Markus Meringer

Algorithms for Chemical Space Enumeration and Applications in Astrobiology

Small Organic Molecules: Chemical Space, Reactions, Catalysis and Autocatalysis Workshop

Department of Statistics, University of Oxford, UK

March 24, 2020
Outline

- **Data structures and algorithms**
  - representation of chemical structures in a computer
  - algorithms for enumerating chemical structures
- **Applications in astrobiology**
  - construction and
  - analysis of
  virtual libraries of
  - amino acids
  - nucleotide analogs
  - other biomolecules...
Representing chemical compounds: What precisely are we talking about?

Different levels of abstraction

**Composition**
- $C_4H_4Cl_4$
- molecular formula

**Constitution**
- structural formula

**Configuration**
- Conformation

**Specialization**
- Generalization
From compositions to constitutions

Example: Alkanes $C_nH_{2n+2}$

Typically there are several, mostly very many structural formulas with the same molecular formula

Lists must be

- complete
- non-redundant

Exponential growth!

- $C_4H_{10}$
- $C_5H_{12}$
- $C_6H_{14}$
- $C_7H_{16}$ ...

... 9 isomers (try yourself – it’s fun!)
Applications: relating structure and properties

- From structure to physical, chemical, biological and pharmaceutical properties
  - structure-property relationships, esp. QSAR/QSPR
  - application of such relationships to predict properties of virtual structures (→ inverse QSAR)

- From physical and chemical properties (spectra) to structure
  computer-aided / automated molecular structure elucidation
  “CASE”
Structure elucidation by database searching

- Established approach: use spectral data as molecular fingerprint for a database search

- Problem: only such data can be found that is stored in the database
Sizes of data bases

Structures:
- elements C, H, N, O
- at least 1 C-atom
- standard valencies C:4 H:1 N:3 O:2
- no charges
- no radicals
- only connected structures

Need for techniques to explore virtual chemical space in silico!
The DENDRAL project

- driven by exobiologist J. Lederberg
- initiated in the mid 1960’s
- short for DENDRitic ALgorithm
- included an algorithm for generating acyclic structures
- partially funded by NASA
- aim: identifying unknown organic molecules by analyzing their mass spectra (MS) automatically
- perspective: processing of MS recorded on mars missions
- pioneer project in artificial intelligence, first expert system
- structure generators covering cyclic structures followed: StrGen, CONGEN, GENOA

DENDRAL approach to structure generation

- remove hydrogen
- decompose into superatoms
- strip element symbols
- delete free valencies
- replace chains of bivalent nodes by edges

Generating tree for $C_6H_{10}$ isomers

- masterpiece of computer programming
- especially in consideration of limited hardware resources, operation systems, programming languages available at this time
- however, this approach was very complicated
- particularly not suited to process structural constraints efficiently
Molecular graphs

- **Chemical compounds as molecular graphs**
  - vertices and edges (simple graph)
  - + bond multiplicities (multigraph)
  - + element & atomic state symbols

- **Representation of molecular graphs in a computer:**
  - adjacency matrix
  - - label atoms with numbers
  - - write bond multiplicities into a matrix

- **Idea:** fill adjacency matrix in all possible ways
Chemical compounds in nature and in silico

Chemical compounds
- in nature: atoms are not labeled
- in a computer: atoms have to be labeled

leads to problems
- up to $n!$ different labeled (isomorphic) representations of an unlabeled structure
- deciding whether two labeled structures are isomorphic is computationally expensive
- “graph isomorphism problem”
Discrete mathematicians found solutions

Orderly generation
- principle found by Read in 1978
- reduced the number of isomorphism tests

Fast isomorphism tests
- Luks found polynomial time algorithm in 1982
- note: molecular graphs have valences at most 4 (or maybe 6 for S)
Order on edges of labeled graphs

Order on edges of graphs:

\[ e = (x,y), \ e' = (x',y') \] with \( x < y, \ x' < y' \)

then \( e < e' \), iff

\[ x < x' \] or \( (x = x' \text{ and } y < y') \)

Examples:

\[
(1,2) < (2,3) \\
(1,2) < (1,3)
\]
Order on labeled graphs

Lexicographical order on graphs on n nodes

\[ \gamma = \{e_1, \ldots, e_t\} \text{ with } e_1 < \ldots < e_t \]

\[ \gamma' = \{e'_1, \ldots, e'_t\} \text{ with } e'_1 < \ldots < e'_t \]

then \( \gamma < \gamma' \), iff

(there is an \( i \) with \( e_i < e'_i \) and for all \( j < i \): \( e_j = e'_j \)) or

(\( t < t' \) and for all \( j \leq t \): \( e_j = e'_j \))

Examples: graphs on 3 nodes 1, 2, 3

\{ (1,2), (1,3) \} < \{ (1,2), (2,3) \}

\{ (1,2), (1,3) \} < \{ (1,2), (1,3), (2,3) \}
Generation of labeled graphs

Algorithm: Labeled Generation (γ)

1. Output γ
2. For each edge e > max{e' ∈ γ} do in ascending order of e
   Call Labeled Generation (γ ∪ {e})

Example: graphs on 3 nodes starting with the empty graph, Labeled Generation ({})) produces the output

```
1  2  3
1  2
1  2
3
1  2
1  2
3
1  2
1  2
3
```
Example: labeled graphs on 3 nodes

Labeled Generation (γ)
(1) Output γ
(2) For each edge e > max{e′ ∈ γ} do in ascending order of e
   Call Labeled Generation (γ ∪ {e})
From labeled to unlabeled graphs

How to obtain from labeled graphs ... 

... unlabeled graphs?
Canonical orbit representatives

Solution: Select from each orbit (column) the lexicographically minimal representative

Note: Testing minimality is a rather expensive procedure, up to $n!$ permutations have to be checked
Testing minimality

\( \gamma \) is minimal, iff

for each permutation \( \pi \) of the symmetric group \( S_n \):

\( \gamma \leq \pi(\gamma) \)

Example:

\[ \pi_3(\{(1,2),(2,3)\}) = \{(2,1),(1,3)\} \]

\[ = \{(1,2),(1,3)\} \]

\[ < \{(1,2),(2,3)\} \]

\( \Rightarrow \) not minimal

Note: Using algebraic and group-theoretic methods, costs for testing minimality can be reduced considerably

Generation of unlabeled graphs

Algorithm: Labeled Generation ($\gamma$)

(1) Output $\gamma$
(2) For each edge $e > \max\{e^\prime \in \gamma\}$
   do in ascending order of $e$
   Call Labeled Generation ($\gamma \cup \{e\}$)

Algorithm: Unlabeled Generation ($\gamma$)

(1) If $\gamma$ is minimal in its orbit then
   Output $\gamma$
(2) For each edge $e > \max\{e^\prime \in \gamma\}$
   do in ascending order of $e$
   Call Unlabeled Generation ($\gamma \cup \{e\}$)
Example: unlabeled graphs on 3 nodes

Unlabeled Generation ($\gamma$)

(1) If $\gamma$ is minimal in its orbit then
   Output $\gamma$

(2) For each edge $e > \max\{e' \in \gamma\}$
   do in ascending order of $e$
   Call Unlabeled Generation
   ($\gamma \cup \{e\}$)

... many more refinements and enhancements lead to an efficient generator of molecular graphs ...

A new generation of structure generators

- MOLGEN 3.5 (1997, Win 95)
- MOLGEN 4.0 (1998, UNIX)
- MOLGEN 5.0 (2007, Win, Linux)
- others, e.g. Assemble, OMG

Computational example with constraints

<table>
<thead>
<tr>
<th>Restrictions</th>
<th>no. of isomers</th>
<th>CPU-time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemical formula $C_6H_8O_6$ only</td>
<td>2,558,517</td>
<td>838 s</td>
</tr>
<tr>
<td>no triple bonds</td>
<td>2,434,123</td>
<td>703 s</td>
</tr>
<tr>
<td>hydrogen distribution $1CH_2,2CH_1,3C,4OH$</td>
<td>79,831</td>
<td>25 s</td>
</tr>
<tr>
<td>no substructure -O-O-</td>
<td>35,058</td>
<td>97 s</td>
</tr>
<tr>
<td>hybridization $1Csp3-2H,2Csp3-1H,3Csp2-OH,1Osp2-OH$</td>
<td>990</td>
<td>8 s</td>
</tr>
<tr>
<td>minimal size of rings =5</td>
<td>348</td>
<td>5 s</td>
</tr>
<tr>
<td>contains at least one $CO_3$ branch</td>
<td>15</td>
<td>11 s</td>
</tr>
</tbody>
</table>

### Orders of magnitude of structural spaces and data bases

<table>
<thead>
<tr>
<th>Database</th>
<th>Number of Structures</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMR Shift DB</td>
<td>$10^5$</td>
</tr>
<tr>
<td>NIST MS DB</td>
<td>$10^6$</td>
</tr>
<tr>
<td>PubChem</td>
<td>$10^8$</td>
</tr>
<tr>
<td>GDB-13</td>
<td>$10^9$</td>
</tr>
<tr>
<td>GDB-17</td>
<td>$10^{10}$</td>
</tr>
<tr>
<td>MOLGEN unconstitutional isomers of TRP</td>
<td>$10^{12}$</td>
</tr>
<tr>
<td>quartic graphs on 23 points</td>
<td>$10^{14}$</td>
</tr>
<tr>
<td>molecular graphs (C,H,N,O, ≤150Da)</td>
<td>$10^3$</td>
</tr>
<tr>
<td>molecular graphs (C,H,O, ≤180Da)</td>
<td>$10^5$</td>
</tr>
<tr>
<td>constitutional isomers of TRP</td>
<td>$10^7$</td>
</tr>
</tbody>
</table>

- paper and pencil (e.g. small alkanes)
- object lessons (e.g. 217 isomers of C$_6$H$_6$)
- automated structure elucidation via MS
- automated structure elucidation via NMR
- molecular graphs (C,H,N,O, ≤150Da: 3.7e9)
- molecular graphs (C,H,O, ≤180Da: 6.7e10)
- constitutional isomers of TRP (1.9e13)
- quartic graphs on 23 points (4.3e14)
Sizes of data bases and numbers of molecular graphs

Structures:
- elements C, H, N, O
- at least 1 C-atom
- standard valencies C:4 H:1 N:3 O:2
- no charges
- no radicals
- no stereoisomers
- only connected structures

I would like to generate a saturated "chemistry space" (i.e. list of isomers) for all possible alpha amino acids (NH2-CHR-COOH), where R is restricted to smallish side-chains of carbon (C=4), with additional sulphur (S≤1), oxygen (O≤2), nitrogen (N≤3) and hydrogen and a possible benzyl ring.

No. molecular formulas: 132 ...
No. structures: 24749 ... that's what I call a manageable chemical space.
Amino acid libraries resulting from the studies at UHNAI

Beyond Terrestrial Biology: Charting the Chemical Universe of α-Amino Acid Structures

Markus Meringer, H. James Cleaves II, and Stephen J. Freeland

1German Aerospace Center (DLR), Earth Observation Center (EOC), München Straße 20, D-82234 Oberpfaffenhofen–Wessling, Germany
2Earth-Life Science Institute, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan
3Institute for Advanced Study, 1 Einstein Drive, Princeton, New Jersey 08540, United States
4Blue Marble Space Institute of Science, 2800 Woodley Road NW, no. 549, Washington, D.C. 20016, United States
5Center for Chemical Evolution, Georgia Institute of Technology, Atlanta, Georgia 30332, United States
6NASA Astrobiology Institute, University of Hawaii, 2680 Woodlawn Drive, Honolulu, Hawaii 96822-1839, United States

Abstract: α-Amino acids are fundamental to biochemistry as the monomeric building blocks with which cells construct proteins according to genetic instructions. However, the 20 amino acids of the standard genetic code represent a tiny fraction of the number of α-amino acid chemical structures that could plausibly play such a role, both from the perspective of natural processes by which life emerged and evolved, and from the perspective of human-engineered genetically coded proteins. Until now, efforts to describe the structures comprising this broader set, or even estimate their number, have been hampered by the complex combinatorial properties of organic molecules. Here, we use computer software based on graph theory and constructive combinatorics in order to conduct an efficient and exhaustive search of the chemical structures implied by two careful and precise definitions of the α-amino acids relevant to coded biological proteins. Our results include two virtual libraries of α-amino acid structures corresponding to these different approaches, comprising 121,044 and 3,846 structures, respectively, and suggest a simple approach to exploring much larger, as yet uncomputed, libraries of interest.

Jim Cleaves

156-membered badlist

Number of structures

Number of carbon atoms

coded amino acids
total structures UL
plausible structures UL
total structures CL
plausible structures CL
Application:
Verify a model on selection of the amino acid alphabet

- Model established previously on a small set of known amino acids
  - abiotic
  - coded
  - biosynthetic

- The 20 biologically encoded amino acids cover chemical space optimally in terms of
  - range and
  - evenness
  with respect to 3 properties
  - charge,
  - size and
  - hydrophobicity

Property calculation for the generated library

- **Calculation of physico-chemical properties**
  - hydrophobicity represented by logP (MOLGEN-QSPR)
  - size represented by Van der Waals volume $V_{vdw}$ (MOLGEN-QSPR)
  - charge represented by $pK_a$ (JChem)

... gives a 3D mapping of our amino acid chemical space

20 biologically encoded amino acids colored green
Statistical Analysis

- Adaptive analysis gives insight to the adaptive properties of the amino acid alphabet

Method:
- sample $10^8$ random sets of 20 amino acids from a virtual library of 1913
- compute coverage of chemical space in terms of range and evenness in three dimensions ($\log P$, $V_{vdw}$, $pK_a$)

Results:
- better sets do exist,
- but they are rare,
- and energetically less favorable

6 sets with better coverage

black: meteoritic  red: encoded  blue: both
Simple statistics by basic combinatorics

- 5 of the 6 better sets (~83%) include at least one encoded AA
- the probability that a random set of 20 includes at least one encoded amino acid is only 19%

Latest results: even the subsets of the genetically encoded amino acids show adaptive properties!

Nucleotides

- **Monomeric building blocks of**
  - DNA
  - RNA
- **Structure**
  - linker: phosphate group
  - core: sugar (ribose)
  - base: C, G, A, T or U
- **Idea**
  - generate isomers of ribose
  - and more general analogues of the core structure
  - analyze the resulting nucleoside libraries
“The 227 faces of RNA”

Isomers of ribose

Conclusion: ribonucleosides may have competed with a multitude of alternative structures

Chemical space of general nucleosides

MOLGEN input

- **Formulas**
  - C2-7H5-15O[h=0]0-2O[h=1]2-4Cl -sum O=2-4

- **Rings**
  - ringsize 5-10

- **Bonds**
  - maxbond 2

- **Badlist**
  - BadHetCl: 2 items
  - BadAaNucList: 181 items
  - BadRingList: 13 items
  - BadAromaticsList: 14 items

---

Sizes of libraries

- Number of structures
- Number of C atoms
Analysis includes:

- geometric descriptors
- shape similarity
- synthetic accessibility
- drug-likeness
The rTCA chemical space

Several approaches

- **Database search (Morowitz et al, 2000)**
  - formulas $C_xH_yO_z$, $1 \leq x \leq 6$, $1 \leq y < 99$, $1 \leq z < 99$
  - $x/y \leq 1$, $y/z \leq 2$ for $1 \leq x \leq 3$,
  - $x/y \leq 1$, $y/z \leq 1.5$ for $4 \leq x \leq 6$
  - prescribed $C=O$, forbidden $C-O-C$, $O-O$, no cyclic compounds, no triple bonds
  - retrieved 153 hits in Beilstein, including the 11 members of rTCA

- **Reaction-based structure generation (Zubarev et al, 2015)**
  - 7 reaction types
  - recursively applied until all 11 rTCA compounds were generated (reaction network)
  - delivered a total of 175 structures (actually 221)

Exhaustive enumeration of the rTCA chemical space

Third approach:
- **Formula-based structure generation**
  - Morowitz rules can almost directly be used as input for MOLGEN
  - additional constraints to exclude hydrates and enols
  - generated 876 structures
  - overlap with Morowitz set: 119
  - overlap with Zubarev set: 70
  - overlap with current databases ...

Perspective:
- **Search autocatalytic cycles in generated set(s)**

Meringer, M., Cleaves, H.J. Computational exploration of the chemical structure space of possible reverse tricarboxylic acid cycle constituents. Sci Rep 7, 17540 (2017)
... more chemical space work in progress ...

- libraries of lipids and their potential to form bilayers

- general small molecule space and exploration missions

Can this help to solve the mystery about Earth’s gigantic prebiotic combinatorial chemistry experiment that led to the origin of life?

“This is one of science’s great unsolved problems that is bound to get much more attention in the near future and turn much more computational as everything else in the biosciences.” - Jotun Hein
Acknowledgements

...to the contributors...

...the institutions...

THANKS FOR YOUR ATTENTION!