verbindungen aus und nimmt einen Speicherbedarf von 5 kByte für ein Spektrum an, so würde eine Datenbank mit den Infrarotspektren aller Substanzen mindestens 50 TByte benötigen. Diese gewaltige Datenbank ließe sich bereits heute realisieren.
Gegenwärtig werden vorwiegend Vollspektren in die Datenbanken aufgenommen. Dabei mißt man wellenzahllineare Spektren, in der Regel mit 2 cm^{-1} aufgelöst, und unterwirft sie einem Qualitätssicherungsprozeß. Hierbei werden störende Absorptionen von CO_2 und H_2O eliminiert.
Bibliotheken mit Lininenspektren haben ihre Existenzberechtigung durchaus noch nicht eingebüßt . Die Bandenseparation bildet hierbei das größte Problem. Hierfür gut geeignet sind Savitzky-Golay-Filter [5,6]. Über die damit zugänglichen Ableitungen kann eine gute Peakerkennung realisiert werden. Die Charakterisierung der Banden erfolgt durch Lage, Intensität und Halbwertsbreite.

1.2.3. Suchverfahren

In den ständig wachsenden Datenbeständen spielen effektive Algorithmen zur Auffindung gespeicherter Informationen eine immer größere Rolle. Dabei müssen Zeitaufwand und Genauigkeit gegeneinander aufgewogen werden. Bei einer Literaturrecherche ist es nicht entscheidend, ob wirklich alle Artikel gefunden wurden, bei einer Patentrecherche ist es das sehr wohl.
Der sicherste und zeitaufwendigste Weg, einen Datensatz in einer Bibliothek aufzufinden, ist die sequentielle Suche. Indem die gesamte Datensammlung der Reihe nach mit dem Suchmuster verglichen wird, erhält man mit Sicherheit auch alle in der Datenbasis enthaltenen Informationen.
Mit der invertierten Suche können auch große Datenbanken in kurzer Zeit mit großem Erfolg durchsucht werden [7]. Hierzu ist es jedoch notwendig, zuvor in einem Indizierungslauf eine Indexdatei, hier als invertiertes File bezeichnet, zu generieren. Aus dem Spektrum werden die wichtigen Absorptionsbanden entnommen und ein Verweis in dem zu diesen Wellenlängenbereich gehörigen Record des Inversfiles erzeugt. Zur Suche eines Spektrums zerlegt man es zuerst in seine signifikanten Linien. Über diese greift man auf die einzelnen Records des Inversfiles zu. Sie werden gelesen und ausgewertet. Die graphische Darstellung dieses Prozesses ist der Abbildung 9 zu entnehmen. Der darin am häufigsten auftretende Verweis bezeichnet das Spektrum mit der größten Anzahl von Übereinstimmungen. Es kann jedoch vorkommen, daß mehrere Substanzen mit identischer Anzahl von Verweisen im Ergebnis auftreten, da sowohl Bibliotheks- als auch Probespektrum nicht in alle Linien zerlegt wurden. Somit bietet sich im Anschluß an eine invertierte Suche ein direkter Vergleich der in der Hitliste aufgeführten Spektren mit dem Spektrum der Probe an. Der

Nachteil der Inversfiles ist die aufwendige Indizierungsprozedur bei Erweiterung der Bibliothek und die je nach Wellenzahlbereich stark differierende Zahl der Verweise.

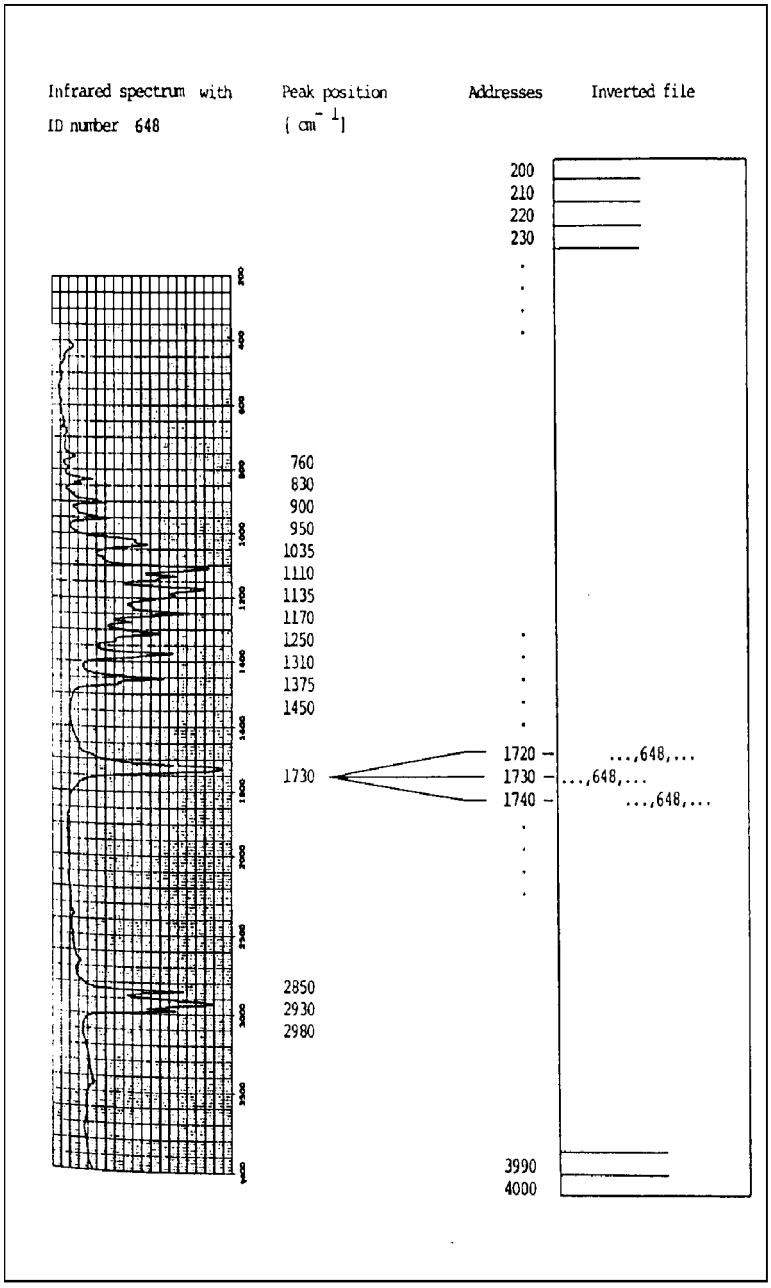


Bild 9 Invertierte Suche in einer IR-Datenbank [23]

Der Vollständigkeit halber soll noch die Hash-Code Methode [8] erwähnt werden.

Der Zugriff auf Spektrendaten ist auch über hierarchische Verfahren möglich. Ein Programm, in dem das erfolgreich angewandt wird, ist der Spec-Finder der Firma Sadtler Research Laboratories [9]. Das Spektrum wird hierbei in 27 Bereiche unterteilt. Die zur Suche genutzten Linien müssen eine Transmission von 70% und weniger aufweisen.

1.2.4. Besonderheiten bei SpecInfo

Da diese Arbeit in ein Interpretationssystem bei SpecInfo einfließen soll, seien einige Ausführungen dazu gestattet.
SpecInfo ist ein umfangreiches Programmsystem zur Suche und Auswertung von NMR-, IR- und Massenspektren. Es handelt sich um die umfangreichste kommerziell verfügbare Spektrensammlung. Diese Datenbank ist über STN-International online zugänglich, kann jedoch auch als Inhouse Version genutzt werden. Zum gegenwärtigen Zeitpunkt verfügt das System über Interpretationswerkzeuge zur Auswertung von NMR- und Massenspektren sowie eine Ähnlichkeitssuche für IR-Spektren.

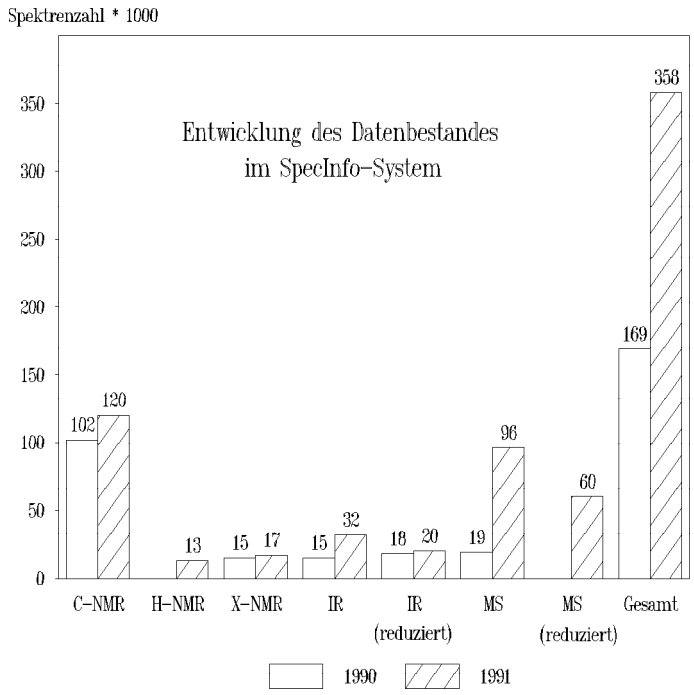


Bild 10 Spektren in der SpecInfo-Datenbank [24]

An den Programmen zur rechnergestützten Interpretation von IR-Daten wird gearbeitet. Zu jedem Spektrum sind Strukturformel, Summenformel, Molekulargewicht, CAS-Registry Nummer, systematischer und Trivialname gespeichert. Die Spektren wurden zum großen Teil bei der BASF gemessen.

1.3. Ähnlichkeitsmaße

Zur Bewertung der Übereinstimmung von gemessenem Spektrum und Bibliotheksspektrum werden Maßzahlen benötigt. Das Bild 11 zeigt das gemessene und das simulierte Spektrum ein und derselben Substanz. Es sind deutlich Störungen durch CO_2- und Wasserbanden zu erkennen. Dazu kommen noch die Differenzen zwischen Simulation und Messung, sowie die zufälligen und gerätetechnisch bedingten Schwankungen. Aus den hier aufgezeigten Störungen leiten sich auch die Anforderungen an die Abstandsmaße ab. Sie müssen gegenüber Störungen tolerant sein, eine starke diskriminatorische Kraft besitzen, charakteristische Merkmale hervorheben und mit geringem numerischen Aufwand berechenbar sein.

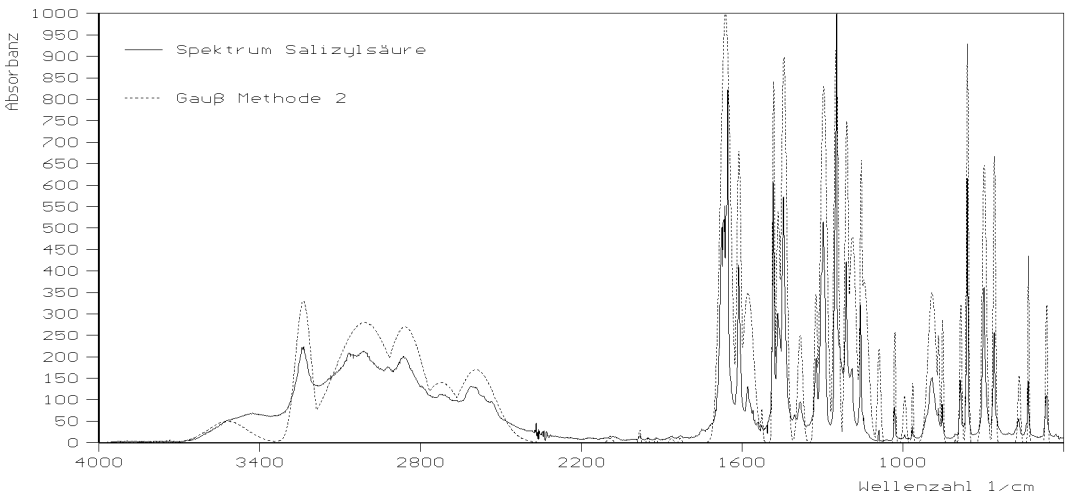


Bild 11 Messung und Simulation des Salizylsäurespektrums

Eine ganze Klasse von Abstands- oder Ähnlichkeitsmaßen ergibt sich aus dem Mahalanobis-Abstand:

$$d_{P,R} = \sqrt[D]{\sum_{i=1}^{N} (I_{P,i} - I_{R,i})^D}$$

D...Dimension
N...ZahlderDatenpunkte (4)

Durch Variation der Dimension erhält man für D=1 die Manhatten-Distanz (auch city-block):

$$d_{P,R} = \sum_{i=1}^{N} | I_{P,i} - I_{R,i} | \quad (5)$$

Der Euklidische Abstand hat die Dimension 2 :

$$d_{P,R} = \sqrt[2]{\sum_{i=1}^{N} (I_{P,i} - I_{R,i})^2} \quad (6)$$

Als ein Ähnlichkeitsmaß kann auch der Korrelationskoeffizient betrachtet werden.

$$r = \frac{\sum_{i=1}^{N} (I_{P,i} - \bar{I}_P) * (I_{R,i} - \bar{I}_R)}{\sqrt{\sum_{i=1}^{N} (I_{P,i} - \bar{I}_P)^2 * \sum_{i=1}^{N} (I_{R,i} - \bar{I}_R)^2}} \quad (7)$$

$I_{P,i}$...Intensität des Probenspektrums
$I_{R,i}$...Intensität des Referenzspektrums

Diese Maße eignen sich zum Vergleich von Vollspektren. In der Arbeit von Kwiatkowski [10] wird ein Maß für den Vergleich von Linienspektren angegeben.

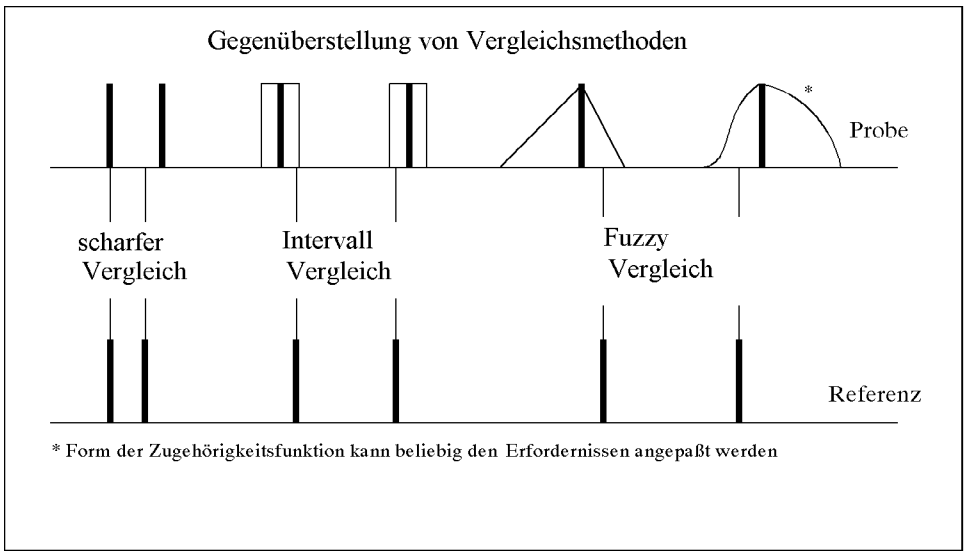


Bild 12 Unterschiede der Vergleichsmethoden

1.4. Fuzzy Logik zum Spektrenvergleich

Der Vergleich von Linienspektren hat gegenüber dem von Vollspektren den Vorteil, sehr schnell ausführbar zu sein. Das Problem hierbei besteht in der Durchführung. Die Repräsentation der Spektren erfolgt durch Vektoren. In der Regel weisen Proben- und Referenzvektor eine voneinander unterschiedliche Anzahl von Elementen auf. Daher ist es, um Störeinflüssen zu begegnen, nicht möglich, die o.g. Ähnlichkeitsmaße anzuwenden. Ein strenger Vergleich, z.B.: Bande bei 1673,2 cm^{-1} in Probe und Referenz enthalten ja/nein (1/0), wird nur in Ausnahmefällen zu einem Ergebnis führen. Diese Verfahrensweise wird in der Fuzzy-Logik als scharf bzw. crisp bezeichnet und ist eine Untermenge der mehrwertigen Logik. Man ist folglich gezwungen, Intervalle für die zu vergleichenden Größen anzugeben. In der Fuzzy-Logik geht man analog vor. Jedoch wird dabei ein Wert nicht durch einen Bereich ersetzt, in dem jeder Punkt die gleiche Wertigkeit besitzt, sondern durch eine zu den "Rändern" hin abnehmende Zugehörigkeitsfunktion oder "membership-function". Durch eine geeignete Wahl jener Funktionen kann auch der scharfe Vergleich bzw. der Intervallvergleich realisiert werden. Den Übergang zwischen den Methoden kann man in Bild 12 erkennen.

Die Bewertung der Anwesenheit einer Bande erfolgt nun nicht mehr mit 1 oder 0, sondern graduell über den Wert der Zugehörigkeitsfunktion an dem Schnittpunkt zwischen scharfem Bibliotheksspektrum und dem fuzzyfizierten

Probespektrum. Diese Vorgehensweise ist nicht allein auf die Bandenlage anwendbar, sondern ebenso auf Intensität und, soweit verfügbar, die Halbwertsbreite. Damit gelingt es, durch eine geeignete Wahl der membership-function das Wissen, um alle Einflüsse auf ein Spektrum in die Suche mit einzubeziehen. So wird die Zugehörigkeitsfunktion für die Linienposition immer "schärfer" sein als die für die Intensität. Die Gestalt der Funktionen unterliegt nur zwei Bedingungen, einerseits müssen sie stetig sein und andererseits einen monotonen Verlauf aufweisen. Die einfachste Form bildet die Dreieckskurve. Ebenso können beliebige andere Funktionen angewandt werden, so sie die obigen Bedingungen erfüllen. Die Parameter der Kurvengleichungen können in einem Trainingslauf an Hand historischen Datenmaterials festgelegt werden.
Die mathematische Beschreibung [11,12] des unscharfen Linienvergleiches wird in den nachfolgenden Formeln gegeben.

$$
\begin{aligned}
L_P &= \{x, m_P(x)\} = \cup L_{P_j} = \{x, \max_{j=1\ldots p} m_{P_j}(x)\} \quad x \in X \\
L_R &= \{x, m_R(x)\} = \cup L_{R_i} = \{x, \max_{i=1\ldots r} m_{R_i}(x)\} \\
&\text{mit: } m_{R_i}(x) = 1 \\
z &= \text{card}\,(L_P \cap L_R)/r = \sum_{i=1}^{r} \min\{m_P(x), m_{R_i}(x)\}/r
\end{aligned}
\tag{8}
$$

L	Menge der Linien
$m_{Pj}(x)$	Zugehörigkeitsfunktion der j-ten Linie der Probe
P	Index Probe
R	Index Referenz
z	Grad der Zugehörigkeit / Sympathiewert

1.5. Neuronale Netze

1.5.1. Überblick

Die Entwicklung neuronaler Netze begann in den 40-er Jahren und erlebte in den 60-ern ihre erste Blüte. Nach Abklingen der Euphorie gerieten sie bis zum Anfang der achtziger Jahre in Vergessenheit. Seit diesem Zeitpunkt werden sie als Hilfsmittel der KI wieder intensiv untersucht [13].

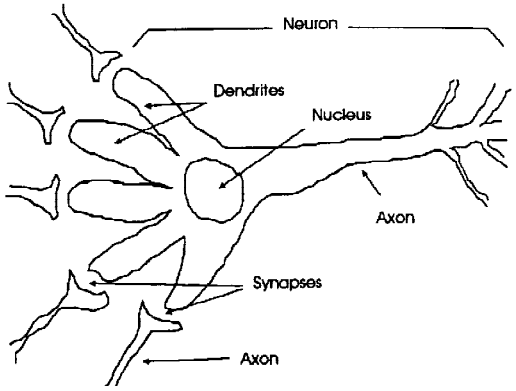


Bild 13 Biologisches Neuron

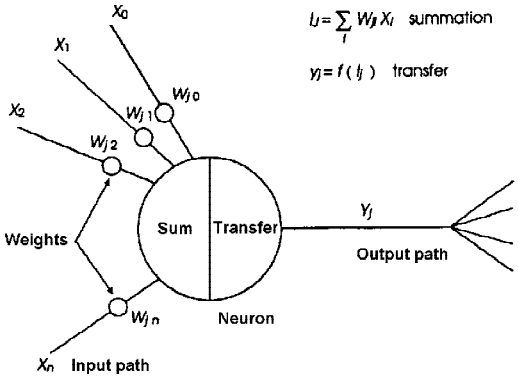


Bild 14 Mathematisches Neuron

Neuronale Netze sind mathematische Algorithmen, welche nach den Vorstellungen ihrer Entwickler, analog zu den Funktionsweisen des Gehirns arbeiten sollen. Ein Signal gelangt über die Synapsen (Gewichte) zu den Dendriten (Input-Pfad). Dabei wird die Signalintensität durch erstere verändert. Im Nukleus (Neuron) erfolgt eine Summation der Signale. Bei Überschreitung eines Schwellenpotentials (Aktivierungsfunktion) feuert das Neuron und leitet ein Signal über das Axon (Output-Pfad) weiter.
Werden Ein- und Ausgänge mehrerer Neuronen verknüpft, erhält man ein neuronales Netzwerk. Um darin Wissen zu speichern, werden in einem iterativen Prozeß die Gewichte verändert. Dazu sind verschiedene Lernstrategien entwickelt worden. Über diese können die Netze klassifiziert werden. Es existieren Methoden zum "unsupervised learning", dem unbeaufsichtigtem Lernen, (Kohonen-Netz, u.a.) und zum "supervised learning", dem beaufsichtigten Lernen, (Perceptrone, u.a.).

1.5.2. Kohonennetze zur Clusterung

Der Algorithmus für dieses Netz wurde von T. Kohonen entworfen und in [14]

vorgestellt. Der Kerngedanke dieser Topologie besteht darin, daß in der Natur eine bestimmte Klasse von Sinneswahrnehmungen auch zu einer topographisch geordneten Neuronenstruktur im Gehirn führt. Die Kohonen-Netze werden auch als "Self-Organizing Maps" (SOM) [15] bezeichnet.
Sie bestehen aus einer Schicht Neuronen, wobei jedes mit seinem Nachbarn und zudem mit jedem Element des Input-Vektors verschaltet ist (Bild 15). Während des Lernens werden die Objekte zufällig dem Netz präsentiert. Dabei stellen sich die Gewichte zwischen den Neuronen und den Eingängen auf Werte ein, die der Wahrscheinlichkeitsdichte der angebotenen Daten nahekommt. Dadurch werden ähnliche Objekte auf benachbarte Neuronen abgebildet.

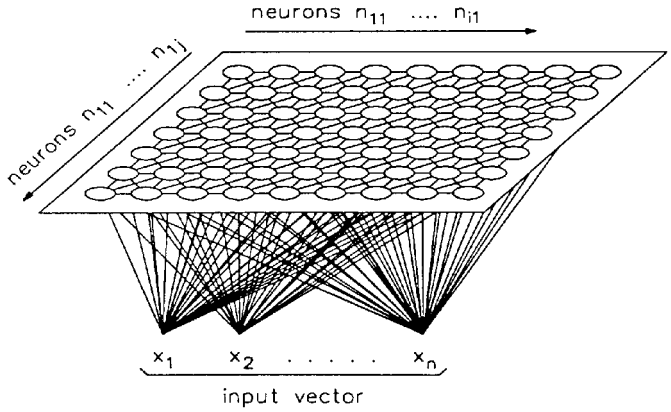


Bild 15 Aufbau eines Kohonen-Netzes

Das Training vollzieht sich nach folgendem Schema [16]:

1. Initialisierung der Gewichte mit zufälligen Werten aus (-0.1, +0,1)
2. Objekt aus Trainingsdaten auswählen und dem Netz präsentieren
3. Berechnung des Euklidischen Abstandes zwischen Eingangsvektor und den Gewichten aller Neuronen
4. Neuron mit dem geringsten Abstand auswählen
5. Gewichte der Neuronen in der Nachbarschaft anpassen gemäß nachfolgender Formel

$$w_{ij}(t+1) = w_{ij}(t) + \tau(t)(x_i - w_{ij}(t)) \quad (9)$$

mit
w_{ij} ... Gewichte des Neurons j
x_i .... Element des Eingangsvektors
τ ..... Lernrate, nimmt mit der Zeit ab
..... Index der zu stimulierenden Nachbarneuronen

6. Wiederholung der Prozedur ab 2. bis σ=0

2. Programme und Ergebnisse

2.1. Genutzte Soft- und Hardware

Die Erstellung der Software erfolgte in der Programmiersprache C. Dabei wurde auf die weitestmögliche Einhaltung des ANSII-Standards geachtet, um die Programme möglichst portabel zu halten. Die Forderung nach umfassender Portabilität ergibt sich aus den verschiedenen Rechnerwelten, für die entwickelt wurde. Das SpecInfo-System ist zum gegenwärtigen Zeitpunkt nur auf VAX unter VMS verfügbar. Die umfangreichen Berechnungen mußten auf einer Workstation RS6000/320 abgearbeitet werden. Die Entwicklung und Testung der Routinen sollte unter MS-DOS auf einem PC erfolgen, da hier bereits einige Erfahrungen im Umgang mit den Entwicklungswerkzeugen vorlagen.
Alle Programme, die nicht auf Grafikausgaben angewiesen sind können, ohne Änderungen übertragen werde. Die grafische Werteausgabe funktioniert nur unter MS-DOS.
Als Rechner standen ein PC 486 33 MHz EISA sowie eine Workstation IBM RS6000 Modell 320 zur Verfügung. Als Entwicklungstools wurden MS-QuickC 2.5 unter MS-DOS 5.0 sowie unter AIX 3.1.5.12 der zum System gehörige C-Compiler xlc (Version 1.1.5.10) genutzt. Da für einige Programme der unter MS-DOS verfügbare Speicherplatz nicht ausreichte, kam auch ein C-Compiler mit DOS-Extender zum Einsatz. Es handelt sich hierbei um GNU gcc/386 1.39 für DOS.
Für die Aufnahme der IR-Spektren konnte ein Spektrometer des Typs SPECORD M82 der Firma Carl Zeiss Jena genutzt werden. Die Bearbeitung der aufgenommenen Daten erfolgte mit der dazugehörigen Software.
Die statistischen Auswertungen sowie die grafischen Darstellungen der Spektren wurden mit dem Programm Multigraf Version 2.6 von WEKA realisiert.

2.2. Datenbestände und -manipulation

Die zur Verfügung stehenden Daten sind im Clear-Text-Format (TU München) gespeichert. Der Zugriff auf einzelne Datenblöcke erfolgt über einen Satz von Schlüsselwörtern. Diese sind nach folgendem Muster Aufgebaut:

"Leerzeichen" "/" "Schlüssel" "Zahl der Informationseinheiten" "Zeilenzahl"

```
 /IDENT          1    1
2
 /NAME           1    1
75-04-7
 /MOLECULS       1    1
 1 1 10
 /MOLNAMES       1    2
 0
ETHYLAMINE
 /ATOMS         10   10
 1 6 0 1 4
 2 6 0 5 8
 3 7 2 9 11
 4 1 0 12 12
 5 1 0 13 13
 6 1 0 14 14
 7 1 0 15 15
 8 1 0 16 16
 9 1 0 17 17
 10 1 0 18 18
 /BONDS         18   18
 1 1 2 1
 2 1 3 1
 3 1 4 1
 4 1 5 1
 5 2 1 1
 6 2 6 1
 7 2 7 1
 8 2 8 1
 9 3 1 1
 10 3 9 1
 11 3 10 1
 12 4 1 1
 13 5 1 1
 14 6 2 1
 15 7 2 1
 16 8 2 1
 17 9 3 1
 18 10 3 1
 /GRTEXT         2    2
2
ETHYLAMINE
 /PEAKS         14   14
 15 10
 16 1
 18 3
 27 13
 28 32
 29 7
 30 100
 31 2
 41 5
 42 9
 43 3
 44 20
 45 19
 46 1
 /IRSPEKB       18   18
 3429.2 400 999
 3369.5 240 960
 3299.1 240 730
 2970.2 60 400
 2876.7 60 270
 1644.2 90 210
 1599.9 34 180
 1454.2 24 50
 1394.4 14 60
 1380.0 22 40
 1144.7 30 40
 1077.2 14 30
 1053.1 12 70
 1017.4 28 60
 955.7 72 90
 879.5 24 70
 799.4 48 80
 692.4 240 110
 /END            0    0
```

Tabelle 2: Ausschnitt aus der Datenbasis

In der nebenstehenden Tabelle ist der Datensatz für Ethylamin wiedergegeben. Eintausend solcher Sätze sind in einer Datei zusammengefaßt. Sie umfaßt Substanzen welche die häufigsten funktionellen Gruppen enthalten und einen repräsentativen Durchschnitt der Spektren organischer Verbindungen darstellen sollen. Die Struktur der Verbindung kann über "Atom-" und "Bond-List" erzeugt werden. In dem uns vorliegenden Datenmaterial ist neben der Peaktabelle des IR-Spektrums noch das Massenspektrum gespeichert. Erstere sind in der Regel mit Wellenzahl, Halbwertsbreite und Intensität gegeben. Bei ca. 30 Spektren liegen keine Werte für die HWB vor. Die Menge der Daten zu einem Schlüssel ist variabel. Sie wird durch die Zahl der Informationseinheiten spezifiziert. Eine Differenz zwischen jener und der Zeilenzahl gibt die Menge der Leerzeilen an. Die Informationen zu einer Substanz müssen immer zwischen "/IDENT" und "/END" stehen. Zur Anordnung einzelner Datenblöcke dazwischen gibt es für einige Schlüssel Empfehlungen, meist können sie wahlfrei angegeben werden. Für die Schlüsselwörter existiert eine Liste [17] mit allen zulässigen Einträgen. Unsere Datenbasis enthält nur einen Teil der zu jeder Verbindung vorliegenden Informationen.

Um die sequentiellen Zugriffsoperationen auf die Daten zu beschleunigen, wurden die uns nicht interessierenden Datenblöcke "/ATOMS", "/BONDS","/GRTEXT" und "/PEAKS" aus der Datenbank entfernt.

Die vermessenen Spektren wurden auf einem SPECORD M82 aufgenommen. Die Präparation der Festsubstanzen erfolgte in KBr, die Messung der Flüssigkeiten in NaCl-Küvetten. Die Tabletten bestanden aus 3 mg Substanz und 997 mg KBr und wurden 3 min. einem Preßdruck von etwa 7500 bar unter gleichzeitiger

.Evakuierung der Preßform ausgesetzt. Die Herkunft der eingesetzten Substanzen ließ sich nicht mehr ermitteln. Da sie jedoch als Testsubstanzen in einem chromatographischen Laboratorium eingesetzt wurden, kann von ihrer Reinheit ausgegangen werden. Für die am häufigsten eingesetzte Salizylsäure wurde trotzdem eine Reinheitsprüfung mittels HPLC vorgenommen. Die in dieser Arbeit gezeigten Bilder und Beispiele beziehen sich, soweit nicht anders angegeben, auf diese Substanz.

2.3. Methodik der Auswertung

Das Resultat einer Spektrensuche bildete immer eine Hit-Liste wie sie in Tabelle 6 zu sehen ist. Darin sind die Substanznamen, die Abstände der Vorschläge und die Platzziffern eingetragen. Die Sortierung erfolgt anhand des zunehmenden Abstandes. Außerdem werden noch Rekonstruktionsmethode und Abstandsmaß verbal aufgeführt. Zur Bewertung der Methoden und Maße wurde ein Platzziffernsystem eingeführt. Es erfolgte ein Vergleich der gemessenen Spektren mit den auf vier verschiedenen Wegen simulierten Spektren (Methode 1 und 2 mit Gauß- und Lorentz-Profil). Die Bewertung erfolgte durch die in 2.4.3. vorgestellten Maße. Der Platz, auf dem sich die Substanz in der Hitliste wiederfand, wurde ausgewertet und in ein Auswertungsformular (Bild 16) eingetragen. Die Platzziffer 50 erhielten Substanzen, die nicht wiedergefunden wurden. Über die Zeilen- und Spaltensummen erfolgte die Einschätzung der Methoden und Maße.

Testsubstanz	Methode 1		Methode 2		Σ
	Gauß	Lorentz	Gauß	Lorentz	
Manhatten-	50	50	50	50	200
Euklidischer-	10	19	4	5	28
Gewichteter-Euklidischer-	8	10	4	3	25
Vier-dimensionaler-Abstand	4	4	1	1	10
Korrelations-abstand	1	1	1	1	4
Σ	73	84	60	60	

Bild 16 Formular zur Auswertung der Datenbankabfragen

2.4. Simulation von Spektren

2.4.1. Methoden zur Simulation

Um einen direkten Vergleich von Vollspektren mit den Bibliotheksdaten realisieren zu können, müssen aus jenen wieder vollständige Spektren erzeugt werden.

Gaußfunktion

$$I_i = I_{max} * \exp\left(-\frac{\ln 2}{(HWB/2)^2} * (\overline{\nu}_i - \overline{\nu}_{max})^2\right) \quad (10)$$

Lorentzfunktion

$$I_i = \frac{I_{max} * (HWB/2)^2}{(HWB/2)^2 + (\overline{\nu}_i - \overline{\nu}_{max})^2} \quad (11)$$

mit:

$\overline{\nu}$ … Wellenzahl

I … Intensität

HWB … Halbwertsbreite

Die für den IR-Bereich typischen Bandenformen (Bild 17) lassen sich durch Gauß- (10) oder Lorentzkonturen (11) mathematisch beschreiben [18]. Üblicherweise wird letztere dafür genutzt. Eine der Realität sehr nahe kommende Form ergibt sich aus der Kombination von beiden. In der vorliegenden Arbeit kamen die Funktionen in der ursprünglichen Form zum Einsatz.

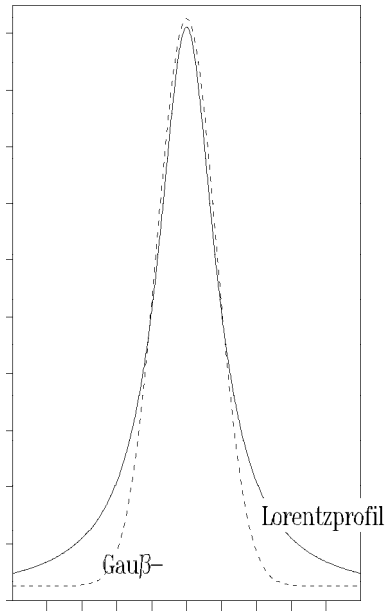


Bild 17 Gauß- und Lorentzprofil

Neben der Kontur einer einzelnen Bande beeinflußt die Art der Aggregation die Gestalt des rekonstruierten Spektrums. Es wurden zwei gegensätzliche Algorithmen entwickelt.

Der erste beruht auf der Annahme, daß bei der Spektrenzerlegung überlappende Banden separiert wurden. Bei der Rekonstruktion wurden nacheinander die Banden einzeln generiert und durch Summation zum Spektrum vereinigt. Die Simulation erfolgte in einem Bereich von ± 2*HWB um das Maximum herum. Eine Normierung vervollständigte diese Funktion. Im weiteren Text wird dieses Verfahren als Methode 1 bezeichnet.

Bei Methode 2 werden alle Peaks gleichzeitig bearbeitet. Die Simulation beginnt am Maximum. Von da aus erfolgt eine Verbreiterung der Banden um jeweils eine Einheit. Sie wird so lange wiederholt bis alle benachbarten Linien miteinander verbunden sind und ein geschlossener Kurvenzug entstanden ist.

Der Nachteil der ersten Verfahrensweise ist, daß hohe Anforderungen an die

Spektrenzerlegung gestellt werden müssen. In den zur Verfügung stehenden Unterlagen über die Datenbank finden sich keine näheren Angaben über die Art der Generierung von Peaklisten. Die Unterschiede der Verfahren (Bild 18) zeigen sich insbesondere in Bereichen überlappender breiter Banden.

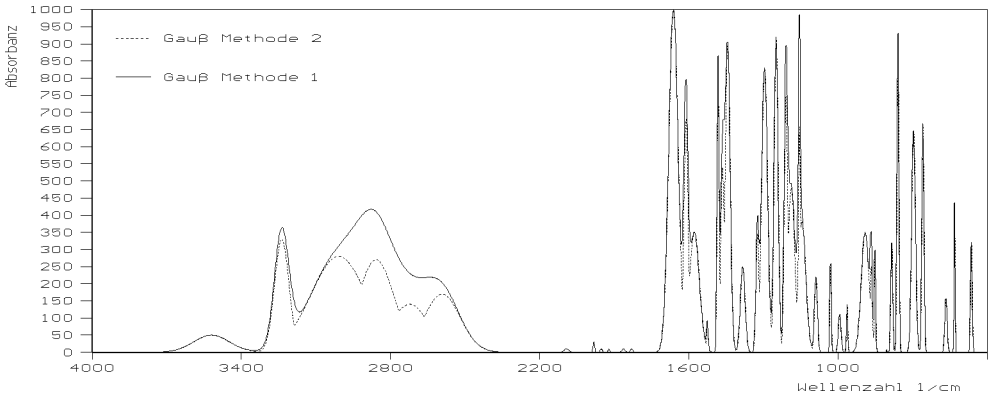


Bild 18 Unterschiede der Methoden zur Spektrenrekonstruktion

2.4.2. Bestimmung der Übereinstimmungsmaße

Zur Bestimmung des zum Vergleich von Vollspektren am besten geeigneten Maßes mußten alle Algorithmen zusammen mit den verschiedenen Simulationsmethoden getestet werden. Für jede der sieben zur Verfügung stehenden Substanzen waren daher 20 Testläufe notwendig. Die Meßergebnisse wurden anfänglich ohne Vorverarbeitung (außer Konvertierung und Normierung) eingesetzt. Dabei zeigte es sich, daß die Spektren der Xylene durch Absorptionen des Küvettenmaterials gestört waren. Die Untergrundkompensation führte zu "negativen" Banden, welche bei der Normierung eine Verschiebung der Basislinie bewirkten. Der Einsatz dieser Spektren sollte die Fehlertoleranz der Abstandsmaße testen. Es erfolgte eine Nachbearbeitung jener Spektren unter Verwendung der Firmware des Spektrometers. Anschließend wurden sie nochmals der Testprozedur unterworfen.

Die Wichtungsfunktion für den modifizierten Euklidischen Abstand wurde in Anlehnung an [19], den Aussagen aus der Linienverteilung in Bild 2 dieser Arbeit und unter Berücksichtigung der Störungen durch CO_2 und Wasserdampf erstellt. Sie setzt sich aus den in Tabelle 3 aufgeführten Intervallen zusammen.

Wellenzahlbereich	Wichtungsfaktor
4000 - 3600	0.5
3600 - 2800	1.0
2800 - 2375	0.75
2375 - 2300	0.5
2300 - 600	1.0
600 - 400	0.5

Tabelle 3 Werte der Wichtungsfunktion

2.4.3. Vergleich simulierter und realer Spektren

Den Hauptteil der Arbeit stellt das Programm "find" dar. Es vereinigt die verschiedenen Methoden zur Rekonstruktion von Vollspektren aus Peaklisten und fünf unterschiedliche Abstandsmaße. Während des Suchvorgangs können die Spektren auf dem Bildschirm eins MS-DOS Rechners angezeigt werden. An Abstandsmaßen stehen zur Verfügung:

- Manhatten-Abstand
- Euklidischer-Abstand
- Vier-dimensionaler-Abstand
- Korrelationsabstand
- Gewichteter euklidische Abstand.

Um eine Ordnung der Resultate mit abnehmender Ähnlichkeit auch für den Korrelationskoeffizienten zu erhalten, wurde der Korrelationsabstand eingeführt. Dieser entspricht *1-Korrelationskoeffizient*.
Bei dem letzten Maß wird der euklidische Abstand mit einer von der Wellenzahl abhängigen Wichtungsfunktion kombiniert. Durch wiederholte Päparation und Messung von Salizylsäure (Bild 19) sollte die Abhängigkeit der Varianz eines Spektrums von der Wellenzahl ermittelt werden, um daraus die Wichtungsfunktion ableiten zu können.

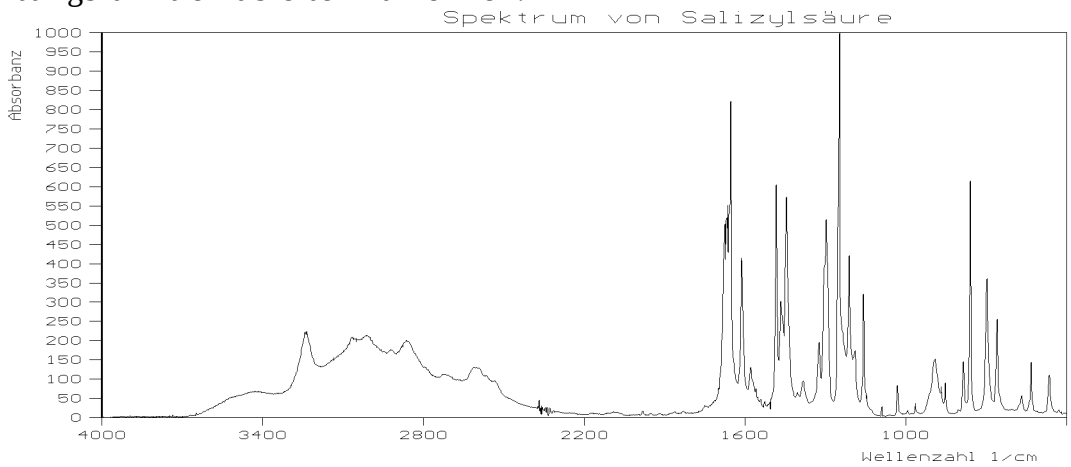


Bild 19 Spektrum von Salizylsäure

Das FT-IR Spektrometer konnte nicht genutzt werden, da keine Beschreibung des Datenformates und kein Konvertierungsprogramm vorlagen. Die Messungen mußten somit auf einem klassischen sequentiellen Gerät erfolgen und waren daher sehr zeitaufwendig. Jede Messung dauerte etwa eine Stunde. Daraus resultierend konnte nur eine Serie von 11 Spektren gemessen werden. Es zeigt sich, daß die Standardabweichung (Bild 20), in den Bereichen der Störungen, einen relativ konstanten Wert annimmt. Dieser ist nahezu unabhängig von den vorkommenden Banden. Die Wichtungsfunktion dämpft diesen Bereich der

Störungen und solche Abschnitte mit geringer Informationsdichte, die anderen gehen unverändert in die Bewertung ein. Beim Vergleich mit dem Spektrum (Bild 19) wird deutlich, daß die Intensität der Banden mit einem großen Wiederholungsfehler behaftet ist. Man kann es sehr gut an den hohen Werten der Varianz in der Umgebung einer Bande erkennen. Daraus resultiert auch ein Teil der Differenzen zwischen realem und rekonstruiertem Spektrum. Die Normierung der Spektren erfolgt über die Intensität. Folglich kommt es zu einer Verschiebung der Spektren gegeneinander. Es ist in den Bildern 11 und 21 gut zu erkennen. Einen
möglichen Ausweg bildet die Normierung auf die Fläche.

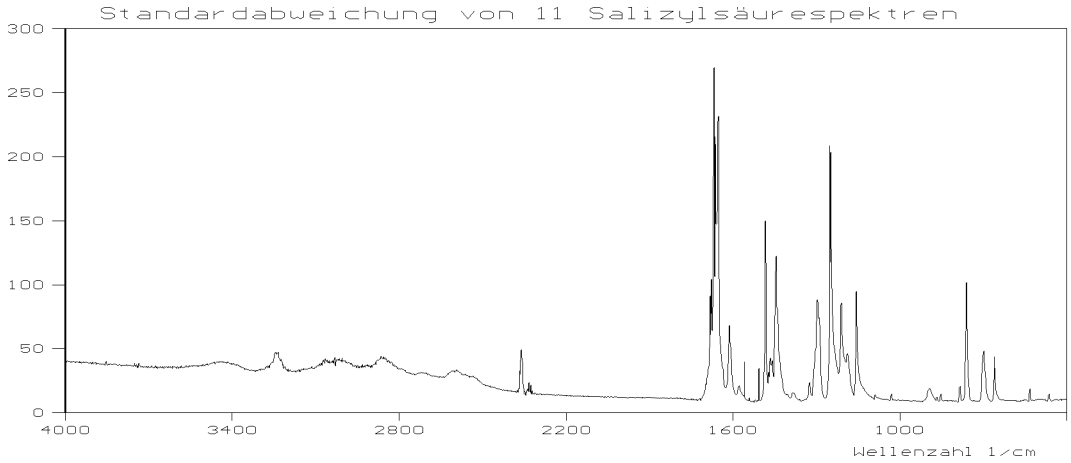


Bild 20 Standardabweichung in Abhängigkeit von der Wellenzahl

Die Berechnungen benötigen für 1000 Spektren auf dem PC wie auch auf der Workstation etwa zwei Minuten, auf einem anderen Rechner (386sx/16 MHz) ohne Coprozessor dagegen ca. sieben Stunden. Das Programm benötigt etwa 250 kByte Speicher wovon etwa 120 kByte auf die Daten entfallen. Die Darstellung der Spektren erfolgt als Extinktionsspektren. Die Grafikausgabe, siehe Bild 21, ist momentan nur auf dem PC möglich. Voraussetzung dafür ist eine VGA-Grafikkarte.

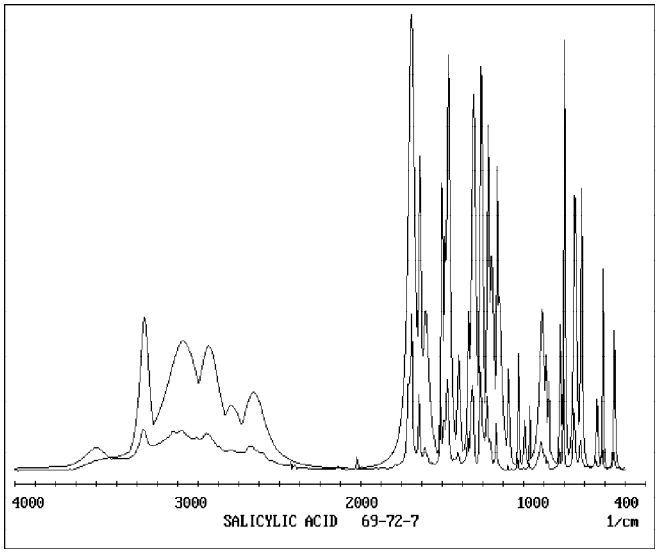


Bild 21 Darstellung
der Spektren auf dem Bildschirm

Die Pfade zu den Unterverzeichnissen für Daten und Ergebnisse sowie die Bibliothek werden durch die umseitig aufgeführten Umgebungsvariablen spezifiziert. Sie müssen vor dem Programmlauf gesetzt werden, am Besten in der Datei autoexec.bat (MS-DOS) bzw. .login (UNIX csh).

IRBIBO=C:\IR\BIBO.DAT
IRDAT=C:\IR\DAT
IRERG=C:\IR\ERG

Diese Verfahrensweise erscheint auf den ersten Blick etwas umständlich, ist jedoch erforderlich, um die Flexibilität zu gewährleisten. In einem Netzwerk können so die Daten verschiedener Nutzer auch in unterschiedlichen Verzeichnissen abgelegt werden. Die Bibliothek kann hingegen auf einem zentralen Netzlaufwerk allen zugänglich gemacht werden. Die Funktionen des Programmes "find" sind so konzipiert, daß sie ohne Änderungen in zukünftigen Rechnerprogramme integriert werden können. Die einzelnen Module tauschen die Daten nur über definierte Schnittstellen aus. Wenn Zugriffe auf globale Variablen erfolgen, so sind diese im Funktionskopf dokumentiert.

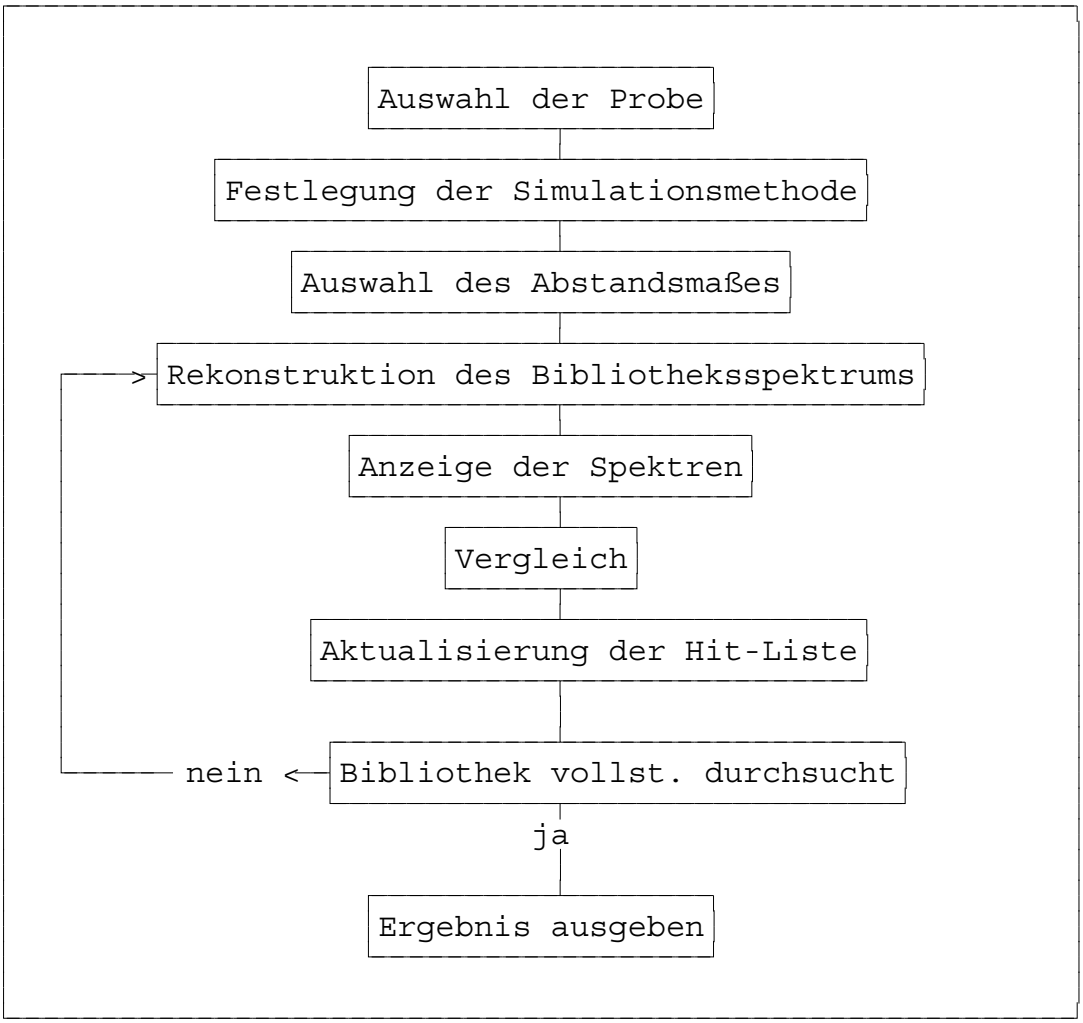


Bild 22 Programmablaufplan des Programmes "find"

2.4.4. Ergebnisse

Die Auswertung der Datenbankabfragen mit unterschiedlichen Abstandsmaßen und Methoden zur Rekonstruktion der Spektren ergab einige interessante Ergebnisse.

Die Art der Spektrenrekonstruktion hat bei den vorliegenden Daten keinen Einfluß auf die Wiederfindung eines Spektrums innerhalb der Datenbank. Die Werte in Tabelle 4 deuten nur für das Lorentzprofil bei Methode 2 auf eine systematischen Unterschiede zu den anderen Methoden hin. Es gilt hierbei jedoch zu bedenken, daß es sich hier nur um einen Ausschnitt aus einer größeren Spektrensammlung handelt. Mit Vergrößerung der Spektrenzahl steigt auch die Wahrscheinlichkeit des Auftretens ähnlicher Spektren. Daraus ergeben sich kleinere Differenzen bei den Abständen in der Hitliste. Möglicherweise erweist sich dann eine der vier getesteten Rekonstruktionsmethoden als die Bessere. Dabei dürfen die Abstandsmaße nicht aus den Augen verloren werden. Es bestehen starke Wechselwirkungen zwischen Maß und Simulation. So fallen die Unterschiede zwischen den Simulationen bei Verwendung des Korrelationsabstandes geringer aus. Die Aussagen zur Übertragung der Rekonstruktionsmethode auf eine große Datenbank gelten ebenso für die Abstandsmaße.

	Rekonstruktion				Maße				
	Methode 1		Methode 2		M	E	GE	V	K
	G	L	G	L					
Benzoesäure	214	224	199	201	200	200	200	198	40
Hydrochinon	59	78	201	201	200	111	108	114	6
Salizylsäure	201	181	139	155	200	142	200	130	4
ortho-Xylen	54	54	54	54	200	4	4	4	4
meta-Xylen	54	54	54	54	200	4	4	4	4
para-Xylen	54	54	54	54	200	4	4	4	4
Naphthalin*	58	59	59	67	200	5	5	4	29
Naphthalin**	63	88	65	146	200	40	26	4	92
ortho-Xylen***	54	54	54	54	200	4	4	4	4
meta-Xylen***	54	54	54	54	200	4	4	4	4
para-Xylen***	60	63	60	71	200	28	21	4	4
Σ	925	963	993	1111	2200	546	580	474	195

* Christiansen-Effekt 15 min Mahldauer
** Christiansen Effekt 30 min Mahldauer
*** Grundlinienstörung durch Küvettenmaterial

G ... Gauß-
L ... Lorentz-Profil

M ... Manhatten-
E ... Euklidischer-
GE ... Gewichteter Euklidischer-
V ... Vier-Dimensionaler-Abstand
K ... Korrelationsabstand

Tabelle 4 Übersicht über alle Tests

Im Gegensatz zu der Simulation läßt sich bei den Maßen ein relativ klarer Sieger bestimmen. Die Summen der Platzziffern des Korrelationsabstandes sind die mit Abstand kleinsten. In der Anlage 5 sind die Werte für alle Testläufe aufgeführt. In Tabelle 5 sind die Plätze der Substanzen in den Hitlisten bei Verwendung jenes Maßes für die verschiedenen Rekonstruktionsmethoden eingetragen. Selbst die Störungen der Grundlinien verfälscht die Ergebnisse bei keiner Substanz. Die Folgen des Christiansen-Effektes sind nicht mehr kompensierbar. Das ist auch verständlich, handelt es sich doch um eine Verzerrung der Intensitätsverhältnisse. Bei den Grundlinienstörungen ist es nur eine Parallelverschiebung des Spektrums. Die Verhältnisse der Intensitäten zueinander bleiben davon unberührt. Hier erweist sich der Korrelationsabstand den reinen Abstandsmaßen als überlegen. Auch die nahezu unvermeidlichen Störungen durch CO_2 und Wasser haben offensichtlich keinen großen Einfluß.

	Methode 1		Methode 2	
	Gauß	Lorentz	Gauß	Lorentz
Benzoesäure	14	24	1	1
Hydrochinon	2	2	1	1
Salizylsäure	1	1	1	1
ortho-Xylen	1	1	1	1
meta-Xylen	1	1	1	1
para-Xylen	1	1	1	1
ortho-Xylen*	1	1	1	1
meta-Xylen*	1	1	1	1
para-Xylen*	1	1	1	1
Naphthalin** 15 min	5	6	6	12
Naphthalin** 30 min	6	27	9	50
Σ	34	66	24	71

* Grundlinienstörung ** Christiansen-Effekt

Tabelle 5 Platzziffern bei der Suche mit dem Korrelationsabstand

Aus den Ergebnissen lassen sich auch Forderungen an die Spektren der zu findenden Verbindungen ableiten. Die Ansprüche sind dank der guten diskriminatorischen Kraft des Korrelationsabstandes relativ gering. Die Korrektur einer gestörten Grundlinie sollte erfolgen. Dabei kann auf höchste Präzision sicherlich verzichtet werden. Spektren, in denen der Christiansen-

Effekt auftritt müssen nochmals vermessen werden. Eine Verlängerung der Mahldauer hat, wie an den beiden Naphthalin-Spektren sichtbar, nicht immer Erfolg. Eine Veränderung des Einbettungsmittels, KI oder KCl anstelle KBr, ist eine weitere Möglichkeit dazu. Die Absorptionen durch CO_2 sollten entfernt werden. Bei den vorliegenden 1000 Spektren spielte diese Störung zwar keine Rolle, in einer bedeutend größeren Datenbasis ist sie sicher nicht mehr zu vernachlässigen.

Salizylsäure Korrelationsabstand Gauß-Profil Methode 2

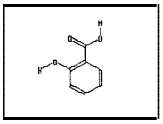
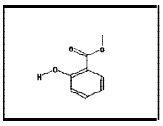
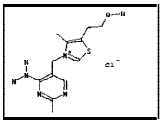
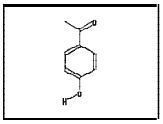
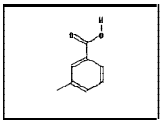
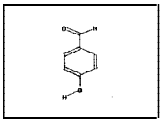
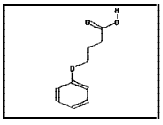
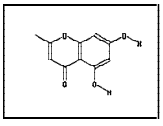
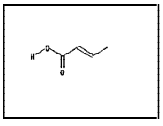
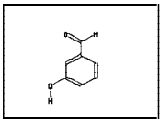
1	0.1295	Salicylic acid	
2	0.3745	Salicylic acid methyl ester	
3	0.3865	Thiamin, salt	
4	0.3927	Ethanone, 1-(3-hydroxyphenyl)-	
5	0.3975	3-Methyl-benzoic acid	
6	0.4093	Benzaldehyde, 4-hydroxy-	
7	0.4117	Butanonic acid, 4-phenoxy	
8	0.4127	4H-1-Benzopyran-4-one, 5,7-dihydroxy-2-methyl	
9	0.4272	trans-Crotonic acid	
10	0.4276	Benzaldehyde, 3-hydroxy	

Tabelle 6 Ausschnitt aus einer Hit-Liste und die dazu gehörenden chemischen Strukturen

Die Art der Vorbehandlung der Spektren hängt auch von den dafür zur Verfügung stehenden Programmen ab. Im Falle ihrer automatischen Anwendung sollte im voraus geklärt sein, wie sie sich auf das Resultat der Suche auswirken. Heutige Programme bieten eine solche Vielfalt an Möglichkeiten, daß man schnell versucht ist, alle immer anzuwenden. Der dadurch möglicherweise entstehende systematische Fehler könnte größer sein als der eigentlich zu beseitigende. Das Resultat könnte daher verfälscht werden.
Aus diesen Ergebnissen kann für die vorliegenden Daten der Korrelation-sabstand im Zusammenspiel mit der Rekonstruktionsmethode 2 unter Verwendung des Gauß-Profils als optimales Suchverfahren angesehen werden.

2.5. Unscharfer Linienvergleich

2.5.1. Algorithmus für Fuzzy-Vergleich

Den Ausgangspunkt für diesen Teil bildet die Arbeit von Blaffert [20]. In der zu entwickelnden Funktion sollte der Vergleich von Peaklisten miteinander, somit ohne den zeitaufwendigen Schritt der Spektrenrekonstruktion, unter maximaler Ausnutzung vorhandener Informationen realisiert werden.
Die Zugehörigkeitsfunktionen für Intensität, Wellenzahl und Halbwertsbreite entsprechen einer Gauß-Kurve. Diese Funktion wurde gewählt, da die Varianzen der Parameter eines Spektrums als normalverteilt angenommen werden können. Gemäß der Theorie unscharfer Mengen ist die genaue Gestalt der "membership-function" von sekundärer Bedeutung, sie muß sowieso in einem Lernprozeß an die aktuellen Daten angepaßt werden. Mit Sicherheit hätte hier eine einfache Dreieckskurve den Anforderungen entsprochen. Da ein Naturwissenschaftler jedoch mit der Verteilungsfunktion und ihrer Anwendung in der Statistik vertraut ist, steigt die Akzeptanz des Verfahrens gegenüber der Verwendung der "Hütchenfunktion".

$$\text{Fuzzy} - \text{Abstand} = 1-(f_L * m_L + f_I * m_I + f_{HWB} * m_{HWB})$$
$$\text{mit: } f_L + f_I + f_{HWB} = 1 \qquad (12)$$
$$f\ldots\text{Aggregationsfaktoren}$$

Die Breite der Zugehörigkeitsfunktionen und die Aggregationsfaktoren (Gl.12) wurden in einer Reihe von Testläufen empirisch bestimmt. Die gefundenen Werte stellen eine erste Näherung dar, die für diese Vergleichsmethode

notwendige Funktion zur Zerlegung eines Spektrums in eine Peakliste stand noch nicht zur Verfügung. Die Linienspektren wurden manuell aus den Vollspektren gewonnen und sind daher mit einem großen Fehler behaftet.

```
for(i=0; i<Zahl der Bibliothekslinien; i++)
   {
   for(j=0; i< Zahl der Linien in Probe; j++)
      {
      nächstliegende Bande ermitteln
      Differenzen von Wellenzahl, Intensität und HWB errechnen
      Werte der Zugehörigkeitsfunktionen bestimmen und summieren
      }
   }
   Summen der Zugehörigkeitswerte normieren
   Aggregation der Werte
```

Bild 23 Algorithmus zum unscharfen Spektrenvergleich

2.5.2. Resultate

Dieser Algorithmus wurde ausgiebig bis jetzt nur an dem Spektrum einer Substanz getestet. Das Spektrum wurde dazu manuell in eine Peakliste zerlegt. Die Halbwertsbreite konnte auf diesem Weg nicht mit der erforderlichen Genauigkeit gewonnen werden. Die Parameter für den unscharfen Linienvergleich (Breite der Zugehörigkeitsfunktionen, Aggregationsfaktoren) wurden so gewählt, daß die gesuchte Substanz den ersten Platz belegte. In der Tabelle 7 sind die gefundenen Parameter und das Ergebnis der Suche zu sehen.
Vergleiche, die auf diesen Werten fußten, brachten gute Ergebnisse.Die Aggregationsfaktoren sollten aus der Varianzanalyse der Vergleichsmessungen gewonnen werden. Der Gedanke war, die Varianzanteile von Wellenzahl, Intensität und Halbwertsbreite zu bestimmen und die Aggregationsfaktoren im reziproken Verhältnis dazu zu wählen. Dazu sind jedoch noch weitere Messungen notwendig. Die empirisch gefundenen Werte und die damit durchgeführten Tests lassen den Schluß zu, daß dieses Verfahren auch für große Datenbanken eine hinreichend große Diskriminationsfähigkeit besitzt. Der geringe Rechenzeitaufwand, etwa 1/10 der Zeit des Vollspektrenvergleiches, prädestiniert diese Methode geradezu.

```
Breite-membership-function Wellenzahl: 6 cm-1
Wichtungsfaktor : 0.80
Breite-membership-function Intensität: 5
Wichtungsfaktor : 0.20
Breite-membership-function Halbwertsbreite: 0 cm-1
Wichtungsfaktor : 0.00

Salizylsäure

1  0.4191  SALICYLIC ACID
2  0.4334  1,6-DIISOCYANATO-HEXANE
3  0.5102  METHYLENE BROMIDE
4  0.5821  ACETIC ACID
5  0.5966  TETRAETHOXY-SILANE
6  0.5985  HYDRAZINE
7  0.6014  DIETHYL ETHER
8  0.6108  PYRIDINE, 4-METHYL-, 1-OXIDE
9  0.6178  FORMIC ACID METHYL ESTER
10 0.6325  2-METHYL-2-BUTENE
11 0.6375  BENZENE, (DICHLOROMETHYL)-
12 0.6463  PYRIDINE, 2-METHYL-, 1-OXIDE
13 0.6508  PROPANAL
14 0.6561  2-T-BUTYL-PHENOL
15 0.6572  Ethanamine, 2,2'-oxybis[N,N-dimethyl-
16 0.6578  3(2H)-Benzofuranone
17 0.6590  TRIETHYLORTHOFORMATE
18 0.6604  MALEIC ACID DIETHYL ESTER
19 0.6633  Cyclohexanemethanol, 2-methyl-
20 0.6662  Cyclopentene, 1-pentyl-
```

Tabelle 7 Ausschnitt aus einer Hit-Liste des Fuzzy-Vergleiches

2.6. Clusterung mit Kohonen-Netzen

2.6.1. Aufbau des Netzes

Übliche Verfahren der Indizierung einer Spektrendatenbank nutzen einzelne Banden als Schlüssel für den Zugriff. Das setzt eine Zerlegung der Spektren voraus. Damit einher geht ein Verlust an Informationen. In sehr großen Datenbanken mit vielen ähnlichen Spektren könnte folglich ein Schlüssel auf zwei verschiedene Stoffe verweisen. Es erschien daher sinnvoll, gleich eine umfassende Gruppeneinteilung zur Indizierung heranzuziehen.

Die Kohonen-Netze sind für solche Aufgaben gut geeignet. Die Wissensspeicherung in einem solchen Netz erfolgt in einer Hypermatrix, X-Neuronen * Y-Neuronen * Merkmale. Daraus ergab sich die Notwendigkeit einer sinnvollen Datenreduktion, ein Netz mit 100*100 Neuronen mit allen Merkmalen eines Vollspektrums, 1801 Punkte, würde etwa 70 MByte benötigen. Durch die Beschränkung auf zwei spektroskopisch relevante Bereiche ließ sich die Merkmalsanzahl auf 266 einschränken.

Bereich 1 von 3400 - 2700 cm^{-1} Auflösung 6 cm^{-1}
Bereich 2 von 1400 - 800 cm^{-1} Auflösung 4 cm^{-1}

Das Training eines solchen Netzes ist immer noch mit einem erheblichen Zeitaufwand verbunden. Da es jedoch nur einmal erfolgen muß, erscheint es noch vertretbar. Störend wirkt sich hier die Speichergrenze von MS-DOS aus. Sie kann jedoch unter Verwendung eines Extenders aufgehoben werden. In dieser Arbeit ließ es sich mit dem GNU-C Compiler realisieren. Unter moderneren Betriebssystemen (OS/2, UNIX) existiert eine solche Barriere nicht. Auf der Workstation ließ sich das Training besser durchführen. Die zum Training erforderlichen Daten konnten aus der Spektrenbibliothek zur Verfügung gestellt werden. Eine Iterationsanzahl von 50000 - 100000 gewährleistet, daß alle 1000 Spektren dem Netz hinreichend oft und mit gleicher Häufigkeit vorgespielt wurden. Die Auslastung der Kohonenschicht sollte m. E. 10% nicht übersteigen. Bei zu hoher Dichte können sich die Übergänge zwischen den gelernten Mustern nicht ausprägen. Eine Einordnung neuer Spektren in jene Räume ist dann nicht mehr möglich.

2.6.2. Clusterung

Die Versuche, die 1000 Spektren der Bibliothek in verschiedene Klassen zu gruppieren, waren leider nicht erfolgreich. Das Netzwerk konnte zwar trainiert werden, es bildeten sich jedoch keine Cluster. Ein Teil der Daten war gleichmäßig über die Kohonenschicht verteilt, der Rest konzentrierte sich auf die Ränder. In der Abbildung 24 ist das nicht zu erkennen, da einige Objekte auf den gleichen Neuronen abgebildet werden. Das ist möglicherweise ein Zeichen für die zu großen Unterschiede der Spektrenausschnitte. Um dem zu begegnen, müßte die Kohonenschicht weiter vergrößert werden. Eine Schicht von 100*100 Neuronen überfordert zum gegenwärtigen Zeitpunkt die uns zur Verfügung stehende Workstation. Eine Reduktion der Merkmale, anhand derer gelernt werden soll, ist eine weitere Möglichkeit. Dazu müßte der für die zugrunde liegende Struktur einer Verbindung charakteristische Frequenzbereich bekannt sein. Um diesen zu ermitteln ist eine Anzahl grundlegender Strukturen vorzugeben, die Datenbank nach den Spektren der diese Grundstruktur enthaltenden Stoffe zu durchsuchen und typische Merkmale zu bestimmen. Dieses Verfahren ist sicherlich durchführbar, aber zum Zwecke der Indizierung einer Datenbank wohl zu aufwendig. Diese Prozedur gehört wohl eher in die Spektreninterpretation.

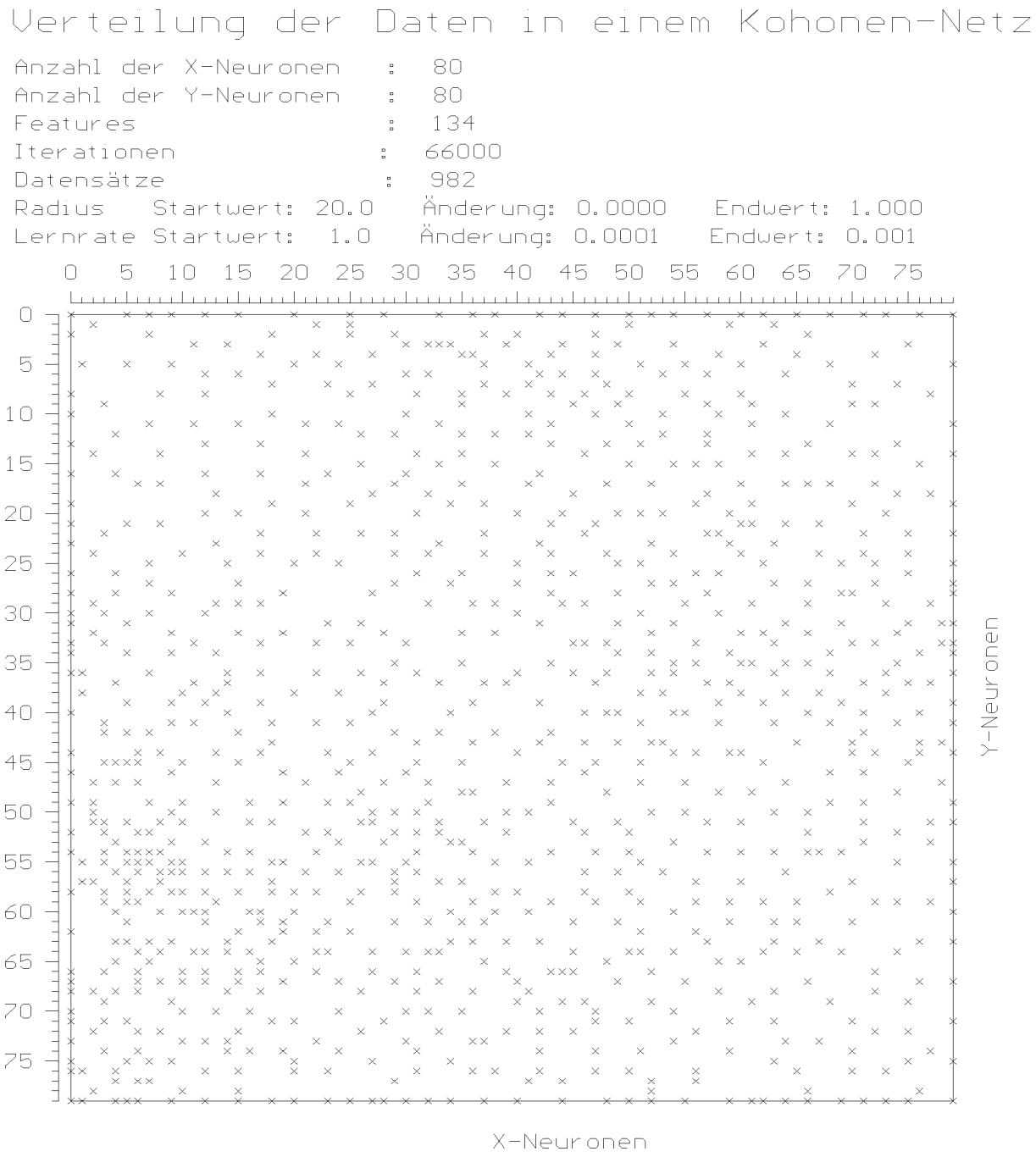


Bild 24 Resultat des Trainings eine Kohonen-Netzes

2.7. Aufwand-Nutzen-Verhältnis

Um endgültige Aussagen zu diesem Punkt treffen zu können, muß die Datenbank beträchtlich vergrößert werden. Zur Suche in den vorliegenden Daten reicht der unscharfe Linienvergleich aus. In einer vergrößerten Datenbasis müssen wahrscheinlich beide Verfahren kombiniert werden. Der hier

genutzte Algorithmus zur Rekonstruktion mit anschließendem Vergleich würde auf der Workstation für die in SpecInfo verfügbaren 20000 Spektren etwa 40 Minuten in Anspruch nehmen. Ein solcher Zeitaufwand erscheint heute nicht mehr akzeptabel. Wie sich die Rechenzeiten auf den für SpecInfo notwendigen VAX-Maschinen verhalten, kann noch nicht gesagt werden. Formal, nach Leistungstests, leisten sie etwa ein Viertel im Vergleich zur RS6000/320. Solche Aussagen sind mit einer gewissen Skepsis zu betrachten. Sollte jedoch die Tendenz stimmen, muß der Rechenzeitaufwand noch gesenkt werden. Die oben erwähnte Kombination der Suchverfahren scheint ein Weg dahin zu sein.

Von den im SpecInfo-System verfügbaren IR-Spektren liegen etwa 2000 ausschließlich als Peaklisten vor. Um sie voll nutzen zu können, ist ein funktionsfähiger Rekonstruktionsalgorithmus unabdingbar.

Die hier durchgeführten Untersuchungen erwiesen sich als sehr zeitaufwendig. Um die eigentlichen Methoden testen zu können, waren eine Vielzahl unterstützender Funktionen notwendig.

3. Zusammenfassung und Ausblick

In dieser Arbeit wurden verschiedene Methoden zur Suche in einer Spektrendatenbank getestet. Um Vollspektren mit Peaklisten vergleichen zu können, ist eine Methode zur Rekonstruktion ersterer notwendig. In den dazu durchgeführten Tests zeigten sich große Unterschiede zwischen den simulierten und den gemessenen Spektren. Die Differenzen waren nahezu unabhängig von der Art der Bandenprofile und der Rekonstruktionsmethode. Das gewählte Abstands- oder Ähnlichkeitsmaß war entscheidend für die Erkennung eines Spektrums. Der Euklidische Abstand, welcher sehr oft als Abstandsmaß für Ähnlichkeitssuchen genutzt wird, hat bei der Kombination von Rekonstruktion und Suche versagt. Die Erweiterung dieses Maßes mit einer Wichtungsfunktion brachte keine Verbesserung. Der aus der Mahalanobis-Distanz abgeleitete Abstand vierter Ordnung zeigte keine signifikant besseren Resultate. Der Manhatten-Abstand ist völlig ungeeignet. Der Korrelationsabstand (1-r) zeigte sich als ausreichend stabil gegenüber den Störungen durch Rekonstruktion des Bibliotheksspektrums und den Effekten bei der Messung.

Die Versuche zur Clusterung der Spektren mittels eines Kohonen-Netzes waren nicht erfolgreich. Die vermutliche Ursache ist die große Mannigfaltigkeit der spektralen Merkmale. Um diese in einem solchen Netz zu fassen, müßte die Kohonenschicht m. E. weiter vergrößert werden. Das übersteigt jedoch die momentanen Möglichkeiten unserer Hardware. Einen weiteren Ansatzpunkt bildet die Reduktion der Merkmale. Hier ließe sich gut ein spezielles neuronales Netz einsetzen. Indem man Grundstrukturen vorgibt, kann man das Netz auf die signifikanten Merkmale trainieren. Setzt man diese beim Lernprozeß eines Kohonen-Netzes ein, müßte sich eine Clusterung bei beherrschbarer Netzwerkgröße ergeben. Eine andere Möglichkeit wäre der Einsatz mehrerer kleiner Netze welche auf einzelne Stoffklassen trainiert wurden.

Die Maße und Methoden, die hier für einen kleinen Datenbestand getestet wurden, müssen ihre Tauglichkeit für eine große Sammlung noch unter Beweis stellen. Die größere Zahl der darin enthaltenen Spektren vergrößert die Wahrscheinlichkeit des Auftretens sehr ähnlicher Spektren. Hier muß sich dann die diskriminatorische Kraft des Korrelationsabstandes erweisen. Eine Verbesserung der Suchergebnisse dürfte auch der Übergang von der Normierung auf die maximale Intensität zur Flächennormierung bringen.

Der unscharfe Linienvergleich hat die Erwartungen vollauf bestätigt. Diese Methode läßt sich sehr genau -es stehen sechs Parameter zur Optimierung zur Verfügung- an die vorhandenen Daten anpassen.

4. Abbildungsverzeichnis

5. Literatur

[1] L. E. Kuenzel: Anal. Chem., 23, 1413 (1951)

[2] Borsdorf, Scholz: "Spektroskopische Methoden in der organischen Chemie". Akademie-Verlag, Berlin 1989

[3] Autorenkollektiv: "Analytikum". VEB Deutscher Verlag für Grundstoffindustrie, 8. Auflage, Leipzig 1971

[4] R. S. McDonald, P. A. Wilks: Appl. Spectroscopy, 42, 151 (1988)

[5] Savitzky, Golay: Anal. Chem., 36, 1627 (1964)

[6] Steinier, Termonia, Deltour: Anal. Chem., 44, 1906 (1972)

[7] J. Zupan: Anal. Chim. Acta, 103, 273 (1978)

[8] P. C. Jurs: Anal. Chem., 43, 364 (1971)

[9] R. H. Shaps, J. F. Sprouse: Eur. Spectrosc. News, 32, 39 (1980)

[10] J. Kwiatkowski: Anal. Chim. Acta, 135, 285 (1982)

[11] T. Blaffert: " Ein Gesamtsystem zur Auswertung von Infrarotspektren durch Lernen und Erkennen spektraler Merkmale sowie Zerlegung uns Synthese chemischer Strukturgraphen" Dissertation Hamburg 1990

[12] H. Bandemer, M. Otto: Mikrochim. Acta, II, 93 (1987)

[13] Munk, Madison, Robb: Mikrochim. Acta, II, 505 (1991)

[14] T. Kohonen: "Self-Organization and Assoziative Memory". Springer Verlag, Berlin 1984

[15] Handbook "Neural Computing, NeuralWorks Professional II /PLUS". NeuralWare Inc., 1990

[16] Lohninger: "Modelle neuronaler Netze", Programmbeschreibung

[17] W. D. Ihlefeldt Programmdokumentation TU München: "Keywords für Clear Text Files". Revision 8 vom 23. August. 1990

[18] Günzler, Böck: "IR-Spektroskopie". Verlag Chemie, 2. Auflage, Weinheim 1983

[19] P. F. Dupuis, A. Dijkstr, J. H. van der Maas: Fresenius Z. Anal. Chem., 291, 27 (1978)

[20] T. Blaffert: Anal. Chim. Acta, 161, 135 (1984)

[21] Doerffel, Eckschlager, Henrion: "Chemometrische Strategien in der Analytik". VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1990

[22] Danzer, Than, Molch, Küchler: "Analytik". Akademische Verlagsgesellschaft Geest & Portig K. G., Leipzig 1987

[23] M.Otto Vorlesungsmanuskript: "Chemische Datenbanken / KI" Sommersemester 1992 Freiberg

[24] Firmenschrift SpecInfo Chemical Conzepts Weinheim 1991

6. Abkürzungen

CAS	Chemical Abstract Services
D	Dimension
E	Energie
FFT	Fast Fourier Transform
GC	Gas Cromatography
h	Plancksches Wirkungsquantum 6.626*10-34 Js
HPLC	High Performance Liquid Chromatography
HWB	Halbwertsbreite
I	Intensität
IR	Infrared
JCAMP-DX	Joint Committee on Atomic and Molecular Physical Data-Data-Exchange
KByte	1024 Byte
KI	Künstliche Intelligenz
LC	Liquid Cromatography
MByte	1024 KByte
MS	Massenspektroskopie
N	Anzahl
NMR	Kernresonanzspektroskopie
P	Probe
R	Referenz
r	Kernabstand
r	Korrelationskoeffizient
TByte	1024 MByte
W	Gewichte eines neuronalen Netzes
τ	Frequenz
μ_s	Valenzschwingung symmetrisch
μ_{as}	Valenzschwingung antisymmetrisch
δ	Deformationsschwingung
λ	Dipolmoment
Z	partielle Ableitung