communications in mathematical
match
and in computer chemistry

no. 34, October 1996

ISSN 0340-6253

MATCDY (34) 179-206 (1996)

Mathematical Simulations in Combinatorial Chemistry*

Thomas Wieland†
Lehrstuhl II für Mathematik, University of Bayreuth,
95440 Bayreuth, Germany

Abstract

A novel technique for chemical synthesis in drug research is *combinatorial chemistry*, where usually a set of building-block molecules is attached to a core structure in all the combinatorially possible ways. The resulting set of compounds (called a *library*) can then be systematically screened for a desired biological activity. In this paper we discuss ways and limits of a mathematical simulation of this procedure. At first, two methods for selecting the building-blocks from a given structure pool are presented with the objective to obtain only dissilimar library entries. Next an algorithm is described for the exhaustive and redundancy-free generation of a combinatorial library, illustrated by a single-step and a multi-component reaction. Finally equations for the enumeration of the library sizes are derived and the limits of the *virtual combinatorial chemistry*, i.e. purely in computer and without experiment, are discussed.

1 Introduction

The common ways to develop a new pharmaceutical drug are to extract a natural drug from bakteria, plants or animals, to use the available potential in laboratory and in-house databases or to employ methods of rational drug-design based on mechanism or structure.
With the upcoming of new analysis automata, it became possible to examine several thousands of compounds a day for their biological activity. (This process is usually refered as *screening*.) Together with the necessity of cost reduction in industrial research, this fact has raised the desire for making very large numbers of novel molecules available. So here industrial revolution has its impact on synthetical chemistry, replacing "hand-work" by fast and efficient machines.
In the recent years a novel technique is used for this purpose, which does not aim at the classical objective of synthesizing *one* substance as pure as possible, but which deliberately utilizes the variety to produce a large number of compounds simultaneously. This technique is called *combinatorial chemistry* [1]- [6].
Typically a set of *building-blocks* is taken that is systematically combined with a *core* structure in all the combinatorially possible ways where the actual reactions make use of chemical,

*Supported by the BMBF under 03-KE7BAY-9
†E-mail: thomas.wieland@uni-bayreuth.de

biological or biosynthetical procedures. The set of resulting molecules is called a *combinatorial library*.
Next to the high efficiency this technology derives its elegance also from another aspect: This approach clearly represents the processes in nature when millions of years ago, at the origins of life, the first proteins and carbon hydrates were formed by the combination of simple carbon hydrogens and amino acids.

A crucial issue in combinatorial chemistry is *diversity* [7]-[9]. A large combinatorial library may fulfill the demand for making many compounds available; for an efficient analysis, however, it should be certified that the elements of a library are not too similar in order to avoid one pharmacological class being tested over and over again. Thus the elements of a library should be as diverse as possible to cover a broad variety with the screening – without requiring too many single substances.
So in connection with combinatorial chemistry the notion of *similarity* has come into the focus [7], [8], [10]- [12]. There is a vast number of ways to define and determine similarity of chemical entities. It turned out that similarity – unlike isomorphy, e.g. – cannot be defined generally. It depends, in fact, on the structure as well as on the studied activity what must be considered similar and what must not. Quite often "similarity" is, however, used in the sense of "structural similarity"; we will also make use of this view in the following.

So the main procedure in combinatorial chemistry can be summarized in the following steps:

i) Selection of building-blocks

ii) Generation of the library

iii) Screening of the library for the required activity

Of course the aim of a computer simulation is to perform the steps as complete as possible by the computer. In this paper we will give an overview over some methods which can be used in each of these steps, provide some mathematical generalizations and take a look at the limits of the simulation.

2 Selection of building-blocks

There are several possibilies for selecting the building-blocks [8, 9, 13]. Their application mainly depends on the objective that is sought by the combinatorial library. In this study we will present two methods based on graph theory in conjunction with statistical analysis.
As a simple example set we will consider the 20 natural amino acids (see Fig. 1).

2.1 Basic definitions

In this paper we consider graphs as mappings

$$\gamma : \underline{\mathbf{p}}^{[2]} \to \{0, \ldots, m-1\}, \quad \text{in short } \gamma \in m^{\underline{\mathbf{p}}^{[2]}}$$

where $\underline{\mathbf{p}}^{[2]}$ is the set of pairs of points of the graph, i.e. the set of all 2-subsets of the set $\underline{\mathbf{p}} := \{1, \ldots, p\}$ of points (or, to be exact, the set of the numbers of the p points), $\gamma(\{i,j\}) = k$ means that there is an edge of degree k – a k-fold bond – between the vertices i and j, and $\gamma(\{i,j\}) = 0$ if the two vertices are not connected.

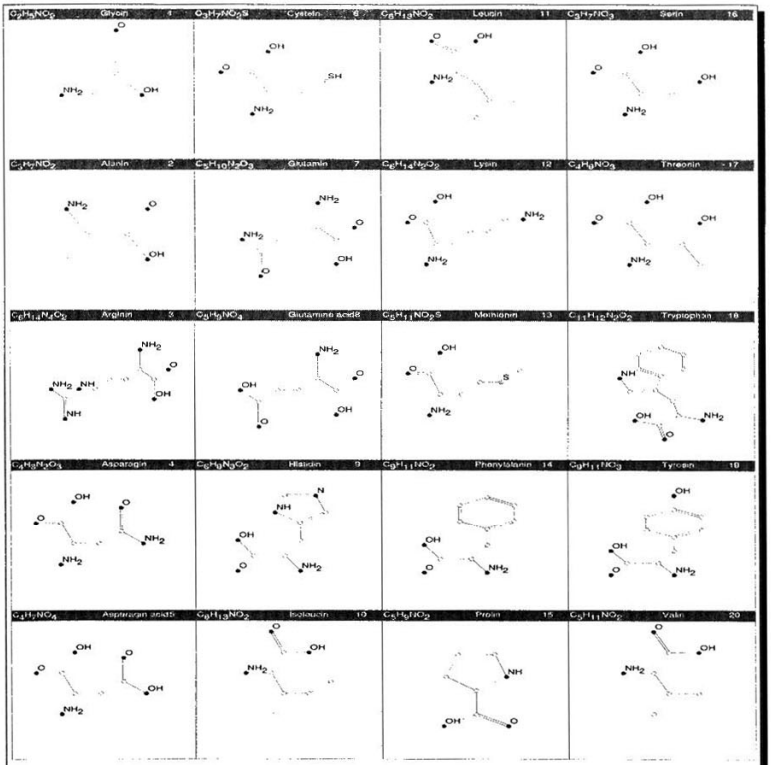


Figure 1: The 20 natural amino acids

For molecules we take the usual model, identifying atoms with vertices and bonds with edges. The atomic types are defined by an additional mapping $\beta : \underline{\mathbf{p}} \to \{E_1, E_2, \ldots\}$ with the E_i representing chemical elements such that a molecular graph is a pair (γ, β) of a graph and a coloring of the vertices with atomic types.

Furthermore we call

$$\eta : T^{[2]} \to \{0, \ldots, m-1\} \text{ with } T \subseteq \underline{\mathbf{p}},\ \forall i, j \in T :\ \eta(\{i,j\}) = \gamma(\{i,j\})$$

a *subgraph* of γ, which we indicate by $\eta \subseteq \gamma$.

2.2 Topological indices

A large number of studies have been carried out on the search for quantitative structure-activity relationships (QSAR), i.e. the search for empirical or theoretical parameters that are directly correlated to some biological response [10], [14]- [19]. Since empirical data are not always available [14, 18] and experiments or quantum chemical calculations are expensive for larger sets of compounds, a lot of interest lies currently in the use of *topological indices* (or graph invariants) [10], [15], [20]-[23] as discrimination criteria and prediction tools.

A topological index is a numerical value computed only from the (hydrogen-suppressed) molecular graph. They are suitable, within a limited range, for modeling molecular properties like flexibility, surface, branching etc. We will use them to investigate structural similarities in a set of potential building-blocks.

The rather simple ones are the numbers of atoms, bonds and rings as well as the molecular weight. In addition, we use the numbers of stereocenters and stereoisomers computed with methods from [24, 25].

A series of important values is based on the *adjacency matrix* A (with $a_{i,j} = 1$ if $\gamma(\{i,j\}) \neq 0$ and $a_{i,j} = 0$ otherwise) and the *connectivity matrix* A' with $a'_{i,j} = \gamma(\{i,j\})$. The *degree* of the i-th vertex is defined as $\delta_i := \sum_{j=1}^{p} a_{i,j}$, whereas the *bond degree* is derived from the connectivity matrix as $\deg_i := \sum_{j=1}^{p} a'_{i,j}$. The *connectivity indices* ${}^k\chi$ and ${}^k\chi^b$ for $k = 0, 1, 2$ after [10] are sums over all paths of length k in the graphs, varying by the use of the adjacency or the connectivity matrix.

$$ {}^0\chi = \sum_{i=1}^{p} (\delta_i)^{-\frac{1}{2}} \tag{2.1} $$

$$ {}^0\chi^b = \sum_{i=1}^{p} (\deg_i)^{-\frac{1}{2}} \tag{2.2} $$

$$ {}^1\chi = \sum_{\text{edges } i,j} (\delta_i \delta_j)^{-\frac{1}{2}} \tag{2.3} $$

$$ {}^1\chi^b = \sum_{\text{edges } i,j} (\deg_i \deg_j)^{-\frac{1}{2}} \tag{2.4} $$

$$ {}^2\chi = \sum_{\text{paths } v_i,v_j,v_k} (\delta_{v_i} \delta_{v_j} \delta_{v_k})^{-\frac{1}{2}} \tag{2.5} $$

$$ {}^2\chi^b = \sum_{\text{paths } v_i,v_j,v_k} (\deg_{v_i} \deg_{v_j} \deg_{v_k})^{-\frac{1}{2}} \tag{2.6} $$

These indices together give a good characterization of the structure, especially with respect to shape, volume, and surface, and have thus been used for a large number of correlations [8, 10, 19].

Another important class of indices is based on the distance matrix D, where each entry $d_{i,j}$ denotes the length of the shortest path from vertex i to vertex j. The distance matrix can be calculated by the Floyd algorithm [26].

The very first application of topological indices was developed by H. Wiener [27]. His index is

$$ W = \tfrac{1}{2} \sum_{i,j}^{p} d_{i,j} = \sum_{i>j}^{p} d_{i,j}. \tag{2.7} $$

Furthermore we use the *Balaban-Index* [28]

$$J = \frac{p}{q+1} \sum_{\text{edges } i,j} (d_i d_j)^{-\frac{1}{2}} \tag{2.8}$$

where p is the number of atoms, q the number of rings and $d_i := \sum_j d_{ij}$.
The *mean square distance index* [29] is due to Balaban and Motoc

$$MSD = \left(\frac{\sum_{i=1}^{d_{\max}} i^2 g_i}{\sum_{i=1}^{d_{\max}} g_i} \right)^{\frac{1}{2}} \tag{2.9}$$

where g_i is number of vertex pairs in γ having distance i, and $d_{\max}$ is the largest entry in D.

Finally we make also use of *information-theoretic* indices [12, 20]. There a set of n elements is partitioned into h classes, denoting the relative frequency of the i-th class by p_i. According to Shannon's relation [30, 31] the mean information content is calculated as $-\sum_{i=1}^{h} p_i \log_2 p_i$.
The mean information content of distances was used by Bonchev and Trinajstić [32] in the form (with the notation from eqns. 2.7 and 2.9)

$$I_D^W = W \log_2 W - \sum_{i=1}^{d_{\max}} g_i \cdot i \log_2 i \tag{2.10}$$

A natural partition of the vertices is given by the orbits of the automorphism group (as already studied by Rashevsky [33]). Let p_i the length of the i-th orbit (with a total of h orbits, say) divided by the number p of vertices. Then the following indices can be defined [12, 19]:

$$IC = -\sum_{i=1}^{h} p_i \log_2 p_i \tag{2.11}$$

$$SIC = IC / \log_2 p \tag{2.12}$$

$$CIC = \log_2 p - IC \tag{2.13}$$

Table 1 shows the calculated index values for the amino acids. Such an amount of numbers is of course rather impracticable for further treatment. A commonly used method for data reduction is *principal component analysis* (PCA). It uses implicit correlations among the values and calculates the eigenvectors in order to obtain a few, new variables, called *factors*, which explain most of the variance in the original data.[1]
In our case, PCA yields three factors explaining 93.2 % of the original variance. The regression values are shown in the following table:

[1] For the calculations in this paper we used the software *SPSS*, version 6.0.1 for Windows.

	Atoms	Bonds	Rings	St-cent.	St-isom.	W	Weight	${}^0\chi$	${}^0\chi^b$	${}^1\chi$	${}^1\chi^b$	${}^2\chi$	${}^2\chi^b$	I_D^W	MSD	Balab.	IC	SIC	CIC
gly	5	4	0	0	1	18	80.010	4.284	3.914	2.270	1.914	1.802	1.311	57.550	1.732	1.917	2.922	0.880	0.400
ala	6	5	0	1	2	29	96.010	5.155	4.784	2.643	2.297	2.488	2.001	109.900	2.000	2.405	2.931	0.792	0.769
arg	12	11	0	4	8	247	188.000	9.560	8.820	5.537	4.835	4.900	3.950	1433.000	4.243	3.032	3.194	0.680	1.506
asn	9	8	0	1	2	96	140.000	7.439	6.699	4.037	3.335	3.851	2.909	479.700	2.828	3.047	3.102	0.759	0.986
asp	9	8	0	1	2	96	140.000	7.439	6.699	4.037	3.335	3.851	2.909	479.700	2.828	3.047	3.078	0.770	0.922
cys	7	6	0	1	2	46	128.100	5.862	5.492	3.181	2.835	2.630	2.156	195.600	2.236	2.666	3.325	0.873	0.483
gln	10	9	0	1	2	136	156.000	8.146	7.406	4.537	3.835	4.192	3.243	719.400	3.317	3.047	3.041	0.704	1.280
glu	10	9	0	1	2	136	156.000	8.146	7.406	4.537	3.835	4.192	3.243	719.400	3.317	3.047	3.005	0.708	1.243
his	11	11	1	6	16	165	164.000	8.268	7.431	5.198	4.199	4.607	3.409	921.000	3.317	2.029	3.504	0.811	0.818
ile	9	8	0	2	4	92	144.000	7.439	7.069	4.091	3.746	3.489	3.021	461.400	2.646	3.207	2.768	0.621	1.692
leu	9	8	0	1	2	96	144.000	7.439	7.069	4.037	3.691	3.851	3.377	479.700	2.828	3.047	2.768	0.621	1.692
lys	10	9	0	1	2	143	160.000	7.983	7.613	4.681	4.335	3.717	3.243	753.700	3.606	2.899	2.781	0.607	1.804
met	9	8	0	1	2	102	160.100	7.276	6.906	4.181	3.835	3.364	2.890	507.100	3.162	2.865	3.022	0.699	1.300
phe	12	12	1	1	2	212	176.000	8.975	7.878	5.698	4.392	4.961	3.437	1238.000	3.606	2.066	2.888	0.639	1.635
pro	8	8	1	2	4	62	124.000	5.983	5.613	3.805	3.459	3.289	2.828	288.600	2.236	1.961	2.814	0.688	1.274
ser	7	6	0	1	2	46	112.000	5.862	5.492	3.181	2.835	2.630	2.156	195.600	2.236	2.666	3.039	0.798	0.768
thr	8	7	0	2	4	65	128.000	6.732	6.362	3.553	3.208	3.347	2.879	303.400	2.449	3.026	3.052	0.747	1.036
trp	15	16	2	4	8	369	216.000	10.840	9.585	7.182	5.609	6.503	4.542	2400.000	3.873	1.794	3.320	0.698	1.435
tyr	13	13	1	1	2	268	192.000	9.845	8.801	6.092	4.803	5.583	3.937	1625.000	3.873	2.106	3.183	0.694	1.402
val	8	7	0	1	2	65	128.000	6.732	6.362	3.553	3.208	3.347	2.879	303.400	2.449	3.026	2.774	0.653	1.474
Mean	9.350	8.650	0.300	1.650	3.550	124.450	146.611	7.470	6.870	4.302	3.677	3.830	3.016	683.558	2.939	2.645	3.026	0.722	1.196
Standard dev.	2.412	2.834	0.571	1.424	3.502	89.753	32.552	1.614	1.386	1.193	0.882	1.110	0.732	589.951	0.697	0.485	0.206	0.080	0.407

Table 1: Topological indices of the 20 natural amino acids from Fig. 1

	f_1	f_2	f_3
gly	-1.83387	1.13563	-1.41418
ala	-1.33258	0.49452	-0.50175
arg	1.25580	-0.08585	1.85784
asn	-0.27493	-0.05936	0.56538
asp	-0.29056	-0.01905	0.54880
cys	-1.00464	1.17446	0.44508
gln	0.17849	-0.57497	0.49897
glu	0.16735	-0.58538	0.46081
his	0.84679	2.60881	1.48386
ile	-0.10004	-1.16748	0.60622
leu	-0.07249	-1.36088	0.09030
lys	0.32192	-1.47271	-0.02419
met	-0.10939	-0.50501	0.30335
phe	0.94399	-0.54371	-1.75631
pro	-0.42065	0.24750	-1.54562
ser	-1.01410	0.44544	-0.00660
thr	-0.51357	0.09556	0.86045
trp	2.38110	1.17295	-1.23013
tyr	1.38372	0.00789	-1.27212
val	-0.51236	-1.00835	0.02984

This shows that structures which differ only slightly also have similar factors, e.g. asparagin (asn) and asparagin acid (asp) – a confirmation of our initial assumption that topological indices are suitable for similarity analysis.
After determining the Euclidian distances, i.e.

$$\sqrt{(f_1^{(i)} - f_1^{(j)})^2 + (f_2^{(i)} - f_2^{(j)})^2 + (f_3^{(i)} - f_3^{(j)})^2},$$

we obtain the distance matrix in Tab. 2. This table reveals the quantitative differences between the single molecules more obviously.
For building-block selection, we can group the molecules together according to these results by putting all structures with an euclidian distance below a certain threshold ε in one group. The value of this threshold and thus the coarseness of the decomposition must be determined in agreement with the requirements of the experiment to be simulated. For $\varepsilon = 1.0$, say, we get the following groups:

{gly}, {ala, ser}, {arg}, {asn, asp, gln, glu, met, thr},
{cys}, {his}, {ile, leu, lys, val}, {phe, tyr}, {pro}, {trp}

A first attempt in experiment may consist of taking one representative from each group; afterwards only those groups need to be examined more precisely the representative of which has shown the desired activity.

2.3 Binary property vectors

For the characterization of diversity lists with binary properties can be considered, too. We took a subset of 120 descriptors from the structure codes of the mass spectra information system *MassLib* [34], the same as used in K. Varmuza's program *ToSIM* [35, 36, 37]. The following classes of properties are taken into account:

- Aromatic compounds (e.g. substructure phenyl)

	gly	ala	arg	asn	asp	cys	gln	glu	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
gly	0	1.22	4.66	2.79	2.75	2.04	3.26	3.24	4.21	3.52	3.41	3.66	2.94	3.26	1.67	1.77	2.83	4.22	3.41	2.9
ala	1.22	0	3.55	1.6	1.57	1.21	2.1	2.08	3.63	2.35	2.32	2.61	1.77	2.8	1.41	0.591	1.64	3.84	2.87	1.79
arg	4.66	3.55	0	2	2.03	2.95	1.8	1.84	2.75	2.14	2.55	2.52	2.11	3.66	3.81	2.99	2.04	3.52	3.13	2.71
asn	2.79	1.6	2	0	0.0463	1.44	0.69	0.695	3.04	1.12	1.4	1.64	0.543	2.67	2.14	1.06	0.41	3.43	2.48	1.12
asp	2.75	1.57	2.03	0.0463	0	1.39	0.729	0.734	3.01	1.17	1.43	1.68	0.574	2.67	2.12	1.02	0.4	3.42	2.47	1.14
cys	2.04	1.21	2.95	1.44	1.39	0	2.11	2.11	2.56	2.52	2.72	3	1.91	3.41	2.27	0.858	1.26	3.78	3.16	2.28
gln	3.26	2.1	1.8	0.69	0.729	2.11	0	0.0411	3.4	0.663	0.921	1.05	0.355	2.38	2.28	1.65	1.03	3.3	2.22	0.941
glu	3.24	2.08	1.84	0.695	0.734	2.11	0.0411	0	3.42	0.657	0.892	1.02	0.328	2.35	2.25	1.64	1.04	3.29	2.2	0.909
his	4.21	3.63	2.75	3.04	3.01	2.56	3.4	3.42	0	3.99	4.31	4.38	3.46	4.52	4.04	3.22	2.93	3.43	3.83	4.13
ile	3.52	2.35	2.14	1.12	1.17	2.52	0.663	0.657	3.99	0	0.552	0.818	0.728	2.66	2.6	1.95	1.35	3.87	2.67	0.726
leu	3.41	2.32	2.55	1.4	1.43	2.72	0.921	0.892	4.31	0.552	0	0.426	0.883	2.26	2.32	2.04	1.71	3.77	2.42	0.567
lys	3.66	2.61	2.52	1.64	1.68	3	1.05	1.02	4.38	0.818	0.426	0	1.11	2.06	2.41	2.34	1.98	3.56	2.21	0.956
met	2.94	1.77	2.11	0.543	0.574	1.91	0.355	0.328	3.46	0.728	0.883	1.11	0	2.31	2.02	1.35	0.913	3.37	2.23	0.7
phe	3.26	2.8	3.66	2.67	2.67	3.41	2.38	2.35	4.52	2.66	2.26	2.06	2.31	0	1.59	2.81	3.06	2.3	0.856	2.35
pro	1.67	1.41	3.81	2.14	2.12	2.27	2.28	2.25	4.04	2.6	2.32	2.41	2.02	1.59	0	1.66	2.41	2.97	1.84	2.02
ser	1.77	0.591	2.99	1.06	1.02	0.858	1.65	1.64	3.22	1.95	2.04	2.34	1.35	2.81	1.66	0	1.06	3.68	2.75	1.54
thr	2.83	1.64	2.04	0.41	0.4	1.26	1.03	1.04	2.93	1.35	1.71	1.98	0.913	3.06	2.41	1.06	0	3.73	2.86	1.38
trp	4.22	3.84	3.52	3.43	3.42	3.78	3.3	3.29	3.43	3.87	3.77	3.56	3.37	2.3	2.97	3.68	3.73	0	1.53	3.84
tyr	3.41	2.87	3.13	2.48	2.47	3.16	2.22	2.2	3.83	2.67	2.42	2.21	2.23	0.856	1.84	2.75	2.86	1.53	0	2.51
val	2.9	1.79	2.71	1.12	1.14	2.28	0.941	0.909	4.13	0.726	0.567	0.956	0.7	2.35	2.02	1.54	1.38	3.84	2.51	0

Table 2: Distance matrix of the principal factors of the topological indices from Tab. 1

- branches in chains and rings (e.g. carbons with three C-neighbors that are not part of a ring)
- Cyclic compounds (e.g. 6-rings)
- double and triple bonds in chains and rings (e.g. a double bonds in a 6-ring)
- elements (e.g. hetero atoms)
- functional groups (e.g. aldehyds)
- special classes of compounds (e.g. methyl ester)

Many of these properties can be described in terms of substructures, and so a substructure search is one of the main algorithms for their determination. We used a procedure from [38]. Other properties were calculated by methods described in [25, 39, 40].

In the evaluation the *Tanimoto* coefficient [41] is employed which measures the similarity of two bit strings as

$$T_{i,j} = \frac{2C_{i,j}}{E_i + E_j}$$

where $C_{i,j}$ is the number of properties that are common in the i-th and in the j-th structure, and E_i is the total number of properties of the i-th molecule. Thus T lies in the range $0 \leq T \leq 1$, showing 1 for complete identity and 0 for maximal dissimilarity.
Again we combined all results to a distance matrix D with $d_{i,j} = 1 - T_{i,j}$, shown in Tab. 3. From this matrix we computed spatial coordinates by *multidimensional scaling*. In this method, the geometric distances are chosen to reflect the pairwise distances as close as possible. Fig. 2 shows the coordinates of the structures in three-dimensional space. Also in this case, in a similar way as for the PCA, we can group closely neighbored compounds together (as partially indicated by the ellipses in Fig. 2).

3 Generation of combinatorial libraries

3.1 Reaction schemes

We would like to start the presentation of our method with a syntax describing the underlying chemical reactions formally, especially the two-component synthesis like

$$\mathrm{A} + \mathrm{B} \rightarrow \mathrm{C}$$

In most cases, subgraphs determine the course of the reaction.

3.1 Definition. *Let* (η_1, β_1) *and* (η_2, β_2) *with* $\eta_1 \in m^{\underline{\mathbf{r}}^{[2]}}$ *and* $\eta_2 \in m^{\underline{\mathbf{s}}^{[2]}}$ *be molecular graphs. A* reaction scheme *is defined as the triple* $((\eta_1, \beta_1), (\eta_2, \beta_2), \rho)$ *(or* (η_1, η_2, ρ) *if the atomic type coloring is clear), where* $\rho : \underline{\mathbf{r}} \times \underline{\mathbf{s}} \rightarrow \mathbb{Z} \cup \{-\infty\}$ *is a mapping with*

$$\rho(i,j) = \begin{cases} k & i \text{ and } j \text{ are connected by a bond of degree } k \\ 0 & i \text{ and } j \text{ remain unconnected} \\ -\infty & \text{one of the atoms } i \text{ or } j \text{ is dropped} \end{cases}$$

•

	gly	ala	arg	asn	asp	cys	gln	glu	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
gly	0	0	0.2	0.333	0.2	0.0476	0.333	0.2	0.391	0.1	0.143	0.1	0.0476	0.238	0.182	0.143	0.143	0.538	0.333	0.143
ala	0	0	0.2	0.333	0.2	0.0476	0.333	0.2	0.391	0.1	0.143	0.1	0.0476	0.238	0.182	0.143	0.143	0.538	0.333	0.143
arg	0.2	0.2	0	0.238	0.2	0.238	0.238	0.2	0.304	0.3	0.333	0.2	0.238	0.333	0.364	0.333	0.333	0.538	0.429	0.333
asn	0.333	0.333	0.238	0	0.238	0.364	0	0.238	0.5	0.429	0.455	0.238	0.364	0.455	0.478	0.273	0.273	0.556	0.364	0.455
asp	0.2	0.2	0.2	0.238	0	0.238	0.238	0	0.478	0.3	0.333	0.3	0.238	0.333	0.364	0.238	0.238	0.615	0.333	0.333
cys	0.0476	0.0476	0.238	0.364	0.238	0	0.364	0.238	0.417	0.143	0.182	0.143	0	0.273	0.217	0.182	0.182	0.556	0.364	0.182
gln	0.333	0.333	0.238	0	0.238	0.364	0	0.238	0.5	0.429	0.455	0.238	0.364	0.455	0.478	0.273	0.273	0.556	0.364	0.455
glu	0.2	0.2	0.2	0.238	0	0.238	0.238	0	0.478	0.3	0.333	0.3	0.238	0.333	0.364	0.238	0.238	0.615	0.333	0.333
his	0.391	0.391	0.304	0.5	0.478	0.417	0.5	0.478	0	0.478	0.5	0.391	0.417	0.417	0.2	0.5	0.5	0.31	0.5	0.5
ile	0.1	0.1	0.3	0.429	0.3	0.143	0.429	0.3	0.478	0	0.0476	0.2	0.143	0.333	0.273	0.238	0.238	0.538	0.429	0.0476
leu	0.143	0.143	0.333	0.455	0.333	0.182	0.455	0.333	0.5	0.0476	0	0.238	0.182	0.364	0.304	0.273	0.273	0.556	0.455	0
lys	0.1	0.1	0.2	0.238	0.3	0.143	0.238	0.3	0.391	0.2	0.238	0	0.143	0.333	0.273	0.238	0.238	0.462	0.429	0.238
met	0.0476	0.0476	0.238	0.364	0.238	0	0.364	0.238	0.417	0.143	0.182	0.143	0	0.273	0.217	0.182	0.182	0.556	0.364	0.182
phe	0.238	0.238	0.333	0.455	0.333	0.273	0.455	0.333	0.417	0.333	0.364	0.333	0.273	0	0.304	0.364	0.364	0.407	0.0909	0.364
pro	0.182	0.182	0.364	0.478	0.364	0.217	0.478	0.364	0.2	0.273	0.304	0.273	0.217	0.304	0	0.304	0.304	0.357	0.391	0.304
ser	0.143	0.143	0.333	0.273	0.238	0.182	0.273	0.238	0.5	0.238	0.273	0.238	0.182	0.364	0.304	0	0.0909	0.63	0.273	0.273
thr	0.143	0.143	0.333	0.273	0.238	0.182	0.273	0.238	0.5	0.238	0.273	0.238	0.182	0.364	0.304	0.0909	0	0.63	0.273	0.273
trp	0.538	0.538	0.538	0.556	0.615	0.556	0.556	0.615	0.31	0.538	0.556	0.462	0.556	0.407	0.357	0.63	0.63	0	0.481	0.556
tyr	0.333	0.333	0.429	0.364	0.333	0.364	0.364	0.333	0.5	0.429	0.455	0.429	0.364	0.0909	0.391	0.273	0.273	0.481	0	0.455
val	0.143	0.143	0.333	0.455	0.333	0.182	0.455	0.333	0.5	0.0476	0	0.238	0.182	0.364	0.304	0.273	0.273	0.556	0.455	0

Table 3: Distance matrix of the Tanimoto coefficients of the property vectors

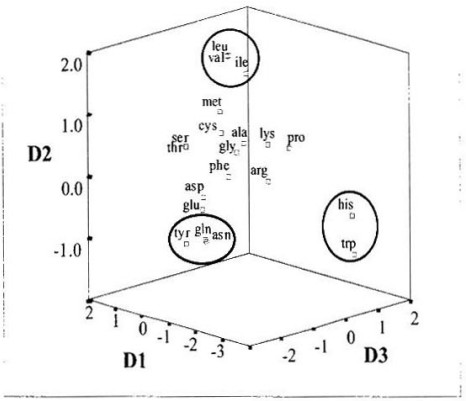


Figure 2: Distribution of the compounds in multidimensional scaling of the distance matrix in Tab. 3

By means of this definition[2] many two component reactions can be described sufficiently. Our main interest in such a reaction is in fact the change of the graphs, and not the experimental aspects (like reaction conditions, catalysts or equilibria).
A corresponding algorithm could be formulated for linking two graphs by a reaction scheme over all reacting subgraphs. As we will not explicitly need such a procedure for library generation, we omit a deeper discussion and just present an example:

3.2 Example. Peptids are protein molecules built from at least two amino acids which play a central role in biochemistry. The joining of the single amino acids is performed by condensation of the acid group (COOH) and the amid group (NH_2). Thus the decisive reaction structure is the acid amid group, which must be contained in both reaction partners.

$$\begin{array}{c} {}^{3}\mathrm{N}\quad \mathrm{O}^{4} \\ \quad | \quad \| \\ -{}^{1}\mathrm{C}-\mathrm{C}^{2}-\mathrm{O}^{5} \end{array}$$

$$\mathbf{I}\colon \eta_1 = \eta_2$$

The condensation is represented by the mapping

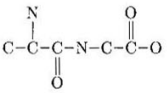
$$\rho = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -\infty & -\infty & -\infty & -\infty & -\infty \end{pmatrix}$$

[2]This definition is a simplification of the situation and is only used for a formalization of the construction problem discussed below. For more sophisticated purposes more comprehensive approaches like the algebra of *bc-&r*-matrices of [42] are necessary.

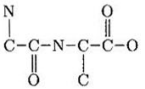
We consider the amino acids

II: Alanin and  **III**: Glycin

Obviously both contain the subgraph η_1. Despite the equality of the subgraphs in the reaction scheme, the order of the initial graphs is essential. Taking alanin as the first one, we obtain:



Glycin as γ_1 yields:



◇

3.2 Group actions and orderly generation

We will introduce some basic definitions and notations from algebra which will be needed furtheron (for more details, see, e.g., [43]).

3.3 Definition. *Let G be a finite group, and Ω a finite non-empty set. A mapping*

$$\Omega \times G \to \Omega,\ (\omega, g) \mapsto \omega^g$$

with $(\omega^g)^{g'} = \omega^{gg'}\ \forall g, g' \in G\, \forall \omega \in \Omega$ and $\omega^{id} = \omega$ is called an action of G on Ω.

•

Group actions give rise to several important sets:

3.4 Definition. *Let G act on Ω, $\omega \in \Omega$ and $\Delta \subseteq \Omega$.*

- $\omega^G := \{\omega^g \mid g \in G\}$ *is called* orbit *of* ω.
- $\Omega /\!\!/ G := \{\omega^G \mid \omega \in \Omega\}$ *is called* set of orbits.
- $\mathcal{T}(\Omega /\!\!/ G)$ *is called* transversal of the orbits *with* $\Omega = \dot{\bigcup}_{t \in \mathcal{T}} G(t)$, *derived from the equivalence class property of the orbits.*
- $\Omega_g := \{\omega \in \Omega \mid \omega^g = \omega\}$ *the* set of fixed points *of* g.
- $C_G(\Delta) := \{g \in G \mid \delta^g = \delta\ \forall \delta \in \Delta\}$ *is called* centralizer *or* pointwise stabilizer *of* Δ *in* G.
- $N_G(\Delta) := \{g \in G \mid \delta^g \in \Delta\ \forall \delta \in \Delta\}$ *is called* normalizer *or* setwise stabilizer *of* Δ *in* G.

•

Let X and Y two finite non-empty sets. Then we set $Y^X := \{f \mid f : X \to Y\}$. If G acts on X, then G also acts on Y^X as

$$Y^X \times G \to Y^X,\ (f,g) \mapsto f^g \text{ with } f^g(x) = f(x^{g^{-1}})$$

Typical sets of this kind will be $\underline{\mathbf{n}} := \{1, 2, \ldots, n\}$. A frequently used group is the *symmetric group* $S_n := \{\pi \in \underline{\mathbf{n}}^{\underline{\mathbf{n}}} \mid \pi \text{ bijective}\}$.

For constructing transversals of orbits, the naive approach is to compare any new element with all previously calculated; but this is completely inappropriate for practical use. A helpful principle is the concept of *orderly generation*, a method that was introduced by R. C. Read [44] and that can be refined considerably [43, 45, 46, 47]. It is based on the fact that total orders on X and Y induce a canonic total order on Y^X, the lexicographic order, so that a *canonic transversal*

$$\mathcal{T}_>(Y^X /\!\!/ G),$$

consisting of the biggest elements of the orbits does exist. The decisive result that can be derived from the general form reads:

3.5 Proposition. *Let $f \in \mathcal{T}_>(Y^X /\!\!/ G)$ and $f_1 \in Y^X$ a* starting piece *of f, i.e. there exists a $t \leq n$ with*

$$f_1(j) = \begin{cases} f(j) & \text{for } j < t \\ 0 & \text{for } j \geq t \end{cases}$$

Then $f_1 \in \mathcal{T}_>(Y^X /\!\!/ G)$.

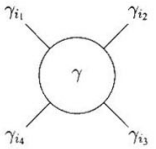
This proposition tells us how to use orderly generation algorithmically: It suffices to expand starting pieces lexicographically without having to re-test them on maximality. This means in the opposite that, if the starting piece is already not canonic, it cannot become a canonic representative by the filling of the remaining places. (For more details see [45, 46].)

3.3 Multiple attachments to a core structure

A special type of reaction scheme is given by a core structure with several reaction sites and a number of ligand compounds.

Let $((\eta_1, \chi_1), (\eta_2, \chi_2), \rho)$ denote a reaction scheme, (γ, β) a molecular graph containing k substructures isomorphic to (η_1, χ_1) with $k > 1$, and $(\gamma_1, \beta_1), \ldots, (\gamma_n, \beta_n)$ a number of molecular graphs; for sake of simplicity we assume that each of them contains exactly one substructure isomorphic to (η_2, χ_2).

The first task is to determine all attachments of the ligands to the sites of the core, where the sites are given by the substructures of the reaction scheme. For $k = 4$, e.g., the situation is:



Topological equivalence among the sites is described by the permutation group $P_\gamma \leq S_k$ which is induced by the automorphism group $Aut(\gamma, \beta) := C_{S_p}(\gamma, \beta)$ of the molecular graph. So P_γ acts on the sites $\underline{\mathbf{k}}$ which shall be assigned with n different ligands. Summarizing our arguments we get:

3.6 Lemma. *The essentially different possibilities to attach n ligand structures, which contain the corresponding subgraph of the given reaction scheme exactly once, to the k different reaction sites of a core structure (γ, β) correspond "one-to-one" and "onto" to a transversal of the action of P_γ on $\underline{\mathbf{n}}^{\underline{\mathbf{k}}}$:*

$$\mathcal{T}(\underline{\mathbf{n}}^{\underline{\mathbf{k}}} /\!/ P_\gamma).$$

So now we can formulate a strategy:

3.7 Algorithm (Attachment of ligands to a core structure).

i) Determine the group P_γ as well as all subgraphs $\zeta_1, \ldots, \zeta_k \subseteq \gamma$ which are isomorphic to η_1.

ii) Compute for $i \in \underline{\mathbf{n}}$ the subgraphs $\zeta^{(i)} \subseteq \gamma_i$ which are isomorphic to η_2.

iii) Use orderly generation in order to obtain the next representative $f \in \underline{\mathbf{n}}^{\underline{\mathbf{k}}}$ under the action of P_γ.

iv) Determine the total graph which yields from the attachment of the ligands $\gamma_{f(1)}, \ldots, \gamma_{f(k)}$ to γ according to ρ, i.e. by combining the graphs, eliminating vertices which have to be dropped and adding the necessary edges.

v) If there are further orbit representatives, go to step iii.

∇

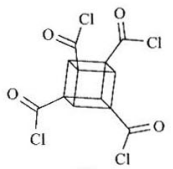
Due to orderly generation in step iii and the uniqueness of the subgraphs of the ligands we only obtain non-isomorphic solutions.

3.4 Single step generation of libraries

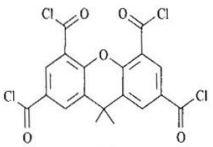
For the generation of a combinatorial library from given building-blocks algorithm 3.7 is perfectly suited, since the basic situation of combinatorial chemistry as described in sec. 1 is just that of this method.[3]

For practical use it is moreover relevant that the multiplicity of a certain building-block can be restricted, i.e. that a (γ_i, β_i) occurs in all compounds of the library at least r and at most s times. This can be reached by an additional test in 3.7 between step iii and step iv. In laboratory, this restriction can be satisfied by an appropriate modification of the reaction conditions.

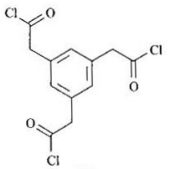
As an example we consider the combinatorial libraries from [5]. The authors used as building-blocks the twenty natural amino acids (cf. Fig. 1) and as core structures some acid chlorides:



IV



V



VI

[3] We assume that each building-block is admissible for each site. In the other case additional rules must be formulated.

a cubane-derivative (structure **IV**), xanthene (**V**) and a benzene triacid chloride (**VI**).
The reaction scheme consists of the substructures

and the matrix

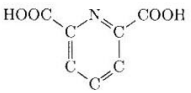
$$\rho = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -\infty & -\infty & -\infty & -\infty & -\infty \end{pmatrix}.$$

The cubane-derivative **IV** is the structure with the highest symmetry, i.e. the largest automorphism group (derived from the symmetry group of the cube), which has 24 elements. Since each automorphism includes a movement of the reacting substructures, we also have $|P_\gamma| = 24$. As there are just four sites, it turns out that $P_\gamma = S_4$.
The xanthene **V** has an automorphism group with four elements. Two of them include the exchange of the methylene groups on the carbon bridge atom, and thus $|P_\gamma| = 2$. Besides the identity, this is the reflection of the rings on the vertical symmetry axis.
The benzene triacid chloride **VI** has cyclic symmetry. Thus P_γ equals the cyclic group C_3, having three elements.

Even though the symmetry situation is a little complicated for one of the three cores only, the advantages of the mathematical concept behind algorithm 3.7 are obvious. The general *Ansatz* with an arbitrary permutation group and the efficient orderly generation (cf. [45, 46]) allows a very rapid generation of the combinatorial libraries in all three cases. The computing speed is about 40 structures per second on a Pentium 90 MHz PC.
Details about the sizes of the libraries are given in section 4. Fig. 3 shows six molecules from each of the three libraries.[4]

3.5 Multiple step methods

It is also possibile to carry out several reactions one after the other using the products of one reation as core structures of the next one. The mathematical situtation is very similar to that of the previous section; due to the differences of the cores the irredundancy remains guaranteed.
As an example we will consider a seven component reaction from [50]. The basic core is pyridin 2,6-dicarbon acid



VII

[4]The 2D placements were automatically calculated by the drawing module of **MOLGEN** [25, 39, 48]. These pictures reveal the current inacurracies of the employed placement algorithm [39, 49] for combinatorial libraries.

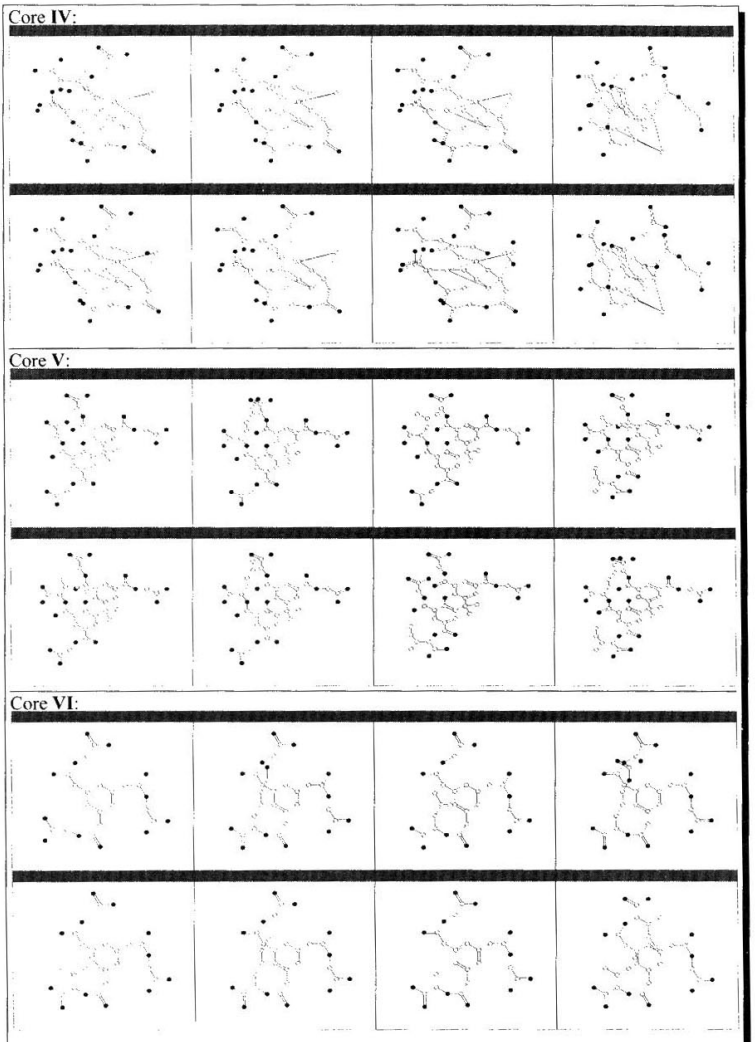
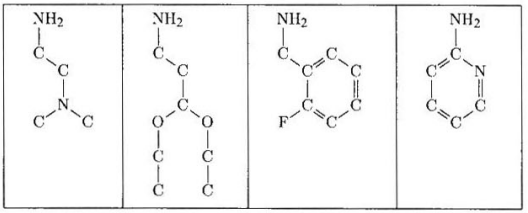


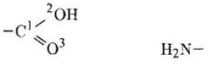
Figure 3: Extracts from the combinatorial libraries produced from the structures **IV**, **V** and **VI** and the natural amino acides from Fig. 1

Our graph-theoretic model of molecules cannot sufficiently represent the aromaticity of this structure. Therefore we use an additional algorithm for eliminating aromatic mesomers (by marking the aromatic bonds, canonical numbering and hashing). To keep the following discussion transparent, we will assume the topological equivalence of the two COOH groups in **VII**. Then P_γ has two elements.
The reaction steps are in particular:

i) In the first step, four amides are added:



The reaction scheme consists of the substructures

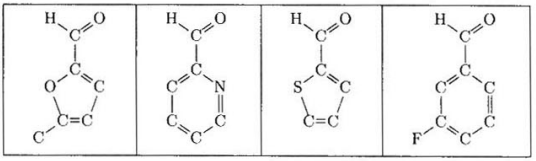


and of the matrix

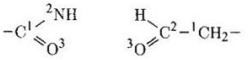
$$\rho = \begin{pmatrix} 1 \\ -\infty \\ 0 \end{pmatrix}.$$

This yields a total of 10 compounds.

ii) The second step takes four aromatic aldehyds as building-blocks:



The corresponding reaction scheme applies the substructures

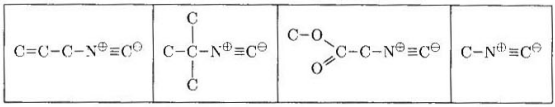


and the matrix

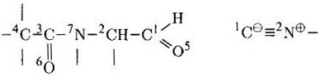
$$\rho = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Selecting all ten molecules from step i as core, one obtains altogether 136 structures.

iii) The last step consists of the addition of four isocyanides (where the charges are considered as an additional label of the atoms, the valences of which are altered respectively):



Here the substructures of the reaction scheme are (a little longer to ensure uniqueness)



and the matrix

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

which may also imply the neutralization of the nitrogen atom with the proton at carbon atom 1 of the second substructure.

Attaching these building-blocks to the cores generated in step ii in all essentially different ways according to the scheme yields a library of 2080 constitutionally different compounds. If stereoisomerism is additionally taken into account (calculated, e.g., after [24]), we obtain 8256 products.

A (randomly selected) extract of 15 molecules is shown in Fig. 4.

4 Enumeration of libaries

In this section we will present methods for the enumeration of the sizes of combinatorial libraries. A key tool for enumeration in algebraic combinatorics is the lemma of Cauchy-Frobenius:

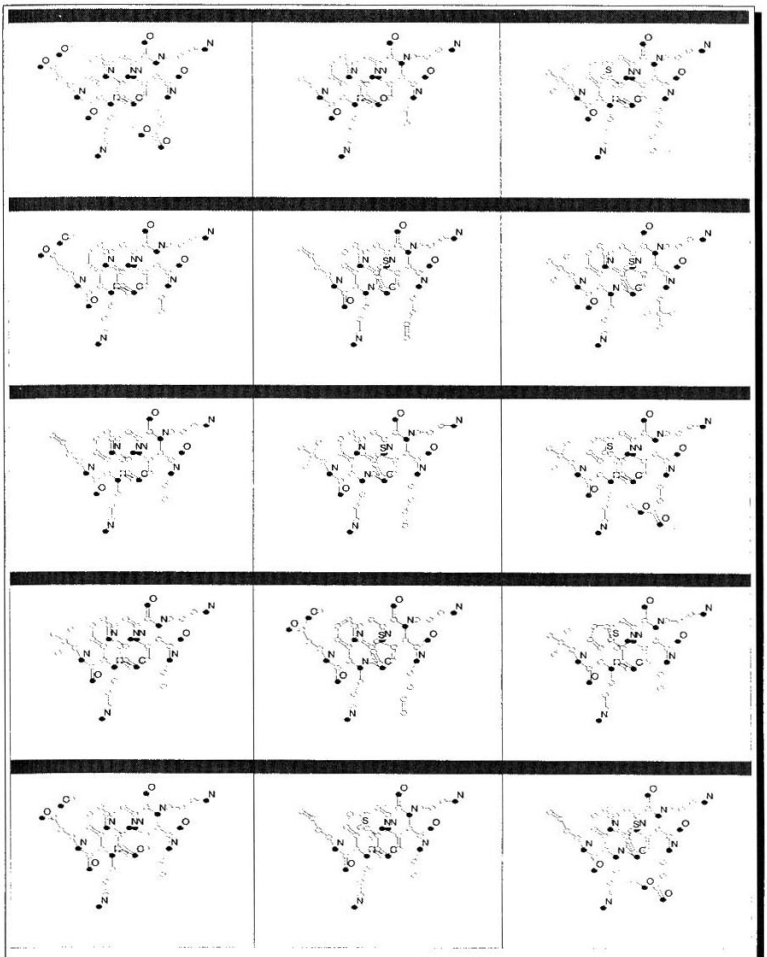


Figure 4: Extract of 15 randomly selected molecules from the library after [50]

4.1 Proposition (The Lemma of Cauchy-Frobenius). *Let G a finite permutation group acting on a finite set X.*

- *The number of orbits of this action is $|X/\!/G| = \frac{1}{|G|}\sum_{g\in G}|X_g|$.*
- *G also acts on Y^X if Y denotes another finite set. For the number of orbits the following equation holds: $|Y^X/\!/G| = \frac{1}{|G|}\sum_{g\in G}|Y|^{c(g)}$, where $c(g)$ is the number of cycles of the permutation g.*

As we saw in lemma 3.6 that the attachment of building-blocks to core structure can be represented by a group action, we can immediately state a result on the size of such libraries:

4.2 Theorem. *Let (γ,β) a core structure with k different reaction sites. $Aut(\gamma,\beta)$ may induce a permutation group $P_\gamma \leq S_k$ among the sites.*
Furthermore a set of n different building-blocks may be given. Then the combinatorial library which can be built from the core structure and the building-blocks according to a corresponding reaction scheme has

$$|\underline{\mathbf{n}}^{\mathbf{k}}/\!/P_\gamma| = \frac{1}{|P_\gamma|}\sum_{\pi\in P_\gamma} n^{c(\pi)}$$

elements.

4.3 Example. Again we consider the example from section 3.4 (taken from [5]) with the cores **IV**, **V** und **VI** and the amino acids (see Fig. 1) as building-blocks.

- The group P_γ for the cubane derivative (**IV**) is the symmetric group S_4 with 24 elements and $k = 4$. Then we get by Theorem 4.2:

$$|\underline{\mathbf{n}}^{\mathbf{4}}/\!/P_\gamma| = \tfrac{1}{24}(n^4 + 6n^3 + 11n^2 + 6n)$$

- For xanthene (structure **V**), we have $k = 4$ and $P_\gamma = \{1, (12)(34)\}$. Here our formula yields:

$$|\underline{\mathbf{n}}^{\mathbf{4}}/\!/P_\gamma| = \tfrac{1}{2}(n^4 + n^2)$$

- In case of the benzene triacid chloride (**VI**) there is $k = 3$ and $P_\gamma = \{1, (123), (132)\}$. So the equation reads:

$$|\underline{\mathbf{n}}^{\mathbf{3}}/\!/P_\gamma| = \tfrac{1}{3}(n^3 + 2n)$$

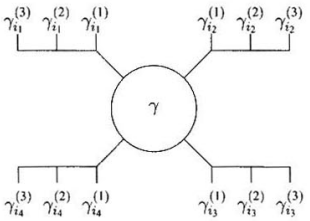
The following table provides an overview over the sizes of libraries depending on the number of building-blocks used:

n	**IV**	**V**	**VI**
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

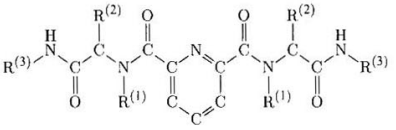
Although there are equally many sites in **IV** and **V**, the libraries with the first one are considerably smaller due to the higher symmetry of the core.

◇

The situation for multi-step procedures as in section 3.5 is a little more complicated. We will now consider the attachment of the building-blocks as a whole process, as if all blocks are added simultaneously. For the construction, this view is less sensible; it is, however, necessary for enumeration, in order to be able to determine the dependencies of the sites correctly.
So the situation can be visualized by the following sketch:



For our pyridine core from [50] this is



VIII

Let k again denote the number of different attachment sites. Furthermore let $\kappa = (\kappa_1, \ldots, \kappa_m) \models k$ be a partition of k (i.e. $\kappa_i \in \mathbb{N}$ and $\sum_{i=1}^{m} \kappa_i = k$) and $k_i := \sum_{s=1}^{i} \kappa_s$ (with $k_0 := 0$) the cumulated entries of the partition. The sites are supposed to be partitioned according to the orbits of the automorphism group of γ, i.e. we assume for P_γ:

$$\forall \pi \in P_\gamma \leq S_k : \quad k_i < r \leq k_{i+1} \Rightarrow k_i < r^\pi \leq k_{i+1} \text{ for } i = 0, \ldots, m$$

By $c_i(\pi)$ we denote the number of cycles of $\pi \in P_\gamma$ on $\{k_i, \ldots, k_{i+1} - 1\}$.
The building-blocks may be elements of the sets $A_1, \ldots, A_m$ with $|A_i| =: n_i$. So we obtain for the sizes of libraries:

4.4 Theorem. *Let core structure and building-blocks be as defined above. Then a bijection exists from the combinatorial library to a tranversal of the orbits of the action of P_γ on the mappings*

$$\Phi := \underline{\mathbf{n_1}}^{\kappa_1} \times \ldots \times \underline{\mathbf{n_m}}^{\kappa_m}.$$

For the numbers of elements of the library the following equation holds:

$$|\Phi /\!\!/ P_\gamma| = \frac{1}{|P_\gamma|} \sum_{\pi \in P_\gamma} \prod_{i=1}^{m} n_i^{c_i(\pi)}$$

4.5 Example. For the library of the multi-component reaction from [50] which we generated in section 3.5, we have $k = 6$, $\kappa = (2, 2, 2)$ and $P_\gamma = \{1, (12)(34)(56)\}$. (This means for the molecular graph **VIII** that the positions of the rests $R^{(1)}$, $R^{(2)}$, and $R^{(3)}$ can be interchanged correspondingly.)
Then we get by Theor. 4.4:

$$|\Phi /\!\!/ P_\gamma| = \tfrac{1}{2}(n_1^2 n_2^2 n_3^2 + n_1 n_2 n_3)$$

The following table shows values for this formula. n_1 and n_2 vary vertically and horizontally, respectively; the third value is kept fixed as $n_3 = 4$. The entry for $(4, 4)$ tells us the well-known size 2080 for 4 building-blocks each (cf. sec. 3.5).

$n_1 \backslash n_2$	1	2	3	4	5	6	7	8	9	10
1	10	36	78	136	210	300	406	528	666	820
2	36	136	300	528	820	1176	1596	2080	2628	3240
3	78	300	666	1176	1830	2628	3570	4656	5886	7260
4	136	528	1176	*2080*	3240	4656	6328	8256	10440	12880
5	210	820	1830	3240	5050	7260	9870	12880	16290	20100
6	300	1176	2628	4656	7260	10440	14196	18528	23436	28920
7	406	1596	3570	6328	9870	14196	19306	25200	31878	39340
8	528	2080	4656	8256	12880	18528	25200	32896	41616	51360
9	666	2628	5886	10440	16290	23436	31878	41616	52650	64980
10	820	3240	7260	12880	20100	28920	39340	51360	64980	80200

◇

A combinatorial enumeration can also obtained, if for single-step procedures not all building-blocks shall be allowed for all possible multiplicities, i.e. if the frequencies shall be restricted. The tool for this task is called *weighted enumeration* (cf. [43], also for proofs).

4.6 Proposition.

i) Weighted form of the Cauchy-Frobenius lemma: *Let G be finite group acting on a finite set X and $W : X \to \mathbb{Q}$ a function. If W is constant on the orbits of G on X (a so-called* weight *function), then we have for any transversal $\mathcal{T}$ of the orbits:*

$$\sum_{t\in\mathcal{T}} W(t) = \frac{1}{|G|}\sum_{g\in G}\sum_{x\in X_g} W(x)$$

ii) Let $w : Y^X \to \mathbb{Q},\ f \mapsto \prod_{x\in X} W(f(x))$ denote the multiplicative weight *for the weight function $W : Y \to \mathbb{Q}$. Then w is constant on the orbits of the permutation group G on Y^X and for any transversal $\mathcal{T}$ of the orbits we have:*

$$\sum_{t\in\mathcal{T}} w(t) = \frac{1}{|G|}\sum_{g\in G}\prod_{i=1}^{|X|}\left(\sum_{y\in Y} W(y)^i\right)^{a_i(g)}$$

where $a_i(g)$ denotes the number of cycles of length i in the permutation g.

iii) Let $c(f,\_) : Y \to \mathbb{N},\ y \mapsto |f^{-1}(\{y\})|$ indicate the content *of the mapping $f \in Y^X$, i.e. $c(f,y)$ denotes how often f takes the value y. Then the number of G-orbits on Y^X, the elements of which have the same content as $f \in Y^X$, is equal to the coefficient of the monomial $\prod_y y^{c(f,y)}$ in the polynomial*

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

Applying this results to combinatorial libraries we obtain:

4.7 Theorem. *Let (γ,β) be a core structure with k different reaction sites. $Aut(\gamma,\beta)$ induces a permutation group, $P_\gamma \leq S_k$, among the sites.*
Furthermore a set of n different building-blocks and a distribution $f \in \underline{\mathbf{n}}^{\mathbf{k}}$ of the blocks may be given.
Then the number of elements of the library, the distributions of which have the same content $c(f,\_)$ as f, is equal to the coefficient of the monomial $\prod_r y_r^{c(f,y_r)}$ in the polynomial

$$\frac{1}{|P_\gamma|}\sum_{\pi\in P_\gamma}\prod_{i=1}^{k}\left(\sum_{r=1}^{n} {y_r}^i\right)^{a_i(\pi)}$$

over the unknowns $y_1,\ldots,y_n$.

4.8 Example. As above, we consider the example from sec. 3.4 (taken from [5]) with the core structures **V** und **VI** and the amino acids as building-blocks.
Then Theor. 4.7 yields the following relations:

- For xanthene (**V**) we obtain the polynomial

$$\tfrac{1}{2}\left((y_1+\ldots+y_n)^4+(y_1^{\,2}+\ldots+y_n^{\,2})^2\right)$$

In the case $n=4$, this leads to the sum

$$\begin{array}{l}6\,y_1y_2^2y_4+6\,y_1y_2y_4^2+6\,y_1y_2^2y_3+y_1^4+y_2^4+y_3^4+y_4^4+6\,y_1y_3^2y_4+\\6\,y_1^2y_2y_3+6\,y_1^2y_2y_4+6\,y_1^2y_3y_4+6\,y_2^2y_3y_4+2\,y_1y_2^3+2\,y_1y_3^3+\\2\,y_1y_4^3+2\,y_1^3y_2+2\,y_1^3y_3+2\,y_1^3y_4+4\,y_1^2y_2^2+4\,y_1^2y_3^2+\\4\,y_1^2y_4^2+2\,y_2y_3^3+2\,y_2y_4^3+2\,y_2^3y_3+2\,y_2^3y_4+4\,y_2^2y_3^2+\\4\,y_2^2y_4^2+2\,y_3y_4^3+2\,y_3^3y_4+4\,y_3^2y_4^2+6\,y_1y_2y_3^2+6\,y_1y_3y_4^2+\\12\,y_1y_2y_3y_4+6\,y_2y_3^2y_4+6\,y_2y_3y_4^2\end{array}$$

The condition that, for instance, only those library elements are of interest which contain the first building-block exactly once corresponds to the summands

$$\begin{array}{l}6\,y_1y_2^2y_4+6\,y_1y_2y_4^2+6\,y_1y_2^2y_3+6\,y_1y_3^2y_4+2\,y_1y_2^3+\\2\,y_1y_3^3+2\,y_1y_4^3+6\,y_1y_2y_3^2+6\,y_1y_3y_4^2+12\,y_1y_2y_3y_4.\end{array}$$

So there are $6+6+6+6+2+2+2+6+6+12=54$ elements of that kind.

- For the benzene triacid chloride (**VI**) the polynomial is

$$\tfrac{1}{3}\left((y_1+\ldots+y_n)^3+2(y_1^{\,3}+\ldots+y_n^{\,3})\right)$$

Considering three building-blocks, say, this means

$$y_1^3+y_1^2y_2+y_1^2y_3+y_1y_2^2+2\,y_1y_2y_3+y_1y_3^2+y_2^3+y_2^2y_3+y_2y_3^2+y_3^3$$

◇

5 Conclusion

The methods and theorems presented above provide a number of tools for the analysis of combinatorial chemistry processes. They should enable the researcher to understand and overview his libraries better and eventually to design his experiments more precisely.
As explained in the introduction, one of our aims is to investigate the use of simulations in combinatorial chemistry. So a few words about the third step, the *screening of the libraries* are necessary. This aspect is indeed the most difficult one. In the view of the present level of research a purely *virtual screening* – only in the computer and without experiment – is impractical due to a large computational expenditure.

A lot of known biological mechanisms are based on the principle that the drug interacts with a usually much bigger biomolecule (protein) (s. e.g. [51, 52]). There the drug is also called *ligand* and the protein *receptor*. The ligand-receptor-interaction depends qualitatively and quantitatively on the spatial structures of both partners.
A possible strategy is to bring some light into the darkness by examination of quantitative structure-activity relationships (QSAR) (see also sec. 2.2). Here a set of sample substances is used to empirically determine one (or several) activity parameters. From the results a correlation with computable structural properties is established by statistical methods. This correlation is

afterwards employed for extrapolation on a larger set of compounds (like a combinatorial library, cf. [10, 15, 19, 23]).
More elaborate techniques are *3D QSAR* [52, 53] and *comparative molecular field analysis (CoMFA)* [53, 54]. Here the library elements are first converted to three-dimensional structures by an appropriate method (like distance geometry programs [55], conformation analysis methods [56], expert systems [57, 58] or force field calculations, e.g. [59]. The crucial feature such a program is required to have is that the computed conformation must be reasonable for the active site, as the usual software packages produce conformations *in vacuo* or in solution.) Then a superposition of the ligand and the binding site is calculated according to the physical fields (steric and electrostatic) of the molecules in order to be able to estimate the activity of the ligand.

So virtual screening and thus *virtual combinatorial chemistry synthesis* is today not in competition with high-throughput screening robots but – as there are many simulation methods in development – may be the reality of tomorrow.

Acknowledgement

The author is grateful to the German Federal Ministry of Education, Science, Research and Technology (BMBF) for financial support and to the large number of researchers in the institute and on the Internet for useful hints and discussions.

References

[1] GALLOP, M.A., R.W. BARRETT, W.J. DOWER, S.P.A. FODOR, AND E.M. GORDON. Applications of combinatorial technologies to drug discovery. 1. Background and peptide combinatorial libraries. *J. Med. Chem.*, **1994**, 37, 1233–1251.

[2] GORDON, E.M., R.W. BARRETT, W.J. DOWER, S.P.A. FODOR, AND M.A. GALLOP. Applications of combinatorial technologies to drug discovery. 2. Combinatorial organic synthesis, library screening strategies, and future directions. *J. Med. Chem.*, **1994**, 37, 1385–1401.

[3] UGI, I. Fast and permanent changes in preparative and pharmaceutical chemistry through multicomponent reactions and their 'libraries'. *Proc. Eston. Acad. Sci. Chem.*, **1995**, 44, 237–273.

[4] WEBER, L., S. WALLBAUM, C. BROGER, AND K. GUBERNATOR. Optimierung der biologischen Aktivität von kombinatorischen Verbindungsbibliotheken durch einen genetischen Algorithmus. *Angew. Chemie*, **1995**, 107, 2452–2453.

[5] CARELL, T., E.A. WINTNER, A.J. SUTHERLAND, J. REBEK, JR., Y.M. DUNAYEVSKIY, AND P. VOUROS. New promise in combinatorial chemistry: synthesis, characterization, and screening of small-molecule libraries in solution. *Chem. & Biol.*, **1995**, 2, 171–183.

[6] FROBEL, K. AND T. KRÄMER. Trendbericht: Kombinatorische Chemie. *Nachr. Chem. Tech. Lab.*, **1996**, 44, 160–162.

[7] JOHNSON, M.A. AND G.M. MAGGIORA, editors. *Concepts and Applications of Molecular Similarity*, New York, NY, 1990. J. Wiley & Sons.

[8] MARTIN, E.J., J.M. BLANEY, M.A. SIANI, D.C. SPELLMEYER, A.K. WONG, AND W.H. MOOS. Measuring diversity: experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.*, **1995**, 38, 1431–1436.

[9] PAVIA, M.R. The chemical generation of molecular diversity. *NetworkScience*, **Aug. 1995**.

[10] KIER, L.B. AND L.H. HALL. *Molecular Connectivity in Structure-Activity Analysis*. Research Studies Press, Chichester, 1986.

[11] JOHNSON, M.A., NAIM. M., V. NICHOLSON, AND C.C. TSAI. Unique mathematical features of the substructure metric approach to quantitative molecular similarity analysis. In KING, R.B. AND D.H. ROUVRAY, editors, *Graph Theory and Topology in Chemistry*, 219–225. Elsevier Science Pub., Amsterdam, 1987.

[12] BASAK, S.C., V.R. MAGNUSON, G.J. NIEMI, AND R.R. REGAL. Determining structural similarity of chemicals using graph-theoretic indices. *Discr. Appl. Math.*, **1988**, 19, 17–44.

[13] SHERIDAN, R.P. AND S.K. KEARSLEY. Using a genetic algorithm to suggest combinatorial libraries. *J. Chem. Inf. Comput. Sci.*, **1995**, 35, 310–320.

[14] AUER, C.M., J.V. NABHOLZ, AND K.P. BAETCKE. Mode of action and the assessment of chemical hazards in the presence of limited data: use of structure-activity relationships (SAR) under TSCA, Section 5. *Environ. Health Persp.*, **1990**, 87, 183–197.

[15] BASAK, S.C. Use of molecular complexity indices in predictive pharmacology and toxicology: a QSAR approach. *Med. Sci. Res.*, **1987**, 15, 605–609.

[16] DEBNATH, A.K., G. DEBNATH, A.J. SHUSTERMAN, AND C. HANSCH. A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test: mutagenicity of aromatic and heterocyclic amines in Salmonella typhimurium TA98 and TA100. *Environ. Mol. Mutagen.*, **1992**, 19, 37–52.

[17] RICHARDS, W.G. *Quantum Pharmacology*. Butterworth, London, 2nd edition, 1983.

[18] STUPER, A.J., W.E. BRUGGER, AND P.C. JURS. *Computer-Assisted Studies of Chemical Structure and Biological Function*. John Wiley & Sons, New York, NY, 1979.

[19] BASAK, S.C., S. BERTELSEN, AND G.D. GRUNWALD. Application of graph-theoretical parameters in quantifying molecular similarity and structure-activity studies. *J. Chem. Inf. Comput. Sci.*, **1994**, 34, 270–276.

[20] BONCHEV, D. *Information Theoretic Indices for Characterization of Chemical Structures*. Research Studies Press, Chichester, 1983.

[21] RANDIĆ, M. *J. Chem. Inf. Comput. Sci.*, **1984**, 24, 164–175.

[22] ROUVRAY, D.H. *Sci. Am.*, **1986**, 255, 40–47.

[23] WIELAND, T. The use of structure generators in predictive pharmacology and toxicology. *Arzneim.-Forsch./Drug Res.*, **1996**, 46 (I), 223–227.

[24] WIELAND, T. Erzeugung, Abzählung und Konstruktion von Stereoisomeren. *MATCH*, **1994**, 31, 153–203.

[25] WIELAND, T., A. KERBER, AND R. LAUE. Principles of the generation of constitutional and configurational isomers. *J. Chem. Inf. Comput. Sci.*, **1996**, 36, 413–419.

[26] FLOYD, R.W. Algorithm 97, Shortest path. *Comm. Assoc. Comp. Mach.*, **1962**, 5, page 345.

[27] WIENER, H. Structural determination of paraffin boiling point. *J. Amer. Chem. Soc.*, **1947**, 69, 17–20.

[28] BALABAN, A.T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.*, **1982**, 89, 399–404.

[29] ROUVRAY, D.H. Should we have designs on topological indices? In KING, R.B., editor, *Chemical Applications of Topology and Graph Theory*, 159–177. Elsevier, Amsterdam, 1983.

[30] SHANNON, C.F. The mathematical theory of communications. *Bell Systems Tech. J.*, **1948**, 27, 379–423.

[31] BRILLOUIN, L. *Science and Information Theory*. Academic Press, New York, NY, 1956.

[32] BONCHEV, D. AND N. TRINAJSTIĆ. Information theory, distance matrix and molecular branching. *J. Chem. Phys.*, **1977**, 67, 4517–4533.

[33] RASHEVSKY, N. Life, information theory and topology. *Bull. Math. Biophys.*, **1955**, 17, 229–235.

[34] HENNEBERG, D. AND B. WEIMANN. *MassLib, Evaluation of low resolution mass spectra series*. Max-Planck-Institut für Kohlenforschung, Mülheim/Ruhr, 1992. Version 7.2.

[35] SCSIBRANY, H. AND K. VARMUZA. *Handbuch zu ToSIM (Software zur Untersuchung von topologischen Ähnlichkeiten in Molekülen)*. Technische Universität, Wien, 1994.

[36] VARMUZA, K. AND H. SCSIBRANY. Clusteranalyse isomerer chemischer Strukturen basierend auf binären Deskriptoren und der Hauptkomponentenanalyse. In *9. CiC-Workshop*, Bitterfeld, 1994. (Posterpräsentation).

[37] VARMUZA, K., W. WERTHER, F. STANCL, A. KERBER, AND R. LAUE. Computer-assisted structure elucidation of organic compounds, based on mass spectra classification and exhaustive isomer generation. In GASTEIGER, J., editor, *Software-Entwicklung in der Chemie 10*, 303–314. GDCh, Frankfurt am Main, 1996.

[38] BENECKE, C. *Algorithmen zur Klassifizierung diskreter Strukturen*. PhD thesis, Universität Bayreuth, 1996. (in Vorbereitung).

[39] BENECKE, C., R. GRUND, R. HOHBERGER, A. KERBER, R. LAUE, AND T. WIELAND. MOLGEN+, a generator of connectivity isomers and stereoisomers for molecular structure elucidation. *Anal. Chim. Act.*, **1995**, 314, 141–147.

[40] KERBER, A., R. LAUE, AND T. WIELAND. Erkennung, Beschreibung und Visualisierung molekularer Strukturen. In *Proceedings des Statusseminars der anwendungsorientierten Verbundprojekte auf dem Gebiet der Mathematik mit Förderung durch das BMBF*. Springer Verlag, Heidelberg, Berlin, 1996. (In Druck).

[41] WILLETT, P. *Similarity and clustering in chemical information systems*. J. Wiley & Sons, New York, NY, 1987.

[42] DUGUNDJI, J. AND I.K. UGI. An algebraic model of contitutional chemistry as a basis for chemical computer programs. *Top. Curr. Chem.*, **1973**, 39, 19–64.

[43] KERBER, A. *Algebraic Combinatorics Via Finite Group Actions*. BI-Wissenschaftsverlag, Mannheim, Wien, Zürich, 1991.

[44] READ, R.C. Every-one a winner. *Ann. Discr. Math.*, **1978**, 2, 107–120.

[45] LAUE, R. Construction of combinatorial objects – a tutorial. *Bayreuther Mathem. Schr.*, **1993**, 43, 53–96.

[46] GRÜNER, T., R. LAUE, AND M. MERINGER. Applications for group actions applied to graph generation. In FINKELSTEIN, L. AND C. KANTOR, editors, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, Providence, RI, 1995.

[47] GRUND, R., A. KERBER, AND R. LAUE. Construction of discrete structures, especially isomers. *Discr. Appl. Math.*, **1996**, 67, 115–126.

[48] GRUND, R., A. KERBER, AND R. LAUE. MOLGEN, ein Computeralgebra-System für die Konstruktion molekularer Graphen. *MATCH*, **1992**, 27, 87–131.

[49] SHELLEY, C.A. Heuristic Approach for Displaying Chemical Structures. *J. Chem. Inf. Comput. Sci.*, **1983**, 23, 61–65.

[50] UGI, I., A. DÖMLING, B. GRUBER, M. HEILINGBRUNNER, C. HEISS, AND W. HÖRL. Formale Unterstützung bei Multikomponentenreaktionen – Automatisierung der Synthesechemie. In MOLL, R., editor, *Software-Entwicklung in der Chemie 9*, 114–128. GDCh, Frankfurt am Main, 1995.

[51] KUBINYI, H. Der Schlüssel zum Schloß. I. Grundlagen der Arzneimittelwirkung. *Pharmazie in unserer Zeit*, **1994**, 23(3), 158–168.

[52] KUBINYI, H. Der Schlüssel zum Schloß. II. Hansch-Analyse, 3D-QSAR und De novo-Design. *Pharmazie in unserer Zeit*, **1994**, 23(5), 281–290.

[53] KUBINYI, H., editor. *3D QSAR in Drug Design. Theory, Methods and Applications*, Leiden, 1993. ESCOM Science Publishers.

[54] CRAMER III., R.D., D.E. PATTERSON, AND A. VITTORIA. *J. Amer. Chem. Soc.*, **1988**, 110, 5959–5967.

[55] CRIPPEN, G.M. *Distance Geometry and Conformational Calculations*. J. Wiley, New York, NY, 1981.

[56] HOFLACK, J. AND P.J. DE CLERCQ. The SCA program: An easy way for the conformational evaluation of polycyclic molecules. *Tetrahedron*, **1988**, 44, 6667–6676.

[57] PEARLMAN, R. Rapid generation of high quality approximate 3D molecular structures. *Chem. Design Autom. News*, **1987**, (2), 5–6.

[58] SADOWSKI, J. AND J. GASTEIGER. From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders. *Chem. Rev.*, **1993**, 93, 2567–2581.

[59] ALLINGER, N.L. MM2. A Hydrocarbon Force Field Utilizing V_1 and V_2 Torsional Terms. *J. Am. Chem. Soc.*, **1977**, 99, 8127–8134.