

Kombinatorische Chemie, eine Herausforderung für Mathematik und Informatik

Ralf Gugisch, Adalbert Kerber, Reinhard Laue,
Markus Meringer, Christoph Rücker

25. September 2001

Abstract

In den letzten Jahren sind in der Chemie ganz neue Methoden der *Massensynthese*¹ entwickelt worden, die mittlerweile ins Zentrum des Interesses und der Diskussion gerückt sind. Sie werden u.a. im pharmazeutischen Bereich und in den Materialwissenschaften angewendet und unter der Überschrift *kombinatorische Chemie* zusammengefaßt. Es geht dabei um die durch Roboter unterstützte und gesteuerte Erzeugung von Hunderten oder gar Tausenden chemischer Substanzen "auf einen Schlag".

Schon der Name *kombinatorische Chemie* läßt Mathematiker und Informatiker aufhorchen. Er zeigt deutlich, daß begleitende mathematisch-kombinatorische Überlegungen im Vorfeld solcher Experimente von Nutzen sein sollten. Da es sich dabei um Massensynthese handelt, sieht der Informatiker sofort, daß bei Experimenten der kombinatorischen Chemie eine riesige Datenflut anfallen wird, die es zu strukturieren gilt.

Wir wollen deshalb hier unsere Erfahrungen schildern, die wir im Rahmen zweier Projekte gemacht haben, und wir werden einige der Probleme aufzeigen, die wir in diesem Zusammenhang lösen wollen. Es handelt sich dabei um ein DFG-Projekt zur Materialforschung (gemeinsam mit Prof. Ziegler, FAN) von Precursorkeramiken, und ein BMBF-Projekt (Federführung Fa. Henkel) zur Optimierung und Analyse solcher Verfahren.

1. Kombinatorische Chemie

Die Entwicklung schneller technischer Screeningmethoden hat neuerdings die Möglichkeit eröffnet, in kurzer Zeit Tausende chemischer Substanzen beispielsweise auf ihre medizinisch-biologischen Eigenschaften zu testen und herauszufinden, ob sie für einen vorgegebenen medizinisch-biologischen (oder auch materialwissenschaftlichen) Zweck geeignet sind oder nicht [1,2]. Man kann deshalb jetzt auch Produkte von Massensynthesen - die sich ja wesentlich durch ihre Vielfalt von den bisherigen Einzelsynthesen unterscheiden - schnell nach medizinisch wirksamen Substanzen durchsuchen. Entsprechende Automaten sind mittlerweile in großer Vielfalt erhältlich.

Interessant für Mathematiker und Informatiker an dieser Neuentwicklung ist natürlich vor allem die Tatsache, daß solche Anlagen *programmgesteuert* sind. Es ist eine Herausforderung sowohl für Mathematiker als auch für Informatiker, die dabei auftretenden Probleme zu behandeln und zu lösen. Es geht dabei neben *der Steuerung der Automaten* vor allem um die Frage, wie man solche Experimente der kombinatorischen Chemie einerseits *im Vorhinein planen* und andererseits *im Nachhinein analysieren* kann. Im Vorhinein will man natürlich wissen, ob eine geplante Massensynthese überhaupt eine interessante Substanz liefern kann, und wie man dies auf möglichst sparsame Weise erreicht. Im Nachhinein ist das Ergebnis des Experiments auf eine geeignete Substanz hin zu durchsuchen. Die Geschichte der Entwicklung von Massensynthesen findet man beispielsweise in [3,4] von I. Ugi, einem Protagonisten der kombinatorischen Chemie.

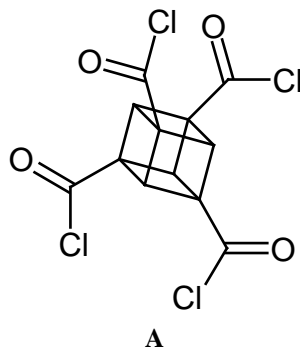
Das Ziel einer kombinatorischen Synthese ist eine sogenannte *kombinatorische Bibliothek* von Molekülen, d.h. eine kontrolliert (und in der Regel automatisch) hergestellte Ansammlung chemischer Substanzen. *Kontrolliert* heißt dabei, daß man die Ausgangsprodukte und die möglichen Reaktionen genau kennt, so daß unter kontrollierten Bedingungen gezielt ganz bestimmte Verbindungen entstehen.

Hierfür gibt es selbstverständlich diverse Verfahren, Mehrschritt- und Einzelschrittmethoden usw. In jedem Fall geht man von mehreren *Grundbestandteilen* und *Reaktionen* zwischen diesen aus. Die Ergebnisse werden nach dem Experiment auf Substanzen durchsucht (in modernem Denglisch: gescreent), welche die gewünschten Eigenschaften haben.

¹ Unter Massensynthese verstehen wir hier die schnelle Synthese sehr vieler Substanzen.

2. Ein konkretes Beispiel

Betrachten wir einmal ein bekanntes Beispiel aus der Literatur [5,6], die Amidierung eines mehrfachen Säurechlorids, etwa des folgenden Cubanderivats **A**:



Man erkennt vier *reaktive Stellen* am Cuban, nämlich die vier Eckpunkte, an denen ein Sauerstoff- und ein Chloratom angezeigt sind. Hier können Amine, und damit auch Aminosäuren angreifen. Man kann sich leicht ein Experiment der kombinatorischen Chemie vorstellen, bei dem Moleküle **A** mit verschiedenen Aminosäuren in Kontakt gebracht werden, so daß Reaktionen stattfinden können. Nehmen wir einmal an, hierzu stehen 20 verschiedenen Aminosäuren zur Verfügung.

Schon ein erster Blick zeigt, daß hierbei wohl recht viele verschiedene Moleküle herauskommen können, die kombinatorische Bibliothek wird also recht groß werden. Frage an den Mathematiker und an den experimentierenden Chemiker: Wie groß kann sie maximal werden? Um diese Frage zu beantworten, muß man die *Symmetriegruppe* des Zentralmoleküls betrachten, die aus den zwölf eigentlichen Drehungen besteht, die das Molekül **A** in sich selbst überführen. Sie zeigt nämlich, mit Hilfe einer geeigneten mathematischen Formel [7], daß die $20 \cdot 20 \cdot 20 \cdot 20 = 160.000$ verschiedenen Reaktionsverläufe zu genau 13.700 verschiedenen Molekülen führen. Das ist die Größe der maximal zu erwartenden kombinatorischen Bibliothek, nennen wir diese einmal die *virtuelle Bibliothek*. Die Bibliotheksgrößen bei 1, 2, 3, ..., 17, 18, 19 Aminosäuren sind entsprechend

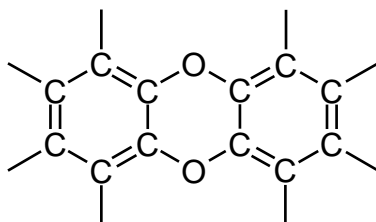
$$1, 5, 15, \dots, 7225, 9045, 11191$$

(vgl. [8,9]). Ein Chemiker kann sich natürlich nicht mit der Größe der virtuellen Bibliothek begnügen. Er wird die Bibliothek selbst brauchen, d.h. zumindest die Strukturformeln der 13.700 Moleküle. Diese Aufgabe stellt eine weitere Herausforderung an Mathematik und Informatik dar, deren Lösung wir uns jetzt zuwenden werden.

3. Die Erzeugung der kombinatorischen Bibliothek

Hierzu verwenden wir ein Verfahren, das in allen Fällen, die in [6] angegeben sind, und in vielen weiteren Beispielen benutzt werden kann. Tatsächlich läßt es sich auch auf scheinbar ganz andere Probleme anwenden, beispielsweise auf die Konstruktion der 22 *Permutationsisomere* des Tetrachlordibenzodioxins, oder auf die Konstruktion fehlerkorrigierender Linearer Codes, kombinatorischer Designs, Spinkonfigurationen und zahlreicher anderer Strukturen aus der Mathematik und den Naturwissenschaften (vgl. [10,11,12]).

Man bedient sich dabei algebraisch-kombinatorischer Methoden, die in der Mathematik klassisch sind, und auf die beispielsweise E. Ruch und Mitarbeiter im Zusammenhang mit der Konstruktion von Permutationsisomeren hingewiesen haben [13,14]. Die Situation ist hier ganz ähnlich der in der kombinatorischen Chemie bei der Konstruktion von Bibliotheken. Sehen wir uns zur Illustration einmal das *Skelett* des Dioxins an:

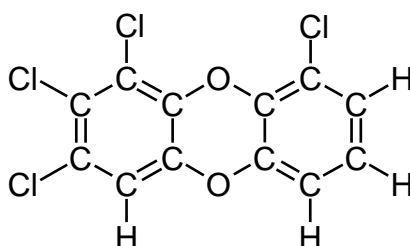


Man sieht hier 8 freie Plätze, auf die 4 Wasserstoff- und 4 Chloratome auf alle möglichen Weisen verteilt werden sollen. Auch hier ist die Symmetriegruppe dieses Skeletts zu berücksichtigen, Mathematiker bezeichnen sie mit V_4 und nennen sie die Kleinsche Vierergruppe. Es gibt jetzt eine weitere Gruppe, die meist als S_8 bezeichnet

wird und diese 8 freien Plätze untereinander vertauscht. Darüber hinaus gibt es noch die Gruppe $S_4 \oplus S_4$, welche die 4 Chloratome untereinander und ebenso die Wasserstoffatome untereinander vertauscht. E. Ruch hat gezeigt, daß die Permutationsisomere des Dioxin umkehrbar eindeutig den $(V_4, S_4 \oplus S_4)$ in S_8 zugeordnet werden können, die Menge dieser Isomere also der Menge von Doppelnebenklassen

$$V_4 \setminus S_8 / S_4 \oplus S_4$$

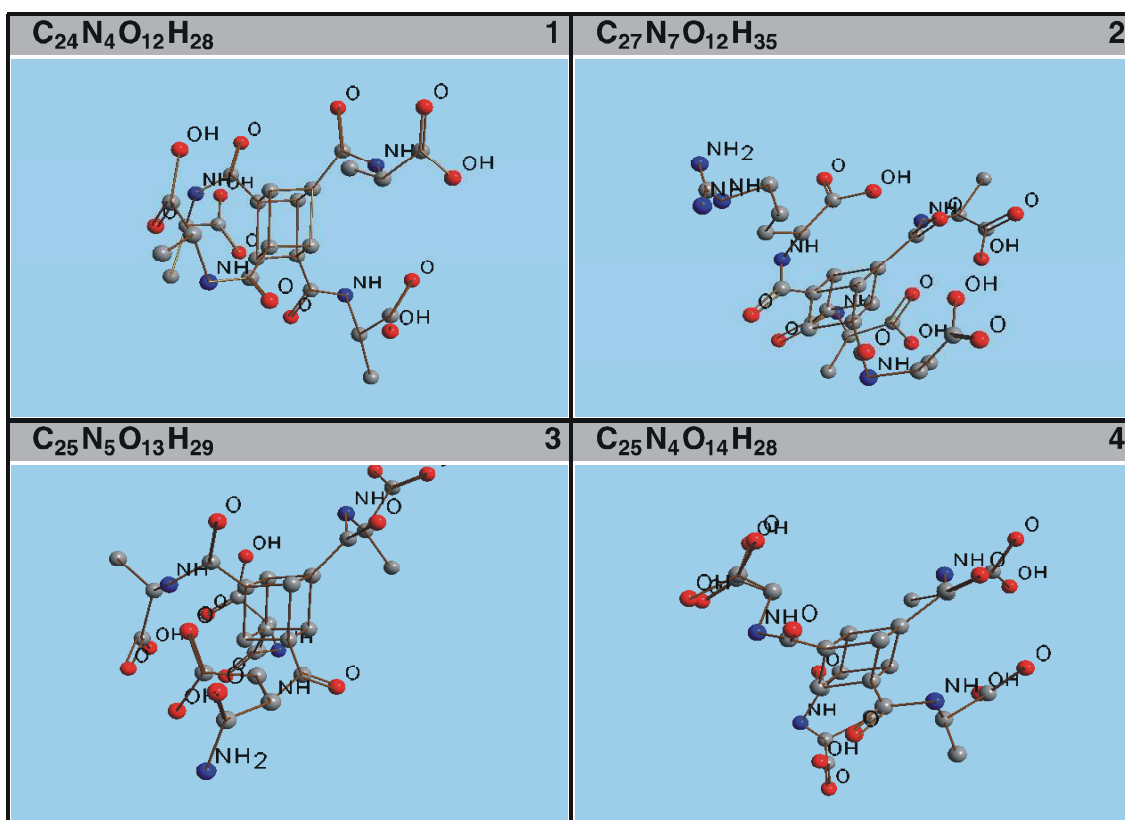
zugeordnet werden können. Man kann zeigen, daß diese Menge aus genau 22 Doppelnebenklassen besteht, und aus Repräsentanten von diesen kann man schließlich die Strukturformeln dieser 22 Moleküle gewinnen. Hier ist eine von diesen:



Zur Konstruktion von Doppelnebenklassen stehen uns verschiedene Algorithmen [15,16] zur Verfügung. Für diese und andere Fälle haben wir ein Softwarepaket [17] entwickelt, eine Anpassung von MOLGEN [18] an die Bedürfnisse der kombinatorischen Chemie, das kombinatorische Bibliotheken generieren kann. Es heißt

MOLGEN-COMB

Hier ist ein Ausschnitt der damit generierten kombinatorischen Bibliothek von Amiden des Cubanderivats **A**:



4. Optimierung kombinatorischer Experimente

Wir sind also in der Lage, für viele Experimente der kombinatorischen Chemie im Vorhinein die Strukturformeln der möglichen Resultate effizient und redundanzfrei zu ermitteln. Diese virtuelle Bibliothek

besteht oft aus vielen Zehntausenden von Strukturformeln, und es ist klar, daß man auch aus Kostengründen nicht alle synthetisieren will, sondern nur solche, bei denen die Chance gut ist, daß sie die gewünschten Eigenschaften haben.

Man kann hierbei wie folgt vorgehen: Man vergleicht die virtuelle Bibliothek mit einer Teilmenge dieser Strukturformeln, für die man schon weiß, ob diese Moleküle für den gewünschten Zweck geeignet sind oder nicht. Nennen wir diese die *reale Bibliothek*. Die Anwendung anspruchsvoller Methoden aus der Stochastik erlaubt nun, mit Hilfe *molekularer Deskriptoren* und Methoden der *Regressionsanalyse* eine Prognose zu wagen, welche Moleküle aus der virtuellen Bibliothek vermutlich ebenfalls geeignet sein können. Hat man die richtigen Deskriptoren zur Hand (genauer: Computerprogramme, die deren Werte für die Moleküle aus der realen und aus der virtuellen Bibliothek berechnen können), dann hat man hier eine gute Chance, mit deutlich größerer Wahrscheinlichkeit als in einer beliebigen Stichprobe, "gute" Moleküle zu finden. Auf diese Weise kann man Experimente der kombinatorischen Chemie planen und optimieren.

Aber auch hier treten sehr ernsthafte Probleme auf, die bisher noch nicht zufriedenstellend gelöst sind - vielleicht auch deshalb noch nicht, weil sie erst durch die kombinatorische Chemie ins Zentrum des Interesses rücken. Eines dieser Probleme ist die sogenannte *Nomenklaturfrage*. Sie ergibt sich, wenn man die Moleküle aus der realen Bibliothek in der virtuellen Bibliothek wiederfinden will. Das ist ganz und gar nicht einfach, denn diese Moleküle sind als Datenfiles gespeichert, die man vergleichen muß. Das erfordert die Umformung von Molekülbeschreibungen in standardisierte Files. Mathematiker und Informatiker kennen dieses Problem aus vielen anderen Situationen, sie nennen es das *Normalformenproblem*. Wir lösen es allgemein für den Fall diskreter Strukturen.

Die ersten Schritte sind dabei bereits gemacht, wir haben einen *Kanonisierer* implementiert, der zumindest auf der *topologischen Ebene* Molekülfiles auf Normalform bringen kann. Ein Kanonisierer auf der deutlich anspruchsvolleren *geometrischen Ebene* ist in Arbeit und wird in wenigen Wochen zur Verfügung stehen.

Bisher konnte man dieses Problem z.B. durch Ermittlung der sogenannten CAS-Nummer einer jeden Substanz lösen, oder durch Konstruktion von systematischen Namen. Beide Wege sind zumindest sehr kostspielig und führen oft überhaupt nicht zum Erfolg: CAS-Nummern gibt es nur für bereits synthetisierte Substanzen, nicht für Tausende *neuer* chemischer Substanzen. Systematische Namen lassen sich für viele Substanzklassen bisher noch nicht per Computer erzeugen.

Das Projekt ist einerseits natürlich für den Einsatz in der Materialforschung interessant. Andererseits liefert es für Mathematik und Informatik wertvolle richtungsweisende Anregungen. So wird etwa das Gewicht der Geschwindigkeit bei den Algorithmen reduziert durch die Anforderung, besonders flexibel Nebenbedingungen berücksichtigen zu können. Dies erfordert neue Strategien beim Algorithmenentwurf.

Literatur

- [1] B. Eickhoff: Kombinatorische Chemie. *Biotec*, **5-6** (2001), 39-42.
- [2] M. J. Plunkett und J. A. Ellman: Neue Wirkstoffe durch Kombinatorische Chemie. *Spektrum der Wissenschaft*, Spezial **6** (1997), 28-35.
- [3] I. Ugi, A. Dömling, B. Gruber, M. Heilingbrunner, C. Heiss, W. Hörl: Formale Unterstützung bei Multikomponentenreaktionen - Automatisierung der Synthesechemie. *Software-Entwicklung in der Chemie*, **9** (1995), 113-128.
- [4] A. Dömling, I. Ugi: Multikomponentenreaktionen mit Isocyaniden. *Angew. Chem.*, **112** (2000), 3300-3344.
- [5] T. Carell, E. A. Wintner, A. Bashir-Hashemi, J. Rebek, Jr.: Neuartiges Verfahren zur Herstellung von Bibliotheken kleiner organischer Moleküle. *Angew. Chem.*, **106** (1994), 2159-2162.
- [6] T. Carell, E. A. Wintner, A. J. Sutherland, J. Rebek, Jr., Y. M. Dunayevskiy und P. Vouros: New promise in combinatorial chemistry: synthesis, characterization, and screening of small molecule libraries in solution. *Chem. & Biol.*, **2** (1995), 171-183.
- [7] G. Pólya: Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen. *Acta mathematica*, **68** (1937), 145-253.
- [8] T. Wieland: Mathematical Simulations in Combinatorial Chemistry. *MATCH*, **34** (1996), 179-206.
- [9] T. Wieland: Combinatorics of Combinatorial Chemistry. *J. Math. Chem.*, **21** (1997), 41-157.
- [10] A. Betten, H. Friepfänger, A. Kerber, A. Wassermann, K.-H. Zimmermann: Codierungstheorie - Konstruktion und Anwendung Linearer Codes. *Springer Verlag* (1998).
- [11] A. Kerber: Algebraic Combinatorics Via Finite Group Actions. *Wissenschaftsverlag*, Mannheim, Wien, Zürich (1991).
- [12] R. Laue: Construction of combinatorial objects - a tutorial. *Bayreuther Mathematische Schriften*, **43** (1993), 53-96.
- [13] E. Ruch, W. Hässelbarth, B. Richter: Doppelnebenklassen als Klassenbegriff und Nomenklaturprinzip für Isomere und ihre Abzählung. *Theoretica Chimica Acta*, **19** (1970), 288-300.

- [14] E. Ruch, D. J. Klein: Double cosets in chemistry and physics. *Theoretica Chimica Acta*, **63** (1983), 447-472.
- [15] T. Grüner: Strategien zur Konstruktion diskreter Strukturen und ihre Anwendung auf molekulare Graphen. *MATCH*, **39** (1999), 39-126.
- [16] B. Schmalz: Verwendung von Untergruppenleitern zur Bestimmung von Doppelnebenklassen. *Bayreuther Mathematische Schriften*, **31** (1990), 109-143.
- [17] R. Gugisch, A. Kerber, R. Laue, M. Meringer, J. Weidinger: MOLGEN-COMB, a Software Package for Combinatorial Chemistry. *MATCH*, **41** (2000), 189-203.
- [18] R. Grund, A. Kerber, und R. Laue: MOLGEN, ein Computeralgebra-System für die Konstruktion molekularer Graphen. *MATCH*, **27** (1992), 87-131.