

Forschung, Wissenschaft

Darstellung chemischer Strukturen für die computergestützte deduktive Lösung chemischer Probleme *

Wolfgang Schubert und Ivar Ugi **

Organisch-Chemisches Institut der Technischen Universität München

Prof. James Dugundji in Verehrung zum 60. Geburtstag (30. August 1979) gewidmet

Abstract

The computer assisted deductive solution of chemical problems based on a mathematical model of constitutional chemistry requires a computer oriented representation of chemical structures. The canonical numbering of a molecule is used to uniquely represent the molecule as a BE-Matrix. As an additional key to the identification of molecules the partitioned molecular formula is derived from the classes of constitutionally equivalent atoms in a molecule. The CIP-rules are examined, whether they are suitable for automatically specifying the stereochemistry of molecules. It is shown, how the canonical numbering of the atoms in a molecule can be used beyond the applicability of the CIP-rules to specify stereochemical properties. The configurational symbols largely correspond to the R,S-nomenclature generated in this manner.

Einführung

Das Erfassen und Auswerten von Messdaten, numerische Rechnungen sowie Dokumentation und Recherche des vorhandenen Wissens sind die heute üblichen Anwendungen von Computern in den Naturwissenschaften. Über diese bekannten Anwendungen hinaus lassen sich Computer jedoch auch für die Simulation komplexer Systeme einsetzen. Hierbei begnügt man sich nicht mit der schnellen und exakten Berechnung irgendwelcher Größen, sondern man betrachtet die Entwicklung der Zustände von ganzen Systemen bei gegebenem Anfangszustand und unter bestimmten Gesetzmässigkeiten. Der so aufgespannte Zustandsraum eines Systems wird dann nach Zuständen abgesehen, welche für eine bestimmte Aufgabenstellung interessieren.

Es hat sich gezeigt, dass diese Vorgehensweise auch für die allgemeine Lösung chemischer Probleme geeignet ist. Entscheidend für die effiziente Suche nach Lösungen im Zustandsraum eines Systems ist die geeignete Formulierung des zu lösenden Problems. Für chemische Fragestellungen kommt es deshalb insbesondere darauf an, die wesentlichen strukturellen Merkmale von Molekülen eindeutig und in einer computergerechten Form darzustellen. Im vorliegenden Artikel sollen

Konzeptionen vorgestellt werden, nach denen Moleküle im Computer so repräsentiert werden, dass eine effiziente, automatisierte Lösung chemischer Probleme möglich ist.

Automatische Lösung chemischer Probleme

Mit der computergestützten Syntheseplanung [1] und der Interpretation von Massenspektren [2] begann die Anwendung von Computern bei der Lösung nicht-numerischer chemischer Probleme. Die ersten Syntheseprogramme stellten die Synthesewege retrosynthetisch, d. h. vom Synthesziel ausgehend, dar und beruhten wesentlich auf Retrieval und Manipulation von Information über bekannte chemische Reaktionen. Bei solchen Programmen werden nach den Strukturmerkmalen eines eingegeben Synthesziels solche Reaktionen ausgewählt, die zu den erkannten Untereinheiten der Struktur des Zielmoleküls führen [3].

Um die inhärenten Beschränkungen eines auf Retrieval und Manipulation beruhenden Vorgehens zu überwinden, benötigt man ein System, das Reaktionen und ihre Umkehrungen simuliert, ganz gleich, ob die Reaktionen bereits bekannt oder noch präzedenzlos sind. In jüngster Zeit wurde gezeigt, dass es möglich ist, auf den Grundlagen eines mathematischen Modells [4] der konstitutionellen Chemie Algorithmen und Computerprogramme zu entwickeln, die ohne Benutzung von Information über individuelle chemische Reaktionen für vorgegebene molekulare Systeme ihre denkbaren Reaktions- und Bildungsweisen angeben können. Dies führt nicht nur zu neuen Perspektiven in der computergestützten Syntheseplanung, sondern es werden auch neue Möglichkeiten zur deduktiven Lösung einer Vielzahl weiterer chemischer Probleme eröffnet [3].

Der zentrale Begriff des verwandten mathematischen Modells ist das Ensemble von Molekülen (EM). Ein EM ist eine Kollektion einer beliebigen Anzahl gleicher oder verschiedener Moleküle. Zwischen EM können die Relationen der Identität und der Isomerie bestehen. EM sind identisch, wenn sie in Anzahl und Art der enthaltenen Moleküle übereinstimmen. EM sind isomer, wenn sie die gleiche Bruttozusammensetzung haben.

* Eingegangen am 5. April 1979

** Korrespondenz-Adresse: Prof. Dr. I. Ugi, Organisch-Chemisches Institut, Technische Universität München, Lichtenbergstrasse 4, 8046 Garching, BRD

Während die Relation der Identität von EM bei der Dokumentation chemischen Wissens im Vordergrund steht, beherrscht die Isomerierelation von EM die dynamischen Aspekte der Chemie. Mit diesem verallgemeinerten Isomeriebegriff können insbesondere chemische Reaktionen als Isomerisierungen von EM beschrieben werden. Stellen H_i und H_j die EM der Edukte und Produkte einer Umsetzung dar, ist die zugehörige Reaktion durch die Isomerisierung



gegeben. Wird das EM H_i wiederum zu einem EM H_k isomerisiert, entsteht die Reaktionsfolge



Beliebige Reaktionsfolgen lassen sich durch



wiedergeben.

Durch automatisches Erzeugen solcher Reaktionsfolgen mit dem Computer können u.a. die Reaktionsmöglichkeiten eines EM aufgezeigt werden. Die jeweiligen Retroreaktionen sind ebenfalls darstellbar, so dass z. B. optimale Synthesewege für ein vorgegebenes Synthesziel auf diese Weise gesucht werden können. Die Darstellung von Reaktionen als Isomerisierungen von EM lässt die Behandlung von mechanistischen Elementarschritten und Übergangszuständen ebenso zu wie die von Reaktionen, welche von stabilen Ausgangsstoffen zu stabilen Endprodukten führen.

Für das automatische Erzeugen von Reaktionsfolgen haben sich solche Isomerisierungen von EM als besonders geeignet herausgestellt, die mechanistischen Elementarschritten von Reaktionen entsprechen. In dem von uns entwickelten Programmsystem ASSOR [5] (Allgemeines Simulationssystem Organischer Reaktionen) werden die EM durch mechanistische Elementarschritte isomerisiert. Durch eine Bewertung der EM nach jeder Isomerisierung kann entschieden werden, ob es sinnvoll ist, bestimmte EM für weitere Isomerisierungen heranzuziehen. Dadurch entsteht eine nur mässig verzweigte Reaktionsfolge, die nur solche Re-

aktionen eines EM enthält, die unter den üblichen Bedingungen möglich erscheinen. Die Arbeitsweise von ASSOR sei am Beispiel des Abbaus von Histidin **1** demonstriert.

Bei seinem Abbau zu N-Formylisoglutamin **2** reagiert Histidin mit zwei Molekülen Wasser und spaltet Ammoniak ab. Das zur Simulation erforderliche Ausgangs-EM besteht folglich aus Histidin und zwei Molekülen Wasser. In dem von ASSOR erzeugten, aus 518 Verbindungen bestehenden Reaktionsnetzwerk ist auch jene Reaktionsfolge enthalten, die dem biochemischen Abbau von Histidin über die Urocainsäure **3**, 4-Hydroxyimidazol-5-propionsäure **4** und 4-Imidazol-5-propionsäure **5** zu N-Formylisoglutamin entspricht.

Darstellung der chemischen Konstitution

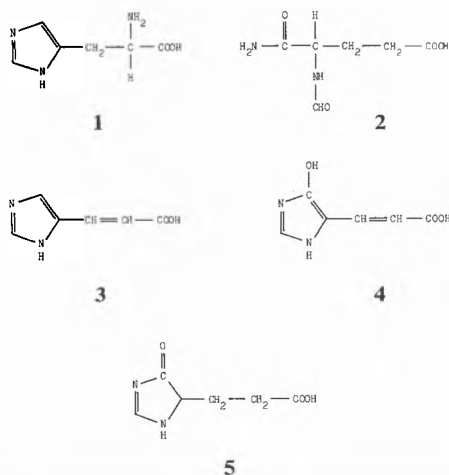
In dem von uns verwandten mathematischen Modell der konstitutionellen Chemie [4] wird die Verteilung der Valenzelektronen in EM durch Bindungs/Elektronen-(BE-)Matrizen dargestellt. Die BE-Matrix eines EM mit n Atomen ist eine symmetrische Matrix der Dimension n mit ganzzahligen Eintragungen. Die Zeilen und Spalten sind den Atomrümpfen der im EM enthaltenen Atome zugeordnet, und zwar gehören die Zeile i und die Spalte i zum Atomrumpf A_i . Die Atomrümpfe sind durch einen der BE-Matrix assoziierten Atomvektor repräsentiert. Die Abb. 1 zeigt eine BE-Matrix und den zugeordneten Atomvektor von Formaldehyd.

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 2 & 1 \\ 0 & 2 & 4 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} \text{H} \\ \text{C} \\ \text{O} \\ \text{H} \end{bmatrix}$$

Abb. 1: BE-Matrix und Atomvektor von Formaldehyd.

Eine ausserdiagonale Eintragung b_{ij} einer BE-Matrix stellt die formale Ordnung der kovalenten Bindung zwischen den Atomen A_i und A_j dar; eine Diagonaleintragung b_{ii} gibt die Anzahl der freien Valenzelektronen des Atoms A_i an. Die Summe aller Eintragungen b_{ij} einer Zeile oder Spalte i ist gleich der Anzahl der Valenzelektronen des Atoms A_i . Die Summe aller Eintragungen einer Zeile i plus einer Spalte i vermindert um die Diagonaleintragung des Atoms A_i ergibt die Anzahl der in den Valenzorbitalen befindlichen Elektronen des Atoms A_i . Die Formalladung des Atoms A_i ergibt sich aus dieser Summe und der Rumpfladung von A_i . Weiterhin ist die Summe aller Eintragungen einer BE-Matrix gleich der Anzahl aller in einem EM vorhandenen Valenzelektronen. Diese Summe hat für alle isomeren EM den gleichen Wert.

Die Darstellung von EM durch BE-Matrizen erfordert die Kennzeichnung eines jeden einzelnen Atoms in den EM, insbesondere auch die formale Unterscheidung chemisch äquivalenter Atome. Hierdurch wird vom



chemischen Standpunkt aus eine nicht erwünschte Mehrdeutigkeit eingeführt, die jedoch wieder vernachlässigt werden muss, wenn EM maschinell identifiziert werden sollen. Da die EM durch die enthaltenen Moleküle gegeben sind, genügt es, das Problem einer eindeutigen Darstellung der chemischen Konstitution für Moleküle zu lösen.

Zwei Moleküle besitzen die gleiche Konstitution, wenn sich ihre BE-Matrizen bis auf Zeilen/Spalten-Permutationen gleichen, denn Zeilen/Spalten-Permutationen einer BE-Matrix entsprechen lediglich der Umnummerierung der Atome von Molekülen. Solche BE-Matrizen, die das gleiche EM repräsentieren, nennen wir äquivalent. Die Tatsache, dass ein Molekül nicht nur durch eine BE-Matrix dargestellt werden kann, hat weitreichende Konsequenzen, sollen Moleküle anhand ihrer BE-Matrizen verglichen bzw. identifiziert werden. Für den Vergleich von BE-Matrizen genügt es nämlich nicht, lediglich die Identität der BE-Matrizen zu prüfen, sondern es muss festgestellt werden, ob sich die BE-Matrizen bis auf Zeilen/Spalten-Permutationen gleichen.

Die beiden nicht identischen BE-Matrizen in Abb. 2 stellen Formaldehyd dar. Dass sie tatsächlich dasselbe Molekül repräsentieren, lässt sich durch Vertauschen der zweiten und vierten Zeile und Spalte der BE-Matrix in Abb. 2 (b) und in der dadurch entstandenen Matrix durch Vertauschen der dritten und vierten Zeile und Spalte zeigen.

$$\begin{array}{cc} \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 2 \\ 0 & 0 & 2 & 4 \end{bmatrix} & \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 2 & 1 \\ 0 & 2 & 4 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \\ \text{(a)} & \text{(b)} \end{array}$$

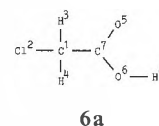
Abb. 2: Äquivalente BE-Matrizen von Formaldehyd.

Um die Äquivalenz von BE-Matrizen festzustellen, könnte man verschiedene Eigenschaften der BE-Matrizen überprüfen, die von Zeilen/Spalten-Permutationen unabhängig sind (Invarianten). Zu diesen Eigenschaften zählen z. B. die Anzahl der Zeilen und Spalten, die Anzahl der von Null verschiedenen Einträge in jeder Zeile und Spalte sowie die Summe aller Einträge. Diesem Verfahren steht allerdings entgegen, dass sich anhand von Invarianten zwar zeigen lässt, dass zwei BE-Matrizen nicht äquivalent sind, umgekehrt sich die Äquivalenz – von Sonderfällen abgesehen – zweier BE-Matrizen aber nicht bestimmen lässt.

Sollen Moleküle anhand ihrer BE-Matrizen verglichen bzw. identifiziert werden, scheiden also die Prüfung auf Identität und der Vergleich von Invarianten aus. Das Permutieren der Zeilen und Spalten einer BE-Matrix ist jedoch zu aufwendig, zumal es für jeden Vergleich wiederholt werden müsste. Eine Lösung dieses Problems besteht darin, die Atome eines Moleküls –

und damit auch die Zeilen und Spalten einer BE-Matrix – einmal nach bestimmten Regeln zu ordnen und dann die BE-Matrizen auf Identität zu prüfen. Die Ordnung der Atome eines Moleküls, die diese Prüfung zulässt, nennen wir kanonisch. Das kanonische Ordnen der Atome eines Moleküls muss folglich immer zu der gleichen BE-Matrix des Moleküls führen. Die Regeln, nach denen die Atome kanonisch geordnet werden, sollten rasch zu einer kanonischen Ordnung führen, leicht formalisierbar und maschinell anwendbar sein. Der beim Chemical Abstracts Service eingesetzte Algorithmus [6] war der erste zur eindeutigen Numerierung der Atome in Molekülen. Er arbeitet insofern ineffektiv, als mehrere Numerierungen parallel entwickelt werden müssen, die schliesslich alle bis auf eine wieder verworfen werden. Das für das Lösen vieler chemischer Probleme wichtige Erkennen der konstitutionellen Äquivalenz von Atomen in einem Molekül ist mit diesem Algorithmus ebenfalls nicht ohne weiteres möglich. Sollte die konstitutionelle Äquivalenz von Atomen mit diesem Algorithmus bestimmt werden, müsste von jedem Atom ausgehend eine Numerierung für das Molekül erzeugt werden. Die Atome, von denen aus die Numerierung begonnen wird, sind dann konstitutionell äquivalent, wenn die aufgrund dieser Numerierung erzeugten Bindungslisten identisch sind. Dieser Weg, die konstitutionelle Äquivalenz von Atomen zu bestimmen, wäre allerdings äusserst aufwendig.

Entsprechend den oben genannten Kriterien haben wir einen Algorithmus CANON [7] entwickelt, der rasch zur kanonischen Ordnung der Atome konvergiert und sowohl von Hand als auch maschinell leicht ausführbar ist. Über das kanonische Numerieren hinaus erkennt der Algorithmus auch konstitutionell äquivalente Atome. Es werden nicht mehrere Numerierungen parallel erzeugt, sondern der Algorithmus konvergiert von vornherein zu nur einer Numerierung. Dass der Algorithmus immer zu einer eindeutigen Bezifferung führt, konnte kürzlich mathematisch bewiesen werden [8]. Der Algorithmus CANON sei am Beispiel der Numerierung von Chloressigsäure 6 demonstriert. Dazu werden die Atome der Chloressigsäure zunächst willkürlich numeriert, vgl. 6a, und anhand dieser Numerierung eine Konnektivitätsliste aufgestellt, vgl. Tabelle 1 (a).



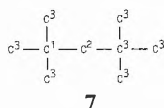
Jedem Atom wird ein Atomindex zugeordnet. Die Atome mit der höchsten Ordnungszahl erhalten den Atomindex 1, die mit der zweithöchsten Ordnungszahl bekommen den Atomindex 2 usw. Die Atomindizes der Nachbaratome eines jeden Atoms werden nach steigenden Werten geordnet und mit dem Atomindex des betrachteten Atoms zum Atomdeskriptor

zusammengefügt, vgl. Tabelle 1 (b). Die Atome werden dann anhand der Atomdeskriptoren lexikographisch geordnet. Dazu werden die Atomindizes in den Atomdeskriptoren von links nach rechts verglichen. Ein Atom kommt in der Ordnung der Atome vor einem anderen, wenn es an einer bestimmten Position im Atomdeskriptor einen kleineren Atomindex besitzt. Sind zwei Atomdeskriptoren verschieden lang, bis zur Länge des kürzeren aber gleich, kommt das Atom mit dem kürzeren Atomdeskriptor vor dem mit dem längeren. Stimmen zwei Atomdeskriptoren in jedem ihrer Atomindizes überein und sind sie gleich lang, lassen sich die zugehörigen Atome nicht ordnen. Die Ordnung der Atome nach ihren Atomdeskriptoren von Tabelle 1 (b) ist in Tabelle 1 (c) gezeigt. Da die Atome noch nicht vollständig geordnet sind, wird ein neuer Zyklus begonnen, indem den Atomen aufgrund der erzeugten Ordnung neue Atomindizes zugeordnet werden. Für die Atome, die noch nicht geordnet werden konnten, werden die Atomdeskriptoren neu aufgestellt und diese Atome danach geordnet. Für die Chloressigsäure stellt man fest, dass durch weitere Zyklen über Tabelle 1 (c) hinaus die Atome (3) und (4) nicht weiter geordnet werden.

Tabelle 1: Erzeugen der kanonischen Numerierung von Chloressigsäure mit dem Algorithmus CANON

1	3,4,2,7	1 - 3:1.3.4.4	2 - 1:	2 - 1:
2	1	2 - 1:3	5 - 2:	5 - 2:
3	1	3 - 4:3	6 - 3:	6 - 3:
4	1	4 - 4:3	1 - 4:	1 - 4:
5	7	5 - 2:3	7 - 5:	7 - 5:
6	7,8	6 - 2:3.4	8 - 6:	8 - 6:
7	1,5 6	7 - 3:2.2.3	3 - 7:4	3 - 7:
8	6	8 - 4:2	4 - 7:4	4 - 8:
(a)	(b)	(c)	(d)	

Können Atome nicht weiter geordnet werden, wird geprüft, ob die Atome durch ihren Ringdeskriptor weiter geordnet werden können. Der Ringdeskriptor wird aus den nach steigenden Werten geordneten Ringgrößen der kleinsten Ringe gebildet, denen das betreffende Atom angehört. Atome, die auch durch Ringdeskriptoren nicht weiter geordnet werden können, sind konstitutionell äquivalent. Konstitutionell äquivalente Atome sind definiert als Atome, die bei einer Abbildung des Moleküls auf sich selbst – ohne Berücksichtigung der stereochemischen Aspekte, jedoch unter Erhalt aller Bindungen – aufeinander abgebildet werden. In 2,2',4,4'-Tetramethylpentan **7** sind z. B. die mit gleichen Zahlen versehenen Atome konstitutionell äquivalent.

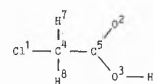


7

Das Erkennen der konstitutionellen Äquivalenz von Atomen eines Moleküls stellt ein wichtiges Zwischen-

resultat des Algorithmus CANON dar, von dem die endgültige Erzeugung einer kanonischen Numerierung abhängt. Um zu einer vollständigen Ordnung der Atome eines Moleküls zu gelangen, wird ein neuer Zyklus des Algorithmus CANON dadurch begonnen, dass in derjenigen Klasse von äquivalenten Atomen, die in der jeweils erreichten Ordnung der Atome am weitesten vorne steht, einem Atom willkürlich ein Atomindex zugeordnet wird, der niedriger ist als der Index der restlichen Atome in der Klasse.

Da die Atome (3) und (4) von **6a** keinem Ring angehören, können sie anhand ihrer Ringdeskriptoren nicht weiter geordnet werden. Die Atome (3) und (4) werden also als konstitutionell äquivalent erkannt. Wird dem Atom (3) daher ein kleinerer Atomindex als dem Atom (4) zugeordnet, kann ein neuer Zyklus begonnen werden. Die Atome von **6a** sind damit aber bereits vollständig geordnet und das Abbruchkriterium des Algorithmus erfüllt. Die kanonische Numerierung **6b** von Chloressigsäure folgt unmittelbar aus den Atomindizes der Tabelle 1 (d).



6b

Für Atome in Ringverbindungen wird gefordert, dass sie nur dann konstitutionell äquivalent sind, wenn sie mit gleichem Atomindex durch die Ringdeskriptoren nicht weiter geordnet werden können. Dieser Schritt sei an der kanonischen Numerierung des Cuneangerüsts **8** näher erläutert.



8a



8b

Tabelle 2: Erzeugen der kanonischen Numerierung des Cuneangerüsts mit dem Algorithmus CANON

1	2,7,8	1 - 1:1.1.1	1 - 3.4.5	1 - 1:1.2.3
2	1,3,6	2 - 1:1.1.1	2 - 4.4.5	3 - 1:1.2.3
3	2,4,5	3 - 1:1.1.1	3 - 3.4.5	5 - 1:1.2.3
4	3,5,8	4 - 1:1.1.1	4 - 3.5.5	7 - 1:1.2.3
5	3,4,6	5 - 1:1.1.1	5 - 3.4.5	4 - 2:1.1.2
6	2,5,7	6 - 1:1.1.1	6 - 4.4.5	8 - 2:1.1.2
7	1,6,8	7 - 1:1.1.1	7 - 3.4.5	2 - 3:1.1.3
8	1,4,7	8 - 1:1.1.1	8 - 3.5.5	6 - 3:1.1.3
(a)	(b)	(c)	(d)	
5 - 1:	5 - 1:	5 - 1:		
1 - 2:2.3.4	3 - 2:	3 - 2:		
3 - 2:1.3.4	1 - 3:3.5.7	7 - 3:		
7 - 2:2.3.4	7 - 3:3.5.6	1 - 4:		
4 - 3:1.2.3	4 - 4:	4 - 5:		
8 - 3:2.2.3	8 - 5:	8 - 6:		
2 - 4:2.2.4	6 - 6:	6 - 7:		
6 - 4:1.2.4	2 - 7:	2 - 8:		
(e)	(f)	(g)		

Ausgehend von der willkürlichen Numerierung des Cuneangerüsts **8a** und den aus den Ordnungszahlen der Atome abgeleiteten Atomindizes werden die Atomdeskriptoren aufgestellt. Alle Atome des Cuneangerüsts haben die gleiche Art und Anzahl von Nachbarn, so dass die Atomdeskriptoren alle gleich lauten, vgl. Tabelle 2(b). Auch ein neuer Zyklus könnte die Atome nicht weiter ordnen. Die Atome werden jedoch durch ihre Ringdeskriptoren weiter geordnet. Diese

Ordnung lässt sich durch neue Zyklen nicht weiter verfeinern, so dass Tabelle 2(d) die Klassen der konstitutionell äquivalenten Atome enthält.

Die kanonische Numerierung des Cuneangerüsts wird erzeugt, indem aus der ersten Klasse von vier äquivalenten Atomen eines einen Atomdeskriptor erhält, der kleiner ist als der der anderen drei Atome. Wählt man Atom (5), lassen sich die Atomdeskriptoren in Tabelle 2(e) und Tabelle 2(f) konstruieren. Es resultiert

(a)		$C_2C_2C_2C_2H_4H_4H_4$	(j)		$C_2CCCCCH_6H_3H_2H_2HH$
(b)		$C_2CCCCCH_6H_3H_2H_2H_2H$	(k)		$C_2C_2C_2CCH_6H_6H_4HH$
(c)		$CCCCCCH_3H_3H_2H_2H_2H_2H$	(l)		$C_2C_2C_2C_2H_6H_6H_4H_2$
(d)		$C_2C_2CCCCCH_6H_4H_3H_2H_2H$	(m)		$C_2CCCCCH_6H_3H_3H_2H_2HH$
(e)		$C_2C_2C_2CCH_6H_4H_4H_3H$	(n)		$C_3C_2CCCH_9H_6H_2H$
(f)		$C_3CCCCCH_9H_3H_2H_2H_2$	(o)		$C_2C_2CCCCCH_6H_6H_3H_2H$
(g)		$C_2CCCCCCH_6H_3H_3H_2H_2H_2$	(p)		$C_3CCCCCH_9H_3H_3H_2H$
(h)		$C_3C_3CCH_9H_6H_3$	(q)		$C_6C_2H_{18}$
(i)		$C_4C_2C_2H_{12}H_4H_2$	(r)		$C_4C_2CCH_{12}H_3H_2H$

Abb. 3: Die Kohlenwasserstoffe C_8H_{18} und ihre partitionierten Bruttoformeln.

schliesslich die kanonische Numerierung **8b**, der die Tabelle 2 (g) zugrunde liegt.

Wählte man ein anderes als das Atom (5), entstünde eine andere, aber vollkommen gleichwertige kanonische Numerierung des Cuneangerüsts, in der jedes Atom genau die gleichen Nachbarn hätte wie in der Numerierung von **8b**. Dies folgt auch unmittelbar aus der Definition der kanonischen Numerierung, nach der die Nachbarn eines jeden Atoms immer die gleichen kanonischen Indices haben müssen.

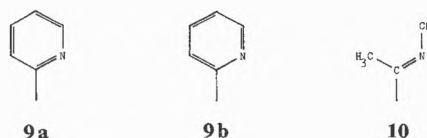
Die Identifizierung anhand ihrer kanonischen BE-Matrizen kann effektiver gestaltet werden, wenn jedem Molekül noch ein Schlüssel zugeordnet wird, unter dem es gesucht werden kann. Im Gegensatz zur kanonischen BE-Matrix braucht dieser Schlüssel nicht eindeutig zu sein. Durch einen Schlüssel wird die Anzahl der tatsächlich anhand ihrer kanonischen BE-Matrizen zu identifizierenden Moleküle auf jene beschränkt, die unter dem gleichen Schlüssel abgelegt sind. Zu solch einem Schlüssel eignet sich die sog. partitionierte Bruttoformel. Sie wird gebildet, indem die Äquivalenzklassen eines Moleküls durch das entsprechende, mit einem Index versehene Elementsymbol bezeichnet werden. Das Elementsymbol gibt die Atomart, der Index die Anzahl der in den Äquivalenzklassen enthaltenen Atome an. Die partitionierte Bruttoformel von n-Octan lautet z. B. $C_2C_2C_2C_2H_6H_4H_4H_4$. Die partitionierte Bruttoformel vereinigt Eigenschaften der Konstitutions- und der Bruttoformel. Sie enthält Angaben über die konstitutionelle Äquivalenz der Atome eines Moleküls, behält aber die Linearität der Bruttoformel bei. Moleküle können daher mittels der partitionierten Bruttoformel weiter unterschieden werden, als es mit der Bruttoformel möglich ist. Von allen 18 Kohlenwasserstoffen der Formel C_8H_{18} , vgl. Abb. 3, können z. B. bis auf das 2,3- und 2,4-Dimethylhexan alle anhand ihrer partitionierten Bruttoformel unterschieden werden. Während der Erzeugung der kanonischen Numerierung mit dem Algorithmus CANON werden die Atome eines Moleküls automatisch in Klassen konstitutionell äquivalenter Atome eingeteilt, so dass die partitionierte Bruttoformel ohne zusätzlichen Aufwand bei der kanonischen Numerierung eines Moleküls miterzeugt wird.

Die Kennzeichnung stereochemischer Merkmale

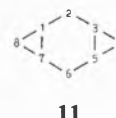
Stereoisomere werden heute im allgemeinen anhand der R,S-Nomenklatur und der daran angelehnten Nomenklaturen bezeichnet. Hierbei bildet eine in mehrere Unterregeln aufgeteilte Sequenzregel von Cahn, Ingold und Prelog (CIP), nach der die Liganden der chiralen Zentren jeweils paarweise verglichen und geordnet werden, das Kernstück [9].

Zieht man die maschinelle Anwendung der CIP-Sequenzregel in Betracht, gilt es zu prüfen, ob es sich bei den CIP-Regeln um ein algorithmisches Verfahren handelt. Danach muss jeder Schritt des Verfahrens aus

dem vorhergehenden Schritt eindeutig bestimmt sein und das Ergebnis nach einer endlichen Anzahl von Schritten vorliegen. Zunächst stellt man fest, dass Definitionen nach CIP wie «übliche» Resonanzformel und «gemittelte» Resonanzformel einer Automatisierung nur sehr schwer zugänglich sind, so dass eine maschinelle Durchführung der CIP-Sequenzregel mit vertretbarem Aufwand nicht möglich erscheint. Will man die CIP-Sequenzregel auf solche Moleküle anwenden, die nicht nur durch eine Valence-Bond-Formel repräsentiert werden können, müssen alle Resonanzformeln des betreffenden Moleküls berücksichtigt werden, um jene, die nach CIP als üblich definiert sind, auswählen zu können. Diese üblichen Resonanzformeln werden durch «Auflösen» der Mehrfachbindungen und «Sättigen» freierwerdender Valenzen expandiert. Durch Mitteilung dieser expandierten Formeln wird schliesslich eine Formel konstruiert, nach der die Liganden geordnet werden. Dieses Verfahren ist nicht nur äusserst aufwendig, sondern es ist auch fehleranfällig, denn die Liganden können nur reproduzierbar geordnet werden, wenn alle üblichen Resonanzformeln einbezogen werden. Wird nur ein Teil dieser Resonanzformeln berücksichtigt, kann die Ordnung der Liganden umgekehrt werden. Nach den CIP-Regeln besitzt z. B. der Ligand **10** eine höhere Priorität als **9**, hingegen **9a** eine höhere als **10** und **9b** eine niedrigere als **10**. Liesse man daher **9b** unberücksichtigt, kehrte sich die Ordnung der Liganden **9** und **10** um.



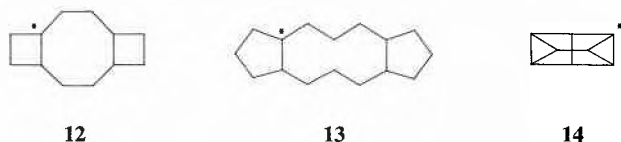
Bei der weitergehenden Prüfung der CIP-Regeln auf die Erfüllung der Kriterien eines algorithmischen Verfahrens zeigt sich, dass es Moleküle gibt, die Chiralitätszentren enthalten, deren Liganden nach der CIP-Sequenzregel jedoch nicht geordnet werden können. Im Tricyclo-[5.1.0.0^{3,5}]-octan **11** besitzt das Atom (1) vier verschiedene Liganden: ein Wasserstoffatom, das tertiäre Kohlenstoffatom (7) und die zwei sekundären Kohlenstoffatome (2) und (8).



11

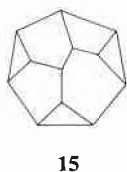
Letztere Atome befinden sich in verschiedener konstitutioneller Umgebung, denn Atom (8) gehört einem Dreiring, Atom (2) jedoch einem Sechsring an. Atome (2) und (8) sind also offensichtlich konstitutionell nicht äquivalent und daher auch die Liganden, die von diesen Atomen ausgehen, verschieden. Nach der CIP-Sequenzregel

sollten die beiden Liganden also zu ordnen sein. Man stellt jedoch fest, dass, gleich wieviele Iterationsschritte durchgeführt werden, die Liganden nicht geordnet werden können. In diesem Falle versagt die CIP-Sequenzregel. Ausser in **11** können z. B. auch die Liganden der jeweils mit * gekennzeichneten Atome in Tricyclo-[8.2.0.0^{4,7}]-dodecan **12**, Tricyclo-[11.3.0.0^{5,9}]-hexadecan **13** und Pentacyclo-[5.1.0.0^{2,6}.0^{3,5}.0^{4,8}]-octan (Cunean) **14** nicht vollständig geordnet werden.



Die Moleküle **11** bis **14** enthalten konstitutionell äquivalente Atome. Allerdings ist die Konstitution dieser Moleküle sowohl mit achiralen als auch mit chiralen Strukturen vereinbar, die eine Spezifikation der Konfiguration verlangen. In den angeführten Fällen ist es jedoch nicht möglich, den Molekülen Konfigurationsymbole nach CIP zuzuordnen, da die konstitutionell verschiedenen Liganden nicht in einer endlichen Anzahl von Schritten geordnet werden können.

Das Cunean **14** enthält ausschliesslich tertiäre Kohlenstoffatome, die untereinander teilweise äquivalent sind. Es ist jedoch denkbar, dass ein Molekül mit nur tertiären Gerüstkohlenstoffatomen konstitutionell vollkommen asymmetrisch ist, d. h. keines der Atome ist dem anderen äquivalent. Ein solches Molekül wäre z. B. **15**.



In diesem hypothetischen Heptacyclododecan ist jedes Atom von anderen unter sich verschiedenen Liganden umgeben. Für keines der Atome ist es jedoch möglich, die Liganden nach der CIP-Sequenzregel zu ordnen, denn jedes Kohlenstoffatom ist von drei Kohlenstoffatomen und einem Wasserstoffatom umgeben, wie weit auch immer man sich von einem Zentralatom entfernt. Das Versagen der CIP-Sequenzregel beruht auf der ihr zugrunde liegenden Annahme, dass die Liganden immer nach einer endlichen Anzahl von Schritten zu ordnen sind. Diese Annahme trifft, wie obige Beispiele zeigen, offensichtlich nicht zu. Der CIP-Sequenzregel fehlt somit das Kriterium der Endlichkeit, eine der Grundbedingungen eines algorithmischen Verfahrens. Die CIP-Regeln sind deshalb nicht nur für die maschinelle Durchführung, sondern auch prinzipiell nicht zum Ordnen von Liganden geeignet.

Diese Tatsache macht es erforderlich, ein algorithmi-

sches Verfahren zu definieren, dass zum Ordnen von Liganden uneingeschränkt geeignet ist. Wir erreichen dies durch folgende

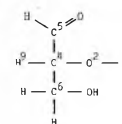
Definition

Die Ordnung von Liganden in einem Molekül ist durch die Ordnung der kanonischen Indizes ihrer α -Atome gegeben.

Durch diese Definition der Ordnung von Liganden über die kanonische Numerierung des Moleküls wird die Beschreibung der Konstitution und der stereochemischen Merkmale auf eine gemeinsame Basis gestellt.

Das Ordnen der Liganden ist demnach nicht nur nach CIP oder ähnlichen Sequenzregeln möglich, sondern auch anhand der kanonischen Numerierung der Atome eines Moleküls. Da Dokumentation in der Chemie heute fast ausschliesslich maschinell betrieben wird, zu speichernde Moleküle also kanonisch numeriert werden müssen, ist es sinnvoll, auch Konfigurationsymbole ohne zusätzliche Prozedur, wie die Anwendung der CIP-Sequenzregel, zu bestimmen. Der Algorithmus CANON eignet sich besonders für diesen Zweck, denn er ist – aufgrund seiner wenigen, klaren Regeln – für die maschinelle Ausführung geeignet und führt fast immer zu den durch die R,S-Nomenklatur eingeführten Konfigurationssymbolen. Verzichtet man auf die Kompatibilität mit eingeführten Nomenklaturesystemen, kann grundsätzlich jede eindeutige Numerierung eines Moleküls zur Kennzeichnung der Stereochemie herangezogen werden [10].

Obwohl der kanonischen Numerierung der Atome eines Moleküls die konzeptionelle Aufteilung des Moleküls in Liganden fremd ist, lässt sich das Konzept der Liganden beibehalten und von einer Ordnung der Liganden sprechen, wenn die Ordnung der Liganden durch die kanonischen Indizes ihrer α -Atome festgelegt wird, d. h. der Atome der Liganden, über die sie an das chirale Zentralatom oder polyzentrische Molekülgerüst direkt gebunden sind. Auf diese Weise ergibt sich aus der Ordnung $2 < 5 < 6 < 9$ der kanonischen Indizes der α -Atome im Glycerinaldehyd **16** die Ordnung der Liganden zu $\text{OH} < \text{CHO} < \text{CH}_2\text{O} < \text{H}$.



16

Im Glycerinaldehyd ist nur ein chirales Zentrum enthalten. Sind in einem Molekül mehrere vorhanden, wird beim Ordnen der Liganden für jedes chirale Zentrum so verfahren, wie es am Glycerinaldehyd gezeigt wurde. Hier manifestiert sich eine bedeutende Eigenschaft des Ordners von Liganden nach kanonischen Indizes der α -Atome von Liganden. Für das

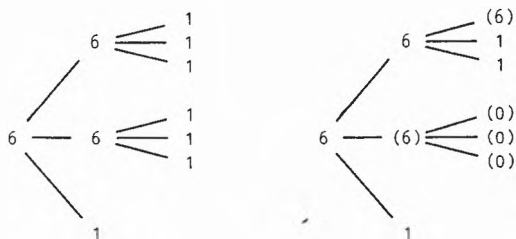


Abb. 5: CIP-Schema zur Ordnung der Liganden $-\text{CH}(\text{CH}_3)_2$ und $-\text{CH}=\text{CH}_2$.

Auch die Ordnung der Liganden **18** bis **21** nach der CIP-Sequenzregel entspricht der Ordnung nach kanonischen Indizes der α -Atome. Dies scheint allgemein für Resonanzstrukturen zuzutreffen, die Kohlenstoffatome im Gerüst enthalten, obwohl Doppelbindungen nach der CIP-Sequenzregel aufgelöst werden. Der Algorithmus CANON umgeht Resonanzformeln, indem die konstitutionellen Aspekte eines Moleküls nur durch die Koordinationszahlen der Atome berücksichtigt werden. Nicht die Anzahl der Bindungen, die von einem Atom ausgehen, wird primär herangezogen, sondern die Anzahl der Nachbarn ist die primär verwandte Grösse. Mehrfachbindungen gehen nur indirekt über die Koordinationszahl in den Algorithmus CANON ein.

Während beim Ordnen von Liganden in den Molekülen **11** bis **15** die CIP-Sequenzregel versagt, ist es mittels kanonischer Indizes der α -Atome der betreffenden Liganden hingegen ohne weiteres möglich. Bezeichnet man z. B. in **12** die Klassen der konstitutionell äquivalenten Atome durch jeweils gleiche Ziffern, die ja ein Zwischenergebnis von CANON sind, vgl. **24**, erkennt man sofort, dass die tertiären Kohlenstoffatome von konstitutionell verschiedenen Atomen umgeben sind. Da durch die kanonische Numerierung der Atome die Ordnung der Äquivalenzklassen erhalten bleibt, geben die Ziffern in **24** auch die Ordnung der Liganden an.



24

Zusammenfassung

Die deduktive Lösung chemischer Probleme mit dem Computer auf der Basis eines mathematischen Modells der konstitutionellen Chemie erfordert eine computergerechte Darstellung chemischer Strukturen. Die kanonische Numerierung eines Moleküls wird zur eindeutigen Darstellung des Moleküls als BE-Matrix herangezogen. Als zusätzlicher Schlüssel zur Identifizierung der Moleküle wird über die Klassen der konstitutionell äquivalenten Atome eines Moleküls die partitionierte Bruttoformel eingeführt. Die CIP-Regeln werden auf ihre Verwendbarkeit zur automatischen Kennzeichnung der Stereochemie von Molekülen hin untersucht. Es wird gezeigt, wie mit der kanonischen Numerierung der Atome eines Moleküls über die Anwendbarkeit der CIP-Regeln hinaus stereochemische Merkmale bezeichnet werden können. Die so erzeugten Konfigurationssymbole entsprechen weitgehend der R,S-Nomenklatur.

Literaturverzeichnis

- 1a I. Ugi und G. Kaufhold: (unveröffentlicht 1966–67), beschrieben in I. Ugi Rec. Chem. Prog., 30 (1969) 289.
- b E. J. Corey, W. T. Wipke, R. D. Cramer and W. J. Howe. J. Amer. Chem. Soc., 94 (1972) 421.
- c W. T. Wipke, in W. T. Wipke, S. Heller, R. Feldmann and E. Hyde: «Computer Representation and Manipulation of Chemical Information», Wiley, New York 1974.
- d H. Gelernter, N. S. Sridharan, A. I. Hart, S. C. Yen, F. W. Fowler and H. I. Shue: Top. Curr. Chem., 41 (1973) 113–150.
- e H. L. Gelernter, A. F. Sanders, D. L. Larsen, K. K. Agarwal, R. H. Boivie, G. A. Spritzer and J. E. Searleman: Science, 4308 (1971) 1041.
- 2a J. Lederberg, G. L. Sutherland, B. G. Buchanan, E. A. Feigenbaum, A. V. Robertson, A. M. Duffield and C. Djerassi: J. Am. Chem. Soc., 91 (1969) 2973–2976;
- b ibid., 91 (1969) 2977–2981.
- c B. R. Kowalski and C. A. Reilly: J. Phys. Chem., 75 (1971) 1402–1411.
- 3a I. Ugi: Intra-Sci. Chem. Rep., 5 (1971) 229.
- b I. Ugi and P. D. Gillespie: Angew. Chem., 83 (1971) 980 bis 981;
- c ibid., 83 (1971) 982–985.
- d I. Ugi, P. Gillespie and C. Gillespie: Trans. N. Y. Acad. Sci., 34 (1972) 416–432.
- e J. Blair, J. Gasteiger, C. Gillespie, P. Gillespie and I. Ugi in W. T. Wipke, S. Heller, R. Feldmann and E. Hyde: «Computer Representation and Manipulation of Chemical Information», Wiley, New York 1974, pp. 129–145.
- f I. Ugi: IBM Nachrichten, 24 (1974) 180–184.
- g I. Ugi, J. Gasteiger, J. Brandt, J. Brunner und W. Schubert: IBM Nachrichten, 24 (1974) 185–189.
- h J. Brandt, J. Friedrich, J. Gasteiger, C. Jochum, W. Schubert, and I. Ugi in E. V. Ludena, N. H. Sabelli and A. C. Wahl: «Computers in Chemical Education and Research», Plenum Publishing Corp., New York 1977, pp. 337–355.
- i J. Brandt, J. Friedrich, J. Gasteiger, C. Jochum, P. Lemmen, W. Schubert and I. Ugi: Pure Appl. Chem., 50, (1978) 1303–1318.
- j I. Ugi, J. Bauer, J. Brandt, J. Friedrich, J. Gasteiger, C. Jochum und W. Schubert: Angew. Chem., 91 (1979) 99 bis 111; Angew. Chem. Int. Ed. Engl., 18 (1979) 123–147.
- k A. Weise: Z. Chem., 13 (1973) 155–156; ibid. 15, (1975) 333–340.
- 4 J. Dugundji and I. Ugi: Top. Curr. Chem., 39 (1973) 19–64.
- 5 W. Schubert: MATCH, im Druck.
- 6 H. L. Morgan: J. Chem. Doc., 5 (1965) 107–113.
- 7 W. Schubert and I. Ugi: J. Am. Chem. Soc., 100 (1978) 37–41.
- 8 J. Dugundji, W. Schubert and I. Ugi: in Vorbereitung.
- 9a R. S. Cahn and C. K. Ingold: J. Chem. Soc. (London) (1951) 612–622.
- b R. S. Cahn, C. K. Ingold and V. Prelog: Experientia, 12 (1956) 81–124.
- c R. S. Cahn, C. K. Ingold und V. Prelog: Angew. Chem., 78 (1966) 413–447.
- 10a H. W. Davis: «Computer Representation of the Stereochemistry of Organic Molecules», Birkhäuser-Verlag, Basel und Stuttgart 1976.
- b W. T. Wipke and T. M. Dyott: J. Am. Chem. Soc., 96 (1974) 4825–4834; ibid., 96 (1974) 4834–4842.
- 11 J. Blair, J. Gasteiger, C. Gillespie, P. D. Gillespie and I. Ugi: Tetrahedron, 30 (1974) 1845–1859.