Mathematics for Combinatorial Chemistry

Th. Grüner, A. Kerber, R. Laue, M. Meringer

Department of Mathematics
University of Bayreuth
D-95440 Bayreuth, Germany

Abstract

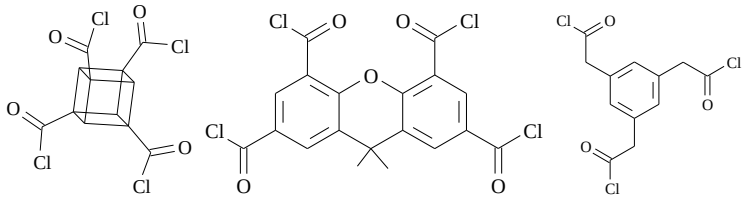
Some of the mathematical methods will be described which are implemented in the software package

MOLCOMB[1]

that allows to simulate combinatorial chemistry by generating combinatorial libraries and to do screening according to geometric substructures.

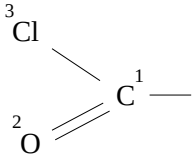
1 Combinatorial Chemistry

To begin with, we consider the examples described in the prominent papers [1] and [2] on combinatorial chemistry. The authors introduce particular combinatorial libraries obtained by starting from the *central molecules*

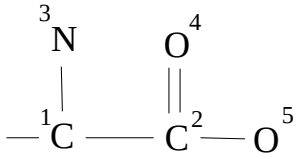


i.e. they start from cubane, xanthene, benzene triacid chlorine as central molecules to which they attach amino acids according to the *reaction scheme* that describes the reaction between the active sites

[1]MOLCOMB is the outcome of a research project supported by the federal ministery for technology, under contract 03 KE7BA 1-4



of the central molecule and the active parts of the aminoacids in question:



Here is the *reaction scheme:*

$$\rho = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -\infty & -\infty & -\infty & -\infty & -\infty \end{pmatrix}.$$

Atom i of the central molecule's active site will be connected to atom j of the amino acid by a bond of multiplicity $\rho_{i,j}$.
It is a variation of the Ugi's be&r-matrices. The aim is to describe the combinatorial libraries that arise from the given central molecule by reaction with a prescribed set of amino acids.

2 A mathematical model

A mathematical model for that situation is well known since Pólya ([3]). It is in fact the very same model that applies to the description of *permutational isomers* corresponding to a given skeleton and a prescribed set of ligands. We introduce two sets:

- S, the set of *active sites* of the central molecule,
- A, the prescribed set of *building blocks*, which are in our example, a set of amino acids.

A reaction of the central molecule with the amino acids is, in mathematical terms, a mapping from S into A, i.e. an element of the following set of all such mappings:

$$A^S := \{f : S \to A\}.$$

For example, if we allow 20 amino acids to react, then there are $20^4 = 160000$ mappings in the case of the cubane as well as in the case of the xanthen or the benzene triacid chlorine.

These 160 000 molecules (better say: the corresponding molecular graphs) are easily generated, but we definitely should *not* put them into a big data bank and do screening immediately, since they are in fact too many, as *not all of them are essentially different* and we should definitely try to save space, since combinatorial libraries can be very big. The reason why not all of them are different is the *symmetry group* of the central molecule. In case of the cubane it is the group T_d, in case of the xanthen it is C_{2v} while the symmetry group of the triacid is C_{3v}.

Hence what we are really after is the set of *orbits* of these groups on the set of mappings. We indicate these sets by

$$T_d \backslash\backslash A^S,\ C_{2v} \backslash\backslash A^S,\ C_{3v} \backslash\backslash A^S.$$

Pólya's theory of enumeration under group action contains enough results to evaluate first of all the *size* of these sets, i.e. the size of the libraries of molecules arising from the central molecules and the amino acids. We obtain the following formulae in terms of the order $|A|$ of the set of amino acids which we would like to take: In the cubane case we get that

$$|T_d \backslash\backslash A^S| = \frac{1}{24}\left(|A|^4 + 6 \cdot |A|^3 + 11 \cdot |A|^2 + 6 \cdot |A|\right).$$

$|A| = 20$ yields the number 8 855. For the xanthen we obtain

$$|C_{2v} \backslash\backslash A^S| = \frac{1}{2}(|A|^4 + |A|^2).$$

$|Y| = 20$ gieves 80 200, while for the triacid we find

$$|C_{3v} \backslash\backslash A^S| = \frac{1}{3}(|A|^3 + 2 \cdot |A|),$$

which gives 2680 in the case when $|A| = 20$. Here is a table for different sizes of A :

$\|A\|$	cubane	xanthene	triacid
1	1	1	1
2	5	10	4
3	15	45	11
4	35	136	24
5	70	325	45
6	126	666	76
7	210	1225	119
8	330	2080	176
9	495	3321	249
10	715	5050	340
11	1001	7381	451
12	1365	10440	584
13	1820	14365	741
14	2380	19306	924
15	3060	25425	1135
16	3876	32896	1376
17	4845	41905	1649
18	5985	52650	1956
19	7315	65341	2299
20	8855	80200	2680

Pólya's theory allows a refinement, the enumeration of the libraries *by weight.* This means that we can derive a multivariate polynomial such that the coefficient of a monomial summand is the number of orbits the weight of the elements of which is just the sequence of exponents of the monomial. For example, in the case of the cubane, the polynomial

$$\frac{1}{24}((\sum_y y)^4 + 6(\sum_y y^2)(\sum_y y)^2 + 6(\sum_y y^4) + 3(\sum_y y^2)^2 + 8(\sum_y y^3))$$

is the desired generating function. For xanthen it is

$$\frac{1}{2}((\sum_y y)^4 + (\sum_y y^2)^2),$$

while for the triacid we obtain

$$\frac{1}{3}((\sum_y y)^4 + (\sum_y y^2)^2).$$

In order to get that in a more explicit form, we can use the *online calculations* offered by the home page of MOLGEN

```
http://www.mathe2.uni-bayreuth.de/molgen4/
```

It yields, for the xanthen case, say, and for 3 admissible amino acids the expression

```
grf_arb

The input was

[[2,1,4,3]]

and

3

The result of the computation is

1 [0,0,4]  2 [0,1,3]  4 [0,2,2]  2 [0,3,1]

1 [0,4]  2 [1,0,3]  6 [1,1,2]  6 [1,2,1]

2 [1,3]  4 [2,0,2]  6 [2,1,1]  4 [2,2]

2 [3,0,1]  2 [3,1]  1 [4]

the computation was finished after 0.01 seconds

on a pentium 133 MHz
```

It shows, for example, that there are exactly 6 elements of type $[1, 2, 1]$ in the library, i.e. which contain exactly one amino acid of type 1, 2 of type 2 and 1 of type 3.

3 The library

The most interesting problem is, of course, the *generation* of the elements in the library itself. In order to solve this we simply need to apply the connection between isomers and double cosets introduced by Ruch/Hässelbarth/Richter ([4], see also [5]) for the construction of permutational isomers of a given ligand partition. It says, for example, that the elements in the library that contain a_1 amino acids of type 1, a_2 amino acids of type 2 . . . are in one-to-one correspondence to the elements of the set of couble cosets

$$T_d \backslash S_4 / S_{a_1} \oplus S_{a_2} \oplus \ldots$$

in the cubane case, while we have for the xanthen case the corresponding bijection to the set of double cosets

$$C_{2v} \backslash S_4 / S_{a_1} \oplus S_{a_2} \oplus \ldots .$$

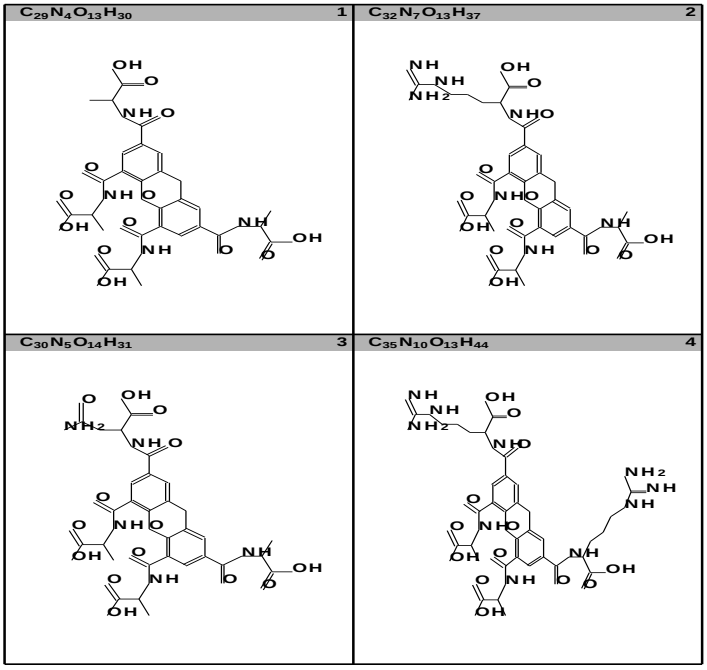
and for the triacid we get

$$C_{3v} \backslash S_3 / S_{a_1} \oplus S_{a_2} \oplus \ldots$$

This fact is used in the software package

$$MOLCOMB$$

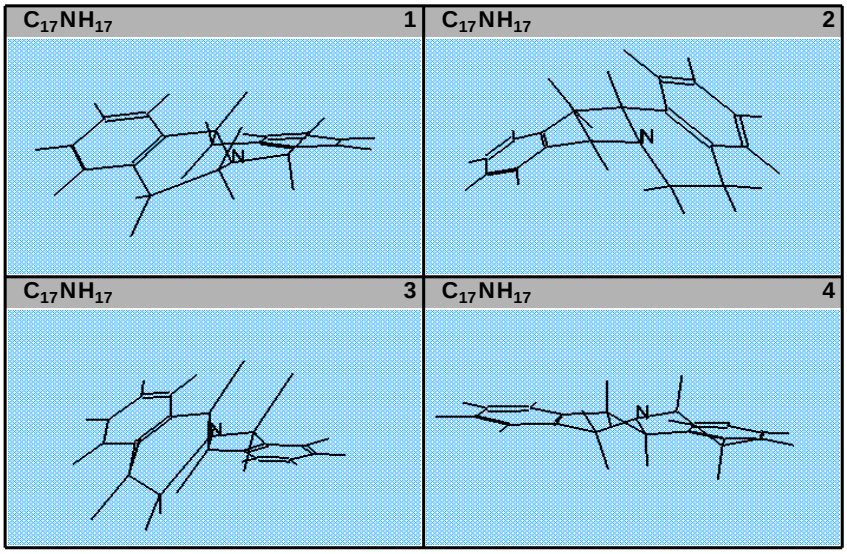
which will be demonstrated. Here are 6 elements of the xanthen library for the case when 3 different amino acids are allowed (see the generating function given above).



More detailed descriptions of the methods used can be found in [6], [7],[8].

4 Conformations

Having the library at hand we may want to do screening for a lead structure. Such a structure is mostly a geometric one, and so we need to evaluate *conformations* for the elements in the library. This is, of course, the most time consuming part of the game. For this purpose we developed an object oriented way of classifying energy minima ([9]) and a mixture of conjugate gradient method of optimization together with an evolutionary approach to the evaluation of energetic minima ([10]). Here is an example of several classes of conformations of berbin which we obtained this way (starting from a random distribution of the atoms in space, so that, at least in principle, we can reach all the minima, and also using atomic tables for the approximate lengths of bonds and sizes of angles).



References

[1] Carell, T., E.A. Wintner, A. Bashir-Hashemi, J. Rebek Jr.. Neuartiges Verfahren zur Herstellung von Bibliotheken kleiner organischer Moleküle. *Angew. Chemie*, **106**, S. 2159-2161, 1994.

[2] Carell, T., 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**, S. 171–183, 1995.

[3] G. Pólya. Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen. *Acta mathematica*, **68**, pp. 145–253, 1937.

[4] E. Ruch, W. Hässelbarth, B. Richter. Doppelnebenklassen als Klassenbegriff und Nomenklaturprinzip für Isomere und ihre Abzählung. *Theoretica Chimica Acta*, **19** (1970), 288-300.

[5] E. Ruch, D. J. Klein. Double cosets in chemistry and physics. *Theoretica Chimica Acta*, **63** (1983), 447-472.

[6] T. WIELAND. Mathematical Simulations in Combinatorial Chemistry. *MATCH*, **34**, pp. 179–206, 1996.

[7] T. WIELAND. Combinatorics of Combinatorial Chemistry. *J. Math. Chem.*, **21**, pp. 141–157, 1997.

[8] T. WIELAND. Konstruktionsalgorithmen bei molekularen Graphen und deren Anwendung. *MATCH*, **36**, pp. 5–155, 1997.

[9] CH. BENECKE. Objektorientierte Darstellung und Algorithmen zur Klassifizierung endlicher bewerteter Strukturen. *MATCH*, **37**, pp. 7–156, 1997.

[10] C. FREY. An Evolutionary Algorithm with Local Search and Classification for Conformational Searching. *MATCH*, **38**, in print, 1998.