Generation of Virtual Amino Acid Libraries for Multiple Applications in Astrobiology

Markus Meringer[1] H. James Cleaves[2,3,4,5] Stephen J. Freeland[6,7]

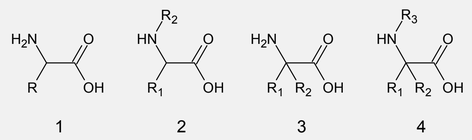
[1]German Aerospace Center (DLR), Earth Observation Center (EOC)
[2]Earth–Life Science Institute (ELSI), Tokyo Institute of Technology [3]Institute for Advanced Study (IAS)
[4]Blue Marble Space Institute of Science (BMSIS) [5]Center for Chemical Evolution, Georgia Institute of Technology
[6]University of Hawaii, NASA Astrobiology Institute (UHNAI) [7]University of Maryland, Baltimore County (UMBC)

Introduction

- Amino acids: “Building blocks of life”
- Genetic code: maps information to function
- Information: stored in DNA, a polymer of nucleotides
- Function: realized by proteins, polymers of amino acids
- Terrestrial life: uses (with very view exceptions) **20** genetically encoded (“coded”) amino acids
- Questions:
  - Why 20? Why these 20?
  - Random result of early evolution on Earth or universal rule?
- Idea: generate comprehensive libraries of virtal amino acids and test hypotheses of [1]

α-Amino Acid Structures

- Coded amino acids have generic structures **1** or **2** (Pro)



- Almost infinite possibilities for side chains **R**, $\mathbf{R_1}$, $\mathbf{R_2}$, $\mathbf{R_3}$
- Numbers of possibilities increase with sizes of side chains [2]

Structure Generator

- Computer program based on methods from graph–theory, combinatorics, group theory and algebra [3]
- Input: molecular formula
- Optional input: structural constraints, e.g. minimum ring size, forbidden and prescribed substructures
- Output: all constitutional isomers that fulffill the constraints
- Software used for this study: MOLGEN 3.5 and MOLGEN 5.0

Isomer Spaces

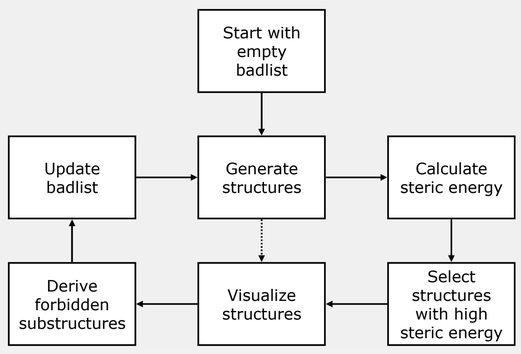
Constitutional isomers of the coded amino acids:

Amino acid	molecular formula	number of isomers		
		all	ringsize $\geq$ **5**	with backbone
Gly	$\mathbf{C_2H_5NO_2}$	84	53	1
Ala	$\mathbf{C_3H_7NO_2}$	391	244	1
Ser	$\mathbf{C_3H_7NO_3}$	1,391	857	2
Cys	$\mathbf{C_3H_7NO_2S}$	3,838	2,422	2
Thr	$\mathbf{C_4H_9NO_3}$	6,836	4,242	4
Asp	$\mathbf{C_4H_7NO_4}$	65,500	25,036	14
Asn	$\mathbf{C_4H_8N_2O_3}$	210,267	81,702	45
Pro	$\mathbf{C_5H_9NO_2}$	22,259	8,462	3 (6)
Val	$\mathbf{C_5H_{11}NO_2}$	6,418	3,973	2
Met	$\mathbf{C_5H_{11}NO_2S}$	86,325	54,575	10
Glu	$\mathbf{C_5H_9NO_4}$	440,821	172,617	71
Gln	$\mathbf{C_5H_{10}N_2O_3}$	1,360,645	539,147	207
Leu, Ile	$\mathbf{C_6H_{13}NO_2}$	23,946	14,866	4
Lys	$\mathbf{C_6H_{14}N_2O_2}$	257,122	162,054	31
His	$\mathbf{C_6H_9N_3O_2}$	89,502,542	13,563,099	902
Arg	$\mathbf{C_6H_{14}N_4O_2}$	88,276,897	36,666,235	3,563
Phe	$\mathbf{C_9H_{11}NO_2}$	277,810,163	25,316,848	571
Tyr	$\mathbf{C_9H_{11}NO_3}$	2,132,674,846	209,838,248	8,309
Trp	$\mathbf{C_{11}H_{12}N_2O_2}$	1,561,538,202,786	64,968,283,073	559,128

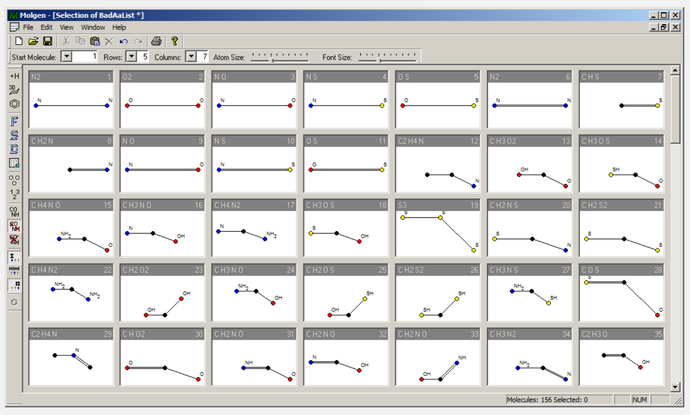
Improvements required:
- Many implausible chemical structures
- All molecular formulas within certain limits should be considered

Plausible Chemical Structures

Create a “badlist” of forbidden substructures:
- Proceed iteratively as depicted in the flow chart
- Use steric energy calculations to find unstable structures

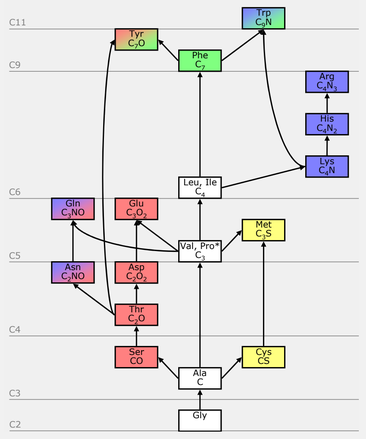


This way a user–defined badlist of 156 substructures was assembled. Some of them are depicted below:



Partial Order on Molecular Formulas

Mathematically, a molecular formula can be considered as a mapping $\boldsymbol{f}$ from a set of chemical elements onto the set of natural numbers, which relates each chemical element $\boldsymbol{X}$ with its multiplicity $\boldsymbol{f(X)}$. For instance $\mathbf{C_2H_5NO_2}$ is represented by the mapping $\boldsymbol{f}$ with $\boldsymbol{f(C) = 2}$, $\boldsymbol{f(H) = 5}$, $\boldsymbol{f(N) = 1}$, $\boldsymbol{f(O) = 2}$ and $\boldsymbol{f(S) = 0}$.
We say $\boldsymbol{f_1}$ is subformula of $\boldsymbol{f_2}$ ($\boldsymbol{f_1 \leq f_2}$), if for all elements $\boldsymbol{X}$ the inequality $\boldsymbol{f_1(X) \leq f_2(X)}$ holds, e.g. $\mathbf{C_2H_5NO_2 \leq C_3H_7NO_3}$.



The figure above represents the **H**–reduced formulas of the coded amino acid's side chains as partially ordered set.

This order can be used to describe the set of molecular formulas defined by a *fuzzy formula*. For instance the fuzzy formula $\mathbf{C_{2-11}H_{5-14}N_{1-4}O_{2-4}S}$ includes all molecular formulas $\boldsymbol{f}$ that fulfill the inclusions $\mathbf{C_2H_5NO_2 \leq f \leq C_{11}H_{14}N_4O_4S}$.

Unique Library

Input for the structure generator:
- Based on a *unique* fuzzy formula $\mathbf{C_{0-5}H_{3-16}N_{0-3}O_{0-2}S_{0-1}R}$ where **R** is a tri–valent macro atom, representing the backbone **1** plus the β-**C** atom
- Two badlists shipped with MOLGEN 5 (cyclic and unsaturated substructures, bridged aromatic substructures)
- Our own customized badlist of 156 substructures
- Allowed ring sizes of 5–10

Results:

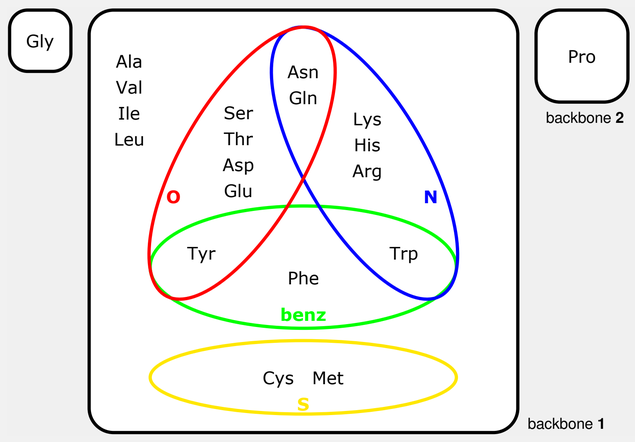
Number of **C** atoms	number of formulas	number of structures		CPU time for plausible structures
		total	plausible	
3	36	5,185	5	0.9 s
4	60	202,682	88	22.7 s
5	84	4,899,064	3,562	3 min 30.1 s
6	108	97,627,979	117,389	19 min 06.9 s
7	132	1,776,370,818	2,868,117	2 h 19 min 43.6 s
8	156	30,987,520,710	58,002,850	45 h 26 min 57.7 s
$\sum$ (3-6 **C**)	288	102,734,910	121,044	23 min 00.6 s
$\sum$ (3-8 **C**)	576	32,866,626,438	60,992,011	48 h 09 min 41.9 s

Problems:
- Gly missing (no **C** in side chain)
- Pro missing (backbone **2**)
- Tyr, Phe, Trp missing (more than 8 **C** atoms in side chain)

Classification of Coded Amino Acids

6 structural properties as classification criteria:
- Side chain: occurrence of **C**, **N**, **O**, **S**, presence of a benzene ring
- Backbone type **1** or **2**

Result: 10 classes as sketched below



Combined Library

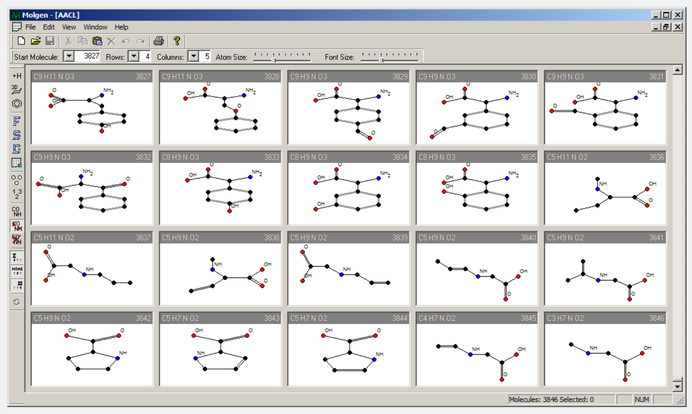
Combined from sublibraries according to the above classification:
- each sublibrary based on its own molecular formula (see table)
- and structural properties defining the class
- plus restrictions already used for the unique library

Results:

Amino acid class	side-chain	back-bone	number of structures	
			total	plausible
Gly			1	1
Phe	$\mathbf{C_{6-7}H_{5-7}}$	1	28	5
Trp	$\mathbf{C_{6-9}H_{6-12}N}$	1	49,296	1,307
Tyr	$\mathbf{C_{6-7}H_{5-7}O}$	1	150	28
Cys, Met	$\mathbf{C_{1-3}H_{3-7}S}$	1	65	28
Asn, Gln	$\mathbf{C_{1-3}H_{3-7}NO}$	1	665	97
Lys, His, Arg	$\mathbf{C_{1-4}H_{3-12}N_{1-3}}$	1	67,597	2,263
Ala, Val, Ile, Leu	$\mathbf{C_{1-4}H_{3-9}}$	1	70	22
Ser, Thr, Asp, Glu	$\mathbf{C_{1-3}H_{3-7}O_{1-2}}$	1	301	84
Pro	$\mathbf{C_{1-3}H_{4-8}}$	2	38	11
$\sum$			118,211	3,846

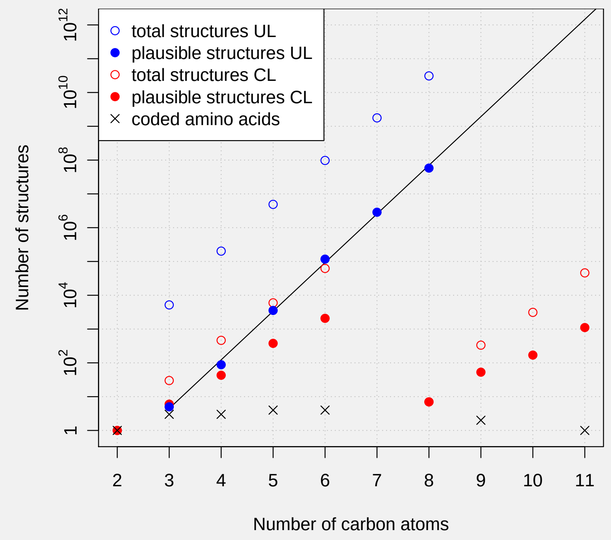
Advantages:
- Moderate size
- All coded amino acids included

Sample Structures

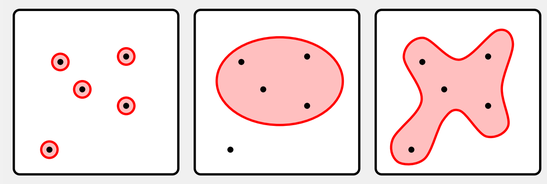


Comparison of the Approaches

Plot of the number of chemical structures generated as a function of the number of **C** atoms for the unique library (UL) and the combined library (CL):



Schematic view of how the different approaches cover the chemical space: isomer space (left), unique library (middle), and combined library (right); black dots represent coded amino acids.



Access

The libraries are freely available for download:
- Unique library (121,044 structures):
  `www.molgen.de/data/AAUL.sdf.zip`
- Combined library (3,846 structures):
  `www.molgen.de/data/AACL.sdf.zip`

The entire work is published in [4].

Applications

- Primary objective: investigations on the selection of the amino acid alphabet [1]; libraries of this study first applied in [5]
- Secondary objectives: use the library compounds as candidate structures for structure elucidation of samples with relevance to astrobiology, e.g. from
  - Prebiotic chemistry experiments
  - Carbonaceous chondrites
  - Future sample return missions
- Applications beyond astrobiology, e.g. in protein engineering

References

[1] G. K. Philip and S. J. Freeland: Did evolution select a nonrandom “alphabet” of amino acids? Astrobiology 11(3): 235-240, 2011.
[2] H. J. Cleaves, 2nd: The origin of the biologically coded amino acids. J. Theor. Biol. 263(4): 490-498, 2010.
[3] M. Meringer: Structure enumeration and sampling. Handbook of Chemoinformatics Algorithms. Edited by J.-L. Faulon and A. Bender, Chapman & Hall: 233-267, 2010.
[4] M. Meringer, H. J. Cleaves and S. J. Freeland. Beyond Terrestrial Biology: Charting the Chemical Universe of α–Amino Acid Structures. J. Chem. Inf. Model. 53(1) 2851-2862, 2013.
[5] M. Ilardo: Natural Selection and the Amino Acid Alphabet. Master's Thesis, University of Hawaii, 2013.

Acknowledgements

The Authors would also like to thank the NASA Astrobiology Institute's Director's Discretionary Fund for seed funding for this project as part of the NASA Astrobiology Institute under Cooperative Agreement No. NNA09DA77A issued through the Office of Space Science.