Materials and Methods

Environmental conditions drive self-organisation of reaction pathways in a prebiotic reaction network

William E. Robinson[1], Elena Daines[1], Peer van Duppen[1], Thijs de Jong[1], and Wilhelm T. S. Huck[1,*]

[1] Institute for Molecules and Materials, Radboud University Nijmegen, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands

[*] Corresponding author. Email: w.huck@science.ru.nl

Materials

D-Threose, L-erythrose, L-erythrulose, D-xylulose, D-ribulose (aqueous solution), D-talose, L-idose (aqueous solution), L-gulose, D-allose, D-altrose and 1,3-dihydroxyacetone were purchased form Carbosynth Ltd, L-(−)-sorbose and D-tagatose, D-psicose were purchased from TCI Europe, $CaCl_2$, NaOH, dihydroxyacetone, paraformaldehyde, glycolaldehyde, glyceraldehyde, ribose, O-ethylhydroxylamine hydrochloride, N,O-bis(trimethylsilyl)trifluoroacetamide, acetonitrile, trifluoroacetic acid and 2,4,-nitrotrophenylhydrazine were purchased from Sigma Aldrich. Pyridine was purchased from Fluorochem. Water was obtained from a Millipore system. Formaldehyde solutions were prepared via sublimation of paraformaldehyde or depolymerisation of paraformaldehyde in water via heating at 60 °C. The concentrations of the resulting formaldehyde solutions were determined by high-performance liquid chromatography (HPLC). All other chemicals were used without further purification.

Instrumentation

HPLC analyses were performed on a Shimadzu Nexera X2 instrument. Conditions: GIST C18 column (2 µm pore size, 75 x 3.0 mm), 40 °C, 0.8 mL min^{--1}, acetonitrile:water (1:1, 0.1% trifluoroacetic acid), 1 µL injection volume, UV-vis detection at 364 nm, or GWS C18 column (5 µm pore size, 250 x 4.6 mm) at 1.0 mL min.$^{-1}$, 40 °C, 1.0 mL min^{--1}, acetonitrile:water (1:1, 0.1% trifluoroacetic acid), 1 µL injection volume, UV-vis detection at 364 nm.

Gas chromatography-mass spectrometric analyses were performed on a JEOL JMS-T100GCv. Gas chromatography conditions: Agilent 7890A gas chromatograph; HP-5MS column, 30 m length, 0.250 mm diameter, 0.25 mm film thickness, He carrier gas, 1 mL min.$^{--1}$, 1 µL split injection (1/10), injector temperature 250 °C. Temperature program: Oven Temp/ °C: 100, 170, 210, 250, 325, Rate/ °C min.$^{-1}$: 0, 14, 4, 15, 60, Time/ min. : 2.33, 0, 0, 0, 3.75. Mass spectrometer conditions: JEOL AccuTOF mass spectrometer, Electron Impact Ionisation mode, Ionisation Voltage 2300 V, sampling rate 10 Hz.

Experimental Methods

Flow reactions: A continuous stirred-tank reactor (CSTR, volume: 411 µL or 439 µL) with five inlets and an outlet was fabricated from polydimethylsiloxane as previously reported.[1] Cetoni Nemesys syringe pumps with Hamilton syringes were used to control input flow rates.

Derivatisation: Derivatisation for HPLC was performed in a similarly to a previously reported method.[2] Samples from the CSTR outlet (35 µL) were dropped directly into a solution consisting of DNPH saturated acetonitrile (300 µL), acetonitrile (97.5 µL), water (65 µL) and HCl solution (2 M, 2.5 µL). The solutions were incubated for at least 30 minutes before HPLC analysis. Derivatisation for GCMS analysis was performed according to reported procedures.[2–4] Samples from the flow reactor outlet (35 μL) were flash-frozen in liquid nitrogen and freeze-dried overnight to give dry to oily residues. To each sample was added a solution of O-ethylhydroxylamine hydrochloride in pyridine (75 μL, 20 g L^{-1}). A solution of dodecane and tetradecane (100 μL, 1.6 mM each in pyridine) was then added to each sample. The samples were then shaken at 70 °C for 30 minutes. After cooling to room temperature, N,O-bis(trimethylsilyl)trifluoroacetamide (25 μL) was added to each sample. The samples were again shaken at 70 °C for 30 minutes. The samples were then cooled to room temperature, followed by centrifugation (3-5 min, 10,000 rpm). The supernatants were decanted into sample vials for analysis by GC-MS (see instrumentation section).

Chromatographic data processing: Peak integration and assignment of raw chromatographic data was performed using a program written in the programming language Python with the packages NumPy[5] and Scipy.[6] Peaks were detected using the first derivative of chromatograms and their integrals were determined using the NumPy function trapz() with subtraction of a baseline linearly interpolated between the beginning and the end of the peak. Peak assignments were performed via comparison to reference samples (Supplementary Figures S3-S5), or via interpretation of peak mass spectra (calibrated samples match known fragmentation patterns Supplementary Figures S6-S8)[7] and retention times, aided by inference from experimental data. Integrals were converted to concentrations using quadratic calibration lines (Supplementary Figures S9-S11). When authentic samples were not available, calibrations were estimated by averaging the calibration factors of compounds of similar carbon chain length to the uncalibrated compound (in cases where two peaks were observed for a compound, the calibration for the peak with the larger integral was used).

Data Analysis Methods

Python programs for the following described data analysis are available at https://github.com/huckgroup/formose-2021.git.

Hierarchical clustering of data: The average compositions and amplitudes for each experiment set were combined into an array. The pairwise dis-similarity between each data set was then determined using a correlation-based metric (Eq 1., scipy.spatial.distance.pdist() using the 'correlation' metric)

Eq 1. $1 - PCC \ = 1 - \frac{(u_i - \overline{u})(v_i - \overline{v})}{\sqrt{(u_i - \overline{u})^2} . \sqrt{(v_i - \overline{v})^2}}$

Where *u* and *v* are both one-dimensional vectors (arrays) of average compound concentrations and amplitudes determined for a given experiment.

Generation of the formose reaction space *in silico*: A set of reaction pathways in line with expected the expected reaction types of the formose reaction was generated using The RDKit (RDKit: Open-source cheminformatics; http://www.rdkit.org, date of access: June 2021). The reactions outlined in Scheme S1 were translated into reaction SMARTS (Extended Data **5**) which were iteratively applied to to a seed set of compounds (glycolaldehyde, formaldehyde, hydroxide and water). Products of each reaction operation were fed into the next iteration. Compounds with a chain length of greater than 6 carbon atoms and the reactions leading to them were removed after every iteration. The resulting network corresponds to a hypothetical case of the formose reaction in which all pathways possible according to the contructing reaction rules are taken. This set of reactions was used as a framework for determining reaction pathways from data.

Pathway analysis: The generated list of reactions for the formose reaction was converted into a networkx DiGraph object.[8] Nodes corresponding to compounds were connected by directed edges to nodes for reactions. The edge direction indicated the role of the compound as either a reactant or a product in the reaction to which it is connected. Nodes corresponding to formaldehyde, hydroxide and water were removed from the graph. Data corresponding to compounds' responses to input modulations (Extended Data **4**) were used as a guide in searching for reaction pathways as described in the main text.

To obtain lists of reactions for each data set of compound concentration amplitudes, the following process was applied. The list of detected compounds was sorted in order of decreasing amplitude. Additional compounds, such as enolates, which could not be detected by the methods used, were appended to the bottom of the list.

From the set of generated formose reactions, those for which reactants were not present in the list of compounds were removed. Reactions whose products were inputs to the system (e.g. dihydroxyacetone or formaldehyde) were also removed.

The construction of a reaction pathway began by determining single shortest paths between each carbon input into the reaction (other than formaldehyde) to a compound from the set of reaction products. Shortest paths were then found between consecutive members of the ampltiude-ordered compound list. The pathways were determined in the direction of high to low amplitude. The resulting list of reactions was checked to make sure all product compounds had reactions leading to them. For each compound without an inbound reaction pathway, a connection to the rest of the reaction scheme was found by finding the shortest path to the compound from a set of those with higher amplitudes.

References

1. Semenov, S. N. *et al.* Rational design of functional and tunable oscillating enzymatic networks. *Nat. Chem.* **7**, 160–165 (2015).

2. Haas, M., Lamour, S. & Trapp, O. Development of an advanced derivatization protocol for the unambiguous identification of monosaccharides in complex mixtures by gas and liquid chromatography. *J. Chromatogr. A* **1568**, 160–167 (2018).

3. Becker, M., Liebner, F., Rosenau, T. & Potthast, A. Ethoximation-silylation approach for mono- and disaccharide analysis and characterization of their identification parameters by GC/MS. *Talanta* **115**, 642–651 (2013).

4. Becker, M. *et al.* Evaluation of different derivatisation approaches for gas chromatographic–mass spectrometric analysis of carbohydrates in complex matrices of biological and synthetic origin. *J. Chromatogr. A* **1281**, 115–126 (2013).

5. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

6. SciPy 1.0 Contributors *et al.* SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat Methods* **17**, 261–272 (2020).

7. Laine, R. A. & Sweeley, C. C. Analysis of trimethylsilyl O-methyloximes of carbohydrates by combined gas-liquid chromatography-mass spectrometry. *Analytical Biochemistry* **43**, 533–538 (1971).

8. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. *7th Python Sci. Conf. (SciPy 2008)* 11–15 (2008).