communications in mathematical
match
and in computer chemistry

no. 37, March 1998

ISSN 0340-6253

MATCDY (37) 157-162 (1998)

THE GENERATION OF MOLECULAR GRAPHS WITH OBLIGATORY, FORBIDDEN, AND DESIRABLE FRAGMENTS

Sergey G. Molodtsov[1]

Institute of Organic Chemistry
Russian Academy of Sciences, Siberian Branch,
Novosibirsk 630090, Russia

1 Introduction

The previous articles [1,2] describe an efficient algorithm and a program for the generation of all connected molecular graphs (their adjacency matrices) with a set of labelled vertices of a given valency and with a set of nonoverlapping fragments. However, the condition nonoverlapping could not always be met. For example, structure elucidation systems may utilize different kinds of spectra in order to obtain data on fragments possibly belonging to the unknown compound. As these data are derived from different sources, fragments sometimes cannot be supposed not to overlap in general. Besides this, it is often known that definite fragments must be absent in the unknown compound. These forbidden fragments can be related either to chemically unstable groups or to structural features found not to belong to a particular unknown. Finally, fragments may be alternative, i. e. at least one fragment from a given set has to be included in the unknown compound. The generators GENOA [3] and COCOA [4][2] seem to account for the noticed constraints. However, there remains the efficiency question. Checking the presence of a fragment only *after* the complete construction of each molecular graph is a waste of computing time. Therefore, all structural constraints should be considered *during* the generation.

The present article describes the procedure of accounting for obligatory, forbidden, and desirable fragments in the generation process of the program GENM. This procedure assures the efficient construction of all graphs (their adjacency matrices) corresponding to given sets of fragments.

[1] Financially supported by Deutsche Forschungsgemeinschaft

[2] This is also true for the new generator MOLGEN 4.0 which is much more efficient and coming out soon

2 Defining the objective

Let us introduce some definitions. All the terms that are not introduced explicitly here are defined in [1,5].

Let $G = (V, E)$ be a *molecular graph* — a multigraph with labelled vertices. A *subgraph* of G is a graph whose vertices and edges are contained in G. An *induced subgraph* of G is a subgraph for which any two vertices are adjacent if and only if they are adjacent in G. A *spanning subgraph* is a subgraph containing all the vertices of G.

Let $G = (V, E_1)$, $H = (V, E_2)$ be graphs and $A = (a_{ij})$, $B = (b_{ij})$ their adjacency matrices. Evidently, a graph H is a spanning subgraph of G if and only if $b_{ij} \leq a_{ij}\ \forall i, j$.

A *fragment* is a molecular graph whose vertices may have a generic label equivalent to any other label. A fragment F and a graph H are *isomorphic* if there exist one-to-one correspondences between their vertices which preserves adjacency and labels.

An *obligatory fragment* of a graph G (denoted $F \subset G$) is a fragment F isomorphic to some subgraph H. In this case one can say that the fragment F is isomorphically embedded in a graph G, and that G contains the fragment F. Note that we can establish the kind of subgraph H. For example: H must be an induced subgraph of G, if two vertices of H belong to different connected components of H then one may be adjacment any way in G, etc.

A *forbidden fragment* of graph G (denoted $\bar{F} \not\subset G$) is a fragment $\bar{F}$ which is not isomorphic to any subgraph of G.

Let a fragment with an assigned positive weight be called a *desirable fragment* and let $\{\tilde{F}_m\}$ be a set of desirable fragments. Denote by $w(\tilde{F}_m)$ the weight of fragment $\tilde{F}_m$. Define the *weight* of the graph by

$$W(G) = \sum_{\tilde{F}_m \subset G} w(\tilde{F}_m)$$

as a sum of weights of the desirable fragments that are contained in the graph G.

Let $\{F_k\}$ be a set of obligatory fragments, $\{\bar{F}_l\}$ — a set of forbidden fragments, $L > 0$ — a number called *a repulse threshold*. Then the generation objective can be formulated as follows: Construct all the connected molecular graphs (their adjacency matrices) with a given set of labelled vertices of a given valency so that

$$\forall\ k:\ F_k \subset G,\ \forall\ l:\ \bar{F}_l \not\subset G,\ \sum_{\tilde{F}_m \subset G} w(\tilde{F}_m) \geq L,$$

i. e. the weight of generated graphs should be greater or equal the repulse threshold.

Note that the desirable fragments can emulate alternative ones. To do it assign to each alternative fragment the weight 1 and set the repulse threshold equal to 1, too.

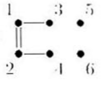
All specified fragments are considered independently from each other, and so obligatory fragments may overlap in the generated graphs.

3 Current checking of isomorphic embedding of a fragment

Let us recall that the algorithm of molecular graph generation in the program GENM is a stepwise procedure. On each step a strongly canonical matrix, being a partially filled matrix, is constructed.

Let A^s be a partially filled matrix constructed at step s. We obtain the matrix $\dot{A}^s$ by filling with zeroes certain places of the A^s which were not yet occupied before. Denote by $\dot{G}^s$ the graph corresponding to $\dot{A}^s$. The example below shows matrices A^2 (dots denote places which are not filled yet), $\dot{A}^2$, and the corresponding graph $\dot{G}^2$:

$$A^2 = \begin{pmatrix} 0 & 2 & 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & \cdot & \cdot & \cdot \\ 0 & 1 & \cdot & 0 & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & 0 & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & 0 \end{pmatrix} \quad \dot{A}^2 = \begin{pmatrix} 0 & 2 & 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \dot{G}^2 =$$



Let $\Gamma(A^s)$ be the set of all the graphs whose adjacency matrix is an admissible complement of the matrix A^s.

3.1 Lemma *The graph $\dot{G}^s$ is a spanning subgraph of each graph $G \in \Gamma(A^s)$.*

This follows immediately from $\dot{a}^s_{ij} \leq a_{ij}$, $\forall\ i, j$, where A is the adjacency matrix of G.

3.2 Corollary *Let F be a fragment such that $F \subset \dot{G}^s$. Then $F \subset G$, for any graph $G \in \Gamma(A^s)$.*

The proof is as follows: if $F \subset \dot{G}^s$ and $\dot{G}^s \subset G$ then $F \subset G$.
And so, if a fragment is contained in graph $\dot{G}^s$, there is no need to repeat the check for isomorphic embedding at further steps within the same branch of the generation tree, and all the graphs generated later at this branch would necessarily contain this fragment.
There are many algorithms for checking isomorph embedding (see [6]). We used an depth-first backtrack procedure, which for every vertex u in a fragment keeps a list of corresponding vertices $\{v\}$ in the graph. On each step the next possible mapping (u, v) preserving vertices labels and connectivity is checked and corresponding lists are rebuilt.

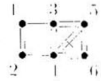
4 Current checking of absence of a fragment

Similarly, one can recognize the absence of a fragment at early stages of the generation. However, even if the fragment F does not belong to $\dot{A}^s$, it could appear there later. To ensure the absence of a fragment one needs another matrix.
Again A^s is a partially filled matrix constructed at step s. Define the matrix $\ddot{A}^s$ as follows:

$$\ddot{a}^s_{ij} = \begin{cases} a^s_{ij}, & \text{in filled places} \\ r'_{ij}, & \text{otherwise} \end{cases}$$

where r'_{ij} is the maximal multiplicity of edges at step s.
By definition the matrix $\ddot{A}^s$ is a complement of A^s. Let $\ddot{G}^s$ be the graph corresponding to $\ddot{A}^s$. The example below shows the same matrix A^2, the adjacency matrix $\ddot{A}^2$ and the corresponding graph $\ddot{G}^2$:

$$A^2 = \begin{pmatrix} 0 & 2 & 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & \cdot & \cdot & \cdot \\ 0 & 1 & \cdot & 0 & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & 0 & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & 0 \end{pmatrix} \quad \ddot{A}^2 = \begin{pmatrix} 0 & 2 & 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 2 & 2 & 2 \\ 0 & 1 & 2 & 0 & 2 & 2 \\ 0 & 0 & 2 & 2 & 0 & 2 \\ 0 & 0 & 2 & 2 & 2 & 0 \end{pmatrix} \quad \ddot{G}^2 =$$



Obviously the following is true:

4.1 Lemma *Any graph $G \in \Gamma(A^s)$ is a spanning subgraph of $\bar{G}^s$.*

4.2 Corollary *Let F be a fragment such that $\bar{F} \not\subset \bar{G}^s$. Then $\bar{F} \not\subset G$ for any graph $G \in \Gamma(A^s)$.*

So, if the graph G^s does not contain the fragment F, then it need not to be checked at further steps of the generation.
The following procedure for accounting of obligatory, forbidden, and desirable fragments is developed on the basis of these two corollaries.

5 Procedure for accounting of obligatory, forbidden, and desirable fragments

Let $\{F_k\}$ be a set of obligatory, forbidden, and desirable fragments, L a repulse threshold, A^s a strongly canonical matrix obtained at step s, p the number of vertices ($s < p$ always). Define parameters $\dot{t}(F_k)$ and $t(F_k)$ as follows:

$$\dot{t}(F_k) = \begin{cases} \min\{t|F_k \subset \dot{G}^t\}, & F_k \subset \dot{G}^{s-1} \\ p, & \text{otherwise} \end{cases}$$

$$\bar{t}(F_k) = \begin{cases} \min\{t|F_k \not\subset \bar{G}^t\}, & F_k \not\subset \bar{G}^{s-1} \\ p, & \text{otherwise} \end{cases}$$

Let $\dot{W}^{s-1}$ be the sum of weights of desirable fragments contained in the graph $\dot{G}^{s-1}$ and W^{s-1} be the sum of weights of desirable fragments contained in the graph $\bar{G}^{s-1}$. Evidently, the weights of the graphs whose adjacency matrices are complements of A^{s-1} will be in the interval from $\dot{W}^{s-1}$ to W^{s-1}.
The procedure for accounting of fragments is inserted into the generation algorithm after the checking of the strong canonicity (point 12 in [1]). The procedure now follows:

```
1. $k = 0$
   $\dot{W}^s = \dot{W}^{s-1}$
   $W^s = W^{s-1}$
2. $k = k + 1$
   if $\dot{t}(F_k) = p$ and $t(F_k) = p$ then
       if $F_k$ is an obligatory fragment then
           if $F_k \not\subset G^s$ then
               goto one step back
           if $F_k \subset \dot{G}^s$ then
               $\dot{t}(F_k) = s$
       else if $F_k$ is a forbidden fragment then
           if $F_k \subset \dot{G}^s$ then
               goto one step back
           if $F_k \not\subset G^s$ then
               $t(F_k) = s$
```

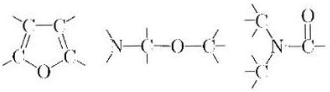
```
      else if  $\dot{W}^s < L$  then      {$F_k$ is a desirable fragment}
         if  $F_k \not\subset G^s$  then
            $W^s = \dot{W}^s - w(F_k)$
            if  $W^s < L$  then
               goto one step back
            $i(F_k) = s$
         else if  $F_k \subset \dot{G}^s$  then
            $\dot{W}^s = \dot{W}^s + w(F_k)$
            $\dot{i}(F_k) = s$
4. if  $k < f$  then                    {$f$ is the number of fragments}
      goto 2
```

As may be seen from the present layout, at every step the minimal and maximal weights of graphs to be generated are calculated. These data allow to control the generation process.

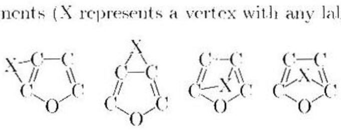
6 Results

The molecular graph generation program GENM was complemented with the above procedure for accounting of obligatory, forbidden, and desirable fragments during the generation process. The procedure was used for the purpose of structure elucidation on infrared spectra [7]. In this work the number of desirable fragments reaches 125. However, let us consider examples which can be easily reproduced.

EXAMPLE 1. We need to generate all graphs with molecular formula $C_7H_8N_2O_3$ with three obligatory fragments,

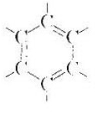


and four forbidden fragments (X represents a vertex with any label):

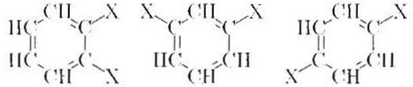


In this example one obligatory fragment can be immediately built in the adjacency matrix. Here the first fragment is chosen as it contains the maximal number of edges. Starting from these data program GENM generates 2778 graphs for 55 sec. (PC 486/66 MHz). Without forbidden fragments the program generates 3365 graphs for 1 min. 23 sec.

EXAMPLE 2. It is required to generate all graphs with molecular formula $C_{10}H_{15}NO$ and with a bisubstituted benzene ring. The last condition is formalized through one obligatory fragment



representing the benzene ring and three alternative fragments



answering for the bisubstitution. There are generated 1 032 graphs in 6 sec. Without alternative fragments 4 468 graphs are generated in 8 sec.
Thus, the present procedure for accounting of fragments during the generation both allows to consider potent structural constraints and reduces the total computing time.

Acknowledgment

The author would like to thank Prof. A. Kerber, Prof. R. Laue for their valuable help and support.

References

1. S. G. Molodtsov, Computer-Aided Generation of Molecular Graphs, *MATCH*, 30 (1994), 213–224.

2. S. G. Molodtsov, Generation of Molecular Graphs with a Given Set of Nonoverlapping Fragments, *MATCH*, 30 (1994), 203–212.

3. R. E. Carhart, D. H. Smith, N. A. B. Gray, J. G. Nourse and C. Djerassi, GENOA: A Computer Program for Structure Elucidation Utilizing Overlapping and Alternative Substructures, *J. Org. Chem.*, 46 (1981), 1708–1718.

4. B. D. Christie and M. E. Munk, Structure Generation by Reduction: A New Strategy for Computer-Assisted Structure Elucidation, *J. Chem. Inf. Comput. Sci.*, 28 (1988), 87–93.

5. F. Harary, Graph Theory, Addison-Wesley, 1969.

6. A. Lingas, Certain Algorithms for Subgraph Isomorphism Problems, *Lect. Notes Comp. Sci.*, 112 (1981), 290–307.

7. V. N. Piottukh-Peletskii, B. G. Derendyaev, S. G. Molodtsov and T. F. Bogdanova, Complete sets of fragment compositions of structures for IR spectrum interpretation using a database. 4. Forming the most probable hypothesis about the structure of the unknown, *J. Struc. Chem.*, 38 (1997), 657–665.