communications in mathematical
match
and in computer chemistry

no. 41, March 2000

ISSN 0340-6253

MATCDY (41) 189-203 (2000)

MOLGEN-COMB, a Software Package for Combinatorial Chemistry

R. Gugisch, A. Kerber, R. Laue, M. Meringer[1], J. Weidinger

University of Bayreuth, Department of Mathematics,

D-95440 Bayreuth, Germany

Abstract

We briefly describe a project devoted to the implementation of the software package MOLGEN-COMB for the *simulation* of combinatorial chemistry and the *optimization* of its experiments.

1 Combinatorial chemistry with central molecule and building blocks

To begin with we recall from [3] our mathematical model of a combinatorial library obtained from a central molecule (for example, the cubane) and building blocks (for example, amino acids): *Attachments* of building blocks to the active sites of the central molecule are considered as elements of the set of mappings

$$Y^X := \{f: X \to Y\},$$

where X is the set of active sites, while Y denotes the set of building blocks. Attachments which describe essentially the same molecules form *orbits* under the action of G, the *symmetry group* of the central molecule, on Y^X :

$$G \times Y^X \to Y^X: (g, f) \mapsto f \circ g^{-1}.$$

[1]financial support by the BMBF under contract 03-KE7BA 1-4 is gratefully acknowledged

The *orbit* of the attachment f was indicated by

$$G(f) := \{f \circ g^{-1} \mid g \in G\},$$

and we denoted the set of all the orbits of G on Y^X as follows:

$$G\backslash\backslash Y^X.$$

The Lemma of Cauchy-Frobenius gives the number of orbits in terms of numbers of fixed points. If we apply it to the action of G on Y^X we obtain that

$$|G\backslash\backslash Y^X| = \frac{1}{|G|} \sum_{g \in G} |Y|^{|\langle g \rangle \backslash\backslash X|},$$

where $|\langle g \rangle \backslash\backslash X|$ denotes the number of orbits of the group $\langle g \rangle$, generated by the element $g \in G$.

This formula allows to evaluate the size of a combinatorial library obtained from a central molecule by attaching building blocks.

In the cubane case the symmetry group is — in mathematical terms — the alternating group A_4, consisting of 12 proper rotations. Hence, the Lemma of Cauchy-Frobenius gives the number of essentially different attachments in terms of the number $|Y|$ of admissible building blocks (=amino acids):

$$|A_4\backslash\backslash Y^X| = \frac{1}{12}\left(|Y|^4 + 11 \cdot |Y|^2\right).$$

In the case of $|Y| = 19$ we obtain, for example, that there exist exactly 11 191 essentially different attachments.

The next step is a refined evaluation of the library using *weighted enumeration,* which means, for example, that we can evaluate the number of elements in the library which contain given building blocks with prescribed multiplicities.

Pólya's Theorem gives the numbers of orbits of G on Y^X by weight (=multiplicities of prescribed building blocks) in terms of the *cycle index polynomial:*

$$Cyc(G, X) := \frac{1}{|G|} \sum_{g \in G} \prod_{i=1}^{|X|} x_i^{|\langle g \rangle \backslash\backslash_i X|},$$

where $|\langle g \rangle \backslash\backslash_i X|$ denotes the number of orbits of length i of $\langle g \rangle$, or, in other words, the number of i-cycles of g on X.

The number of orbits of G on Y^X which consist of mappings f that take the value $y \in Y$ with multiplicity b_y is the coefficient of the monomial $\prod_{y \in Y} y^{b_y}$ in the polynomial

$$\frac{1}{|G|} \sum_{g \in G} \prod_{i=1}^{|X|} \left(\sum_{y \in Y} y^i\right)^{|(g)\%_i X|},$$

which arises from the cycle index by replacing the indeterminate x_i by the i-the *power sum symmetric function* $\sum_y y^i$.

There exists an interesting description (due to H. O. Foulkes) of the generating functions which Pólya obtained, in terms of Schur functions $\{\alpha\}$, which are generating functions for the irreducible characters ζ^α of the irreducible representations $[\alpha]$ of the symmetric group,

$$\{\alpha\} := \frac{f^\alpha}{|X|!} \sum_{\pi \in S_X} \zeta^\alpha(\pi) \prod_{i=1}^{|X|} \left(\sum_y y^i\right)^{a_i(\pi)}.$$

In terms of the multiplicities $(IG \uparrow S_X, [\alpha])$ of these $[\alpha]$ in the representation $IG \uparrow S_X$ of S_X, which is induced from the identity representation of the symmetry group G, we have

$$\frac{1}{|G|} \sum_{g \in G} \prod_{i=1}^{|X|} \left(\sum_{y \in Y} y_i\right)^{a_i(g)} = \sum_{\alpha \vdash |X|} (IG \uparrow S_X, [\alpha]) \cdot \{\alpha\}.$$

($\alpha = (\alpha_1, \alpha_2, \ldots)$ denotes the proper partition that characterizes the corresponding irreducible representation, i.e. $\alpha_i \geq \alpha_{i+1}$, $\alpha_i \in \mathbf{N}$ and $\sum_i \alpha_i = |X|$, the degree of the symmetric group in question.) The cubane case is trivial, we obtain as generating function the sum of Schur functions

$$\{|X|\} + \{1^{|X|}\},$$

as always when the symmetry group of X is the alternating group A_X. The summands correspond to the one-rowed and to the one-column Young diagram, as it is well known from the representation theory of the symmetric group.

Enumeration gives a numeric estimate of what we can expect in the combinatorial library. What we really want is, of course, to get the elements of the library itself, i.e. the structural formulae of these molecules. In order to attack this problem we note that **the desired combinatorial library of molecules is obtained as a transversal, i.e. a complete set of representatives of this set of orbits.**

It was mentioned in [3] that Ruch, Hässelbarth and Richter have shown (see [5],[6]), how these orbits are bijectively related to double cosets: The set of orbits

$$G\backslash\backslash_f Y^X$$

consisting of mappings with the same *weight* (or *content*, as Pólya called it) as f, i.e. which take the values $y \in Y$ with the multiplicities $|f^{-1}(y)|$ (for example in the case of the cubane and the amino acids, $|f^{-1}(y)|$ means the number of active sites, where the amino acid y is attached), can be mapped bijectively onto a set of double cosets as follows:

$$G\backslash\backslash_f Y^X \longrightarrow G\backslash S_X/(S_X)_f\colon\ G(f\circ\pi) \longmapsto G\pi(S_X)_f,$$

where $(S_X)_f$ denotes the stabilizer of f in the symmetric group S_X. In fact, the stabilizer of f in the symmetric group S_X is the direct sum (or direct product, if you prefer) of the symmetric groups $S_{f^{-1}(y)}$ on the inverse images of the values which f takes:

$$(S_X)_f = \oplus_y S_{f^{-1}(y)}.$$

for each element π of the symmetrix group S_X on X.

The crucial point is now that we can obtain a transversal of this set of orbits $G\backslash\backslash_f Y^X$ by constructing a transversal of this set of double cosets and then retranslating its elements into attachments.

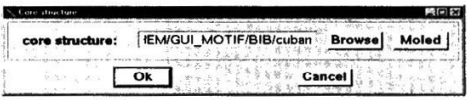
Methods for obtaining transversals of sets of double cosets were described, and we should like to mention that using MOLGEN-COMB, a generator software for this purpose which was developed recently in Bayreuth, can generate the full library of molecules with central cubane and building blocks taken from the set of 19 amino acids, within ten minutes on a work station.

Hence, in order to summarize part I of this paper, we have shown that enumeration under finite group action can cover famous examples of combinatorial chemistry. The method used is Pólya's Ansatz to choose suitable sets X, Y and a finite group G acting on X. The examination of the induced action of G on the set of mappings Y^X allows to evaluate the total number of elements in a combinatorial library arising from a core with the set X as set of active sites by attaching elements of a set Y of building blocks, and with respect to the symmetry group G of X. This number is obtained

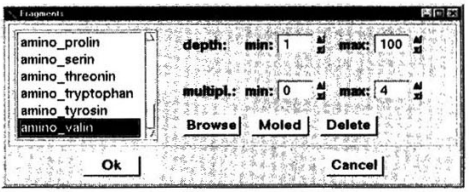
by an easy application of the Cauchy-Frobenius Lemma. Moreover we can obtain (by an application of Pólya's Theorem) a generating function for the elements of the library, which enumerates these elements by weight. Finally we can even construct the library using the double coset reformulation of the orbits.

2 The virtual library

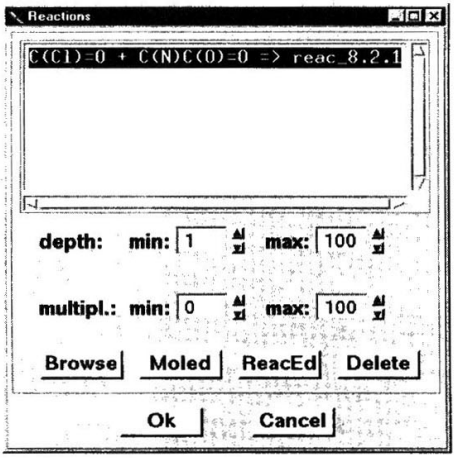
Here is an optical description how the combinatorial library arising from cubane and the attachment of amino acids is obtained by an application of MOLGEN-COMB. We should like to describe it in detail since this shows where problems are. To begin with, the user enters the central molecule:



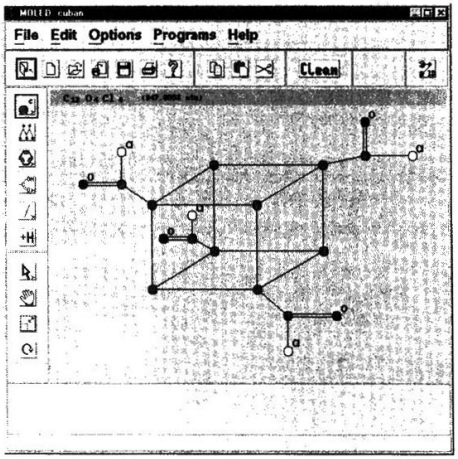
Then the building blocks are put in,



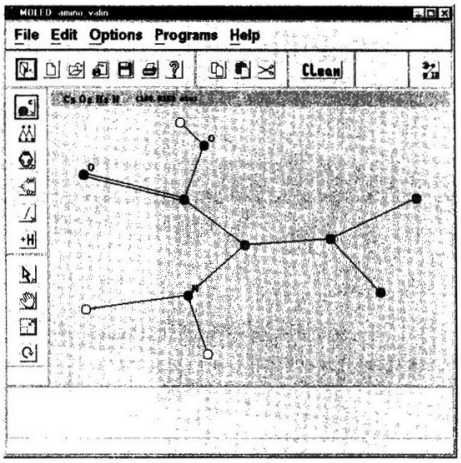
and the admissible reaction(s):



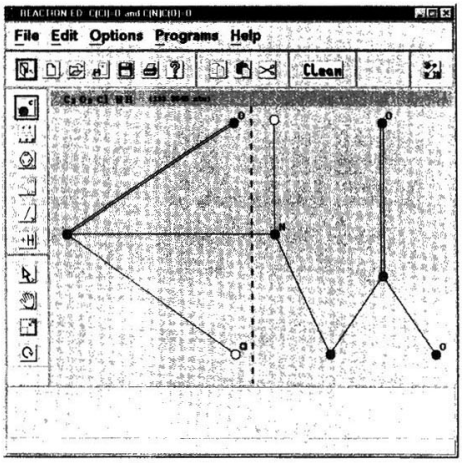
Of course, the central molecule will be shown:



as well as the building blocks:

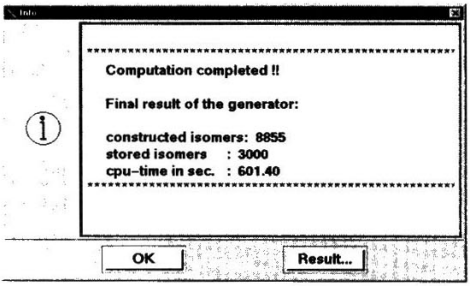


and the reaction, in graphical terms,

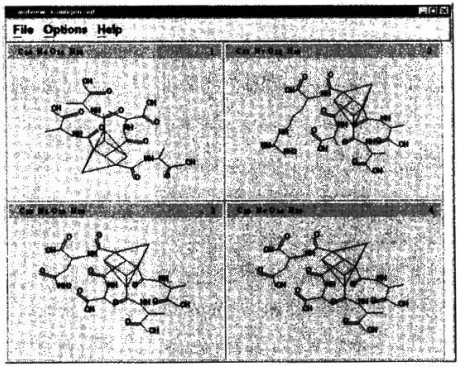


Afterwards we select the admissible building blocks

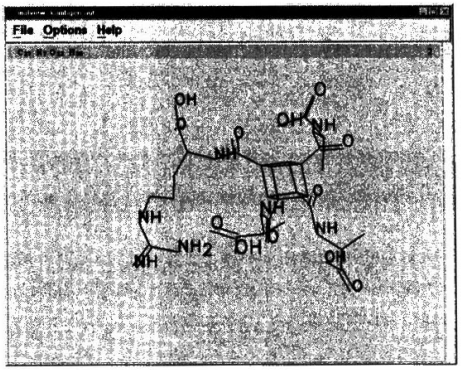
and start the generator which evaluates the combinatorial library in a few minutes



Sequences of elements can be shown:



as well as single elements in a 3D-placement (using the MM2 energy model):



Although we nicely generated the library and stored its elements, **there is a serious problem, since the generator gives 8 855 structural formula, while we obtained 11 191 elements by the Lemma of Cauchy-Frobenius! What has happened and what can be done?** To begin with, let us list the basic facts which lead to this smaller number of elements in the desired library:

- The generator — in its present form, an improved version is under construction has as input *the structural formula of the cubane, i.e. a connected multigraph.*

Hence, all what it can do (at present) is to evaluate *the automorphism group of that graph*, which is, in mathematical terms, the hyperoctahedral group, a wreath product $S_2 \wr S_3$, a group of order 48. The generator restricts the action of this group to the set of active sites, and it is clear that this amounts to an action of the symmetric group S_4.

- But, as it was already mentioned, the symmetry group which we have to take into account is the *alternating group* A_4, since we want to consider proper rotations only. And we should carefully note that this is a *decision*, the reason of which is a chemical argument. So what we usually need and, as we said, a corresponding generator is under construction, is a chance to recognize the symmetry group of the conformation which is a (possibly proper) subgroup of the automorphism group of the underlying graph.

- This group which may be a proper subgroup of the permutation group induced on the active sites by the automorphism group of the underlying graph, must be *the symmetry group of the conformation of the central molecule*. And this is a further, and very serious problem, since there are usually no conformations at hand, nor are there classes of conformations available (like in the cyclohexane case, where we have the chair, the bath tub and the twistor).

- What we can do, of course, is to evaluate a conformation using, say, an energy model like the MM2-model, an optimization and atomic tables for the bond lengths and angles. But this does not solve the problem either, at least it does not solve it directly. The reason is that whenever we do such a calculation, starting from a random distribution of the atoms in space (which makes sense for small molecules in order to at least have a chance to find the global energy minimum), we obtain a result that is, first of all, different from the foregoing, and secondly, is an inexact cube, for example, in the cubane case. And so we are in trouble when we want to evaluate its symmetry group, since we now enter, in a certain sense, *fuzzy mathematics*.

- Finally, even if we have solved the problems just mentioned, we get into prob-

lems since there is no standard data structure that allows routinely to store conformations and to distinguish, say, the two elements of an enantiomeric pair of molecules. So we finally run into a very serious problem of *chemical nomenclature.*

All in all, we have met a bunch of serious problems that have to be solved in the near future. All what we can say is, that we are working on it. For oriented planar graphs see [4], further ideas are presented in [1]. Of course, we can use several tricks, for example to add information about chirality, and to find out this way that there are in fact 11 191 different molecules in the library, but such *tricks should be replaced by finding a normal form for the 3D placements.* **Before this is solved we have a generator of molecular libraries "only" on the topological level.**

3 The virtual library and a real sublibrary

Nevertheless we are now going to continue by describing the use of such *virtual* (i.e. mathematically generated) combinatorial libraries *in order to design optimal combinatorial experiments.*

Suppose we are given a combinatorial library in the sense that we have generated each of its elements and tabulated them in a separate file. We call this library the *virtual library,* in order to distinguish it from its counterpart, the *real library.* This real library is supposed to be a set molecules for which we know the following:

- Each element of the real library is contained in the virtual library,
- and for each element of the real library we have certain parameter values which allow us to distinguish the *good* from the *bad* elements of the real library.

Usually the real library contains a few, maybe a few hundred molecules, while the virtual library quite often contains *several thousand* molecules. **Therefore it is an important and interesting question to obtain a decent prediction which of the further elements in the virtual library might be good as well.** The purpose is to *design an experiment of combinatorial chemistry that synthesizes the possibly good elements of the virtual library and not too many others,* for economic

reasons, of course. The crucial point is that this procedure confronts us with the serious problems that have been listed above, and the solution of which is a major task for the future!

> The first question that is natural to be asked and which is important to get an answer to, is "which is which". We ask which of the elements in the virtual library correspond to the elements of the real library. In mathematical terms, *we want to embed the real library into the virtual one.*

Of course, there are cases in which we can easily get an answer, for example, if both the elements in the virtual and the real library are given by a uniquely defined sequence of reactions, by a *synthesis path*, as we may call it. But it is easy to see that this will already fail in the above cubane case.

We might also try to get the CAS numbers or IUPAC names for the elements in the real and in the virtual library, but this is, first of all, cumbersome and very costly, and, secondly, it will mostly fail, since only part of the virtual library will be found in CAS.

This idea of getting CAS numbers or IUPAC names shows that we are suddenly and deeply involved with problems of *chemical registration*. So let us talk about what can be done and how it should be done.

To begin with, we should note that it clearly should be done automatically. This means that we need a data structure which is uniquely defined and for which there exists a canonical form. Hence neither CAS numbers (which we mostly cannot re-translate into structural formulae) nor IUPAC names (which are not unique) are allowed.

Moreover, the standard way of describing a molecule by its connection table (for which a canonic form can be evaluated, see the Computer Corner of the present issue of MATCH) does not suffice as we can see in our cubane example: There are chiral situations, enantiomeric pairs of molecules that have the same connection table! *It should be emphasized that this is not a mathematical problem, since the mathematical generator does generate both elements of the enantiomeric pair, but when it comes to storing the elements of a pair by a connection table we loose information, the connection tables (in canonical form) are both equal.*

4 The optimization of an experiment

Assume that we are given a virtual library and a real sublibrary. Recall that we assume to know which of the elements in the real library are "good" molecules for a particular purpose. The aim is to find out which of the elements in the virtual library might be good molecules, too, in order to design an experiment of combinatorial chemistry which generates a set of molecules containing good ones with a higher percentage than the total virtual library.

A standard method for doing this is, of course, data mining via regression analysis, using *molecular descriptors*.

The good molecules of the real library are usually classified by the values of certain parameters which distinguish them from the "bad" ones.

Moreover, being aware of their structural formula, we can calculate molecular descriptors for all molecules both in the real and in the virtual library. For example, we can evaluate *topological indices*. Other descriptors, let us call them *geometrical indices*, are based on a theoretical conformation. Theoretical means, the conformation is obtained from the structure formula by some algorithm, for example an energy optimization using the MM2 model.

There exists a huge amount of different descriptors already used in several cases with some kind of success. We provide about 100 topological indices and about 25 geometrical descriptors using a conformation and results on atom charges (provided, for a particular case, by the Gasteiger group in Erlangen). An individual selection of indices is possible. Furthermore we want to provide an interface to include additional descriptors computed by external programs. For example, we can include indices produced by the Cluj-program ([2]), which is able to calculate about 16000 different indices.

Now, in order to do data mining, we look for relations between the (measured) parameter values of the elements in the real library and the (calculated) descriptors of these molecules. The difference to standard methods of statistical data analysis is, that we don't have any preassumptions on the kind of relations. We don't even know, if there are any relations at all. As there is nearly no expert knowledge about

relationships between any kind of structure based descriptors and physical, chemical or biological properties or activities, we want to implement a software package which does an automatic data mining process in order to get as much statistical information out of the real library, as possible.

There are several statistical methods available for this purpose. Well known examples are linear regression analysis, regression trees, neuronal networks.

As a first step, we use regression trees, which are quite fast, can handle a huge amount of variables and are especially qualified for nonlinear relations. Certainly other methods should be used, too, say in parallel, and the best results should be picked.

After the step of data mining, we will get a set of selected virtual descriptors and some kind of prediction of the parameter values based on these descriptors. In addition we should get an idea of the quality of this prediction (in order to be able to compare it with other methods of data mining).

Once we have a set of virtual descriptors together with a prediction, we can compute the descriptors on the virtual library, and then predict the parameter values, i.e. we get *virtual parameter values on the virtual library.* Using these virtual parameters, we can predict, which molecules might be good, and so we can rank the virtual library according to the *predicted quality* of the molecules. Thus the experimental chemist has the chance to synthesize additional molecules, with a higher probability of being good substances. Moreover, at least in principle, this should give us a chance to understand **why certain molecules are good in a certain sense or for a certain purpose.**

References

[1] C. BENECKE, A. KERBER, R. LAUE. *Canonical numbering of 3D-Molecules.* Presented at the Second International Electronic Conference on Computational Chemistry, November 1995. Download from http://www.mathe2.uni-bayreuth.de/molgen4/

[2] L. JÄNTSCHI, G. KATONA, M. V. DIUDEA. *Modeling Molecular Properties by Cluj Indices.* MATCH **41** (2000), 151-188.

[3] A. KERBER, R. LAUE, T. WIELAND. *Discrete Mathematics for Combinatorial Chemistry.* Proceedings DIMACS Workshop on Discrete Mathematical Chemistry, March 23-24, 1998, Rutgers University.

[4] R. LAUE. Construction of combinatorial objects – a tutorial. *Bayreuther Mathematische Schriften*, **43**, pp. 53–96, 1993.

[5] E. RUCH, W. HÄSSELBARTH, B. RICHTER. Doppelnebenklassen als Klassenbegriff und Nomenklaturprinzip für Isomere und ihre Abzählung. *Theoretica Chimica Acta*, **19** (1970), 288-300.

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