Molecular structural dataset of lignin macromolecule elucidating experimental structural compositions
Sudha cheranma devi Eswaran1,2, Senthil Subramaniam3, Udishnu Sanyal1,2, Robert Rallo3, Xiao Zhang1,2,3

Affiliations
1. Bioproducts Sciences and Engineering Laboratory, Washington State University, 2710 Crimson Way, Richland WA 99354 (USA)
2. Voiland School of Chemical Engineering and Bioengineering, Washington State University (USA)
3. Pacific Northwest National Laboratory, 902 Battelle Blvd, Richland, WA 99354 (USA)

Supporting Information

Table of Contents
Method details 2
Lignin structure generation tool architecture and implementation 2
Data Records
Summary of the dataset 4
Structure Visualization in 2D and 3D representations 4
TMAP visualization of SG type structures 6
Data format specifications: 6
References 9

Tables
Table A: Summary of the LGS Dataset ................................................................. 4
Table B: Example of SMILES representation for structure with degree of polymerization as 3 ............. 4

Figures
Figure A: Software architecture of the Lignin structure generator tool................................. 2
Figure B: CDK integration for molecular structure generation........................................... 3
Figure C: Lignin model. These models, derived from 20 monomeric units, represents MWL structures of (a) softwood (predominantly G units) and (b) hardwood (ratio of S and G unit as 1.8) using LGS tool........ 5
Figure D: TMAP visualization for SG Type Structures. Color is based on free phenolic-OH group in the molecule. Clicking on individual datapoint (circle or leaf) in the tree view displays compound information detailing the structural features and link to 3D view. .......................................................... 6
Figure E: JSON File definition for G type structures with DP as 4 ........................................ 7
Figure F: Matrices (*.csv) file definition and corresponding 2D molecular structure representation for G type structure with DP of 4 ............................................................................................................. 8
Figure G: Example MDL MOL file V3000 generated for G type structure with DP as 3 ............. 9
Method details

Lignin structure generation tool architecture and implementation

Lignin Structure (LGS) generator tool was developed as a standalone utility for computing the network of lignin molecular structure and defining a large set of lignin molecules. Tool is implemented using Core Java, major functionality includes modified version of Heap’s algorithm for finding various permutations of lignin monomer, directed graph creation using combination algorithm and creation of topological matrices, integration of CDK (Chemistry Development Toolkit) for molecular structure generation and creation of SMILES notation for structures generated, storing molecular information as MDL Mol files, evaluating the structural features and storing as JSON file. Software architecture of the lignin structure generator tool is provided in Figure A. The tool can be used to generate different structural variations for a given set of experimental observations by configuring the required parameters such as monomer ratio (S, G and H), bond frequencies (β-O-4, β–β, β-5, 4-O-5, 5-5 and DBDO) in project configuration file (project-config.yaml) file. The tool can generate chemically correct and legible 2D structure diagrams of natural lignin for all wood types that includes hardwood, softwood and herbaceous.

Molecular structure formation using CDK

CDK is a widely used open-source cheminformatics toolkit[1]. The CDK provides data structures to represent chemical concepts along with methods to manipulate such structures and perform computations on them. The monomer descriptors were initialized as IAtomContainer object from CDK and linkages between monomer units are generated using IBond object with the edge definition from the directed graphs created. Class MonolignolBase is defined an abstract class for initializing the template of monomer object containing Phenyl propane unit. Class relation diagram Figure B shows the CDK integration for molecular structure creation.
Implementation details

Lignin Structure (LGS) generator tool was implemented in JDK 8 using CDK 2.3, Maven 4.0.0 and MongoDB. Data analysis and charts were generated using DataWarrior v05.02.01, MongoDB and Python libraries (SciPy 1.6.2,3, Matplotlib 3.3.4, tMapView and RDKit 2020.09.1.0). Detailed information on the installation and execution of the Lignin structure generator tool and the source code availability can be found from the GitHub repository (https://github.com/sudhacheran/lignin-structure-generator).
Data Records

Summary of the dataset

Table A presents the high-level summary of the LGS dataset. The dataset contains longer polymer structures. Number of oligomers represents the range of fragmented structures present in the dataset with respect to number of monomers.

Table A: Summary of the LGS Dataset

| Number of monomers | G Type structures | SG Type Structures |
|--------------------|-------------------|--------------------|
|                    | Number of structural variations | Molecular weight range (g/mol) | Number of oligomers | Number of structural variations | Molecular weight range (g/mol) | Number of oligomers |
| 3                  | 4                 | 504 - 520          | 1                   | 6                          | 560 - 576          | 1                   |
| 4                  | 5                 | 688 - 704          | 1                   | 8                          | 722 - 788          | 1 - 2               |
| 5                  | 10                | 873 - 889          | 1                   | 62                         | 957 - 973          | 1 - 2               |
| 6                  | 18                | 1057 - 1073        | 1                   | 131                        | 1169 - 1185        | 1 - 2               |
| 7                  | 24                | 1241 - 1257        | 1                   | 158                        | 1353 - 1369        | 1 - 2               |
| 8                  | 81                | 1409 - 1425        | 1                   | 219                        | 1565 - 1581        | 1 - 2               |
| 9                  | 123               | 1593 - 1609        | 1                   | 203                        | 1777 - 1793        | 1 - 2               |
| 10                 | 144               | 1777 - 1793        | 1                   | 929                        | 1913 - 1961        | 1 - 3               |
| 11                 | 411               | 1913 - 2097        | 1 - 2               | 870                        | 2125 - 2173        | 1 - 3               |
| 12                 | 220               | 2113 - 2297        | 1 - 2               | 1081                       | 2337 - 2385        | 1 - 3               |
| 13                 | 529               | 2281 - 2465        | 1 - 2               | 1109                       | 2521 - 2569        | 1 - 3               |
| 14                 | 356               | 2465 - 2650        | 1 - 2               | 1316                       | 2734 - 2782        | 1 - 3               |
| 15                 | 257               | 2634 - 2818        | 1 - 2               | 1323                       | 2946 - 2978        | 1 - 4               |
| 16                 | 296               | 2834 - 3018        | 1 - 2               | 1561                       | 3130 - 3162        | 1 - 4               |
| 17                 | 296               | 3002 - 3186        | 1 - 2               | 2456                       | 3310 - 3358        | 1 - 4               |
| 18                 | 311               | 3186 - 3370        | 1 - 2               | 5300                       | 3538 - 3722        | 1 - 4               |
| 19                 | 315               | 3370 - 3554        | 1 - 2               | 4720                       | 3706 - 3738        | 1 - 4               |
| 20                 | 263               | 3538 - 3722        | 1 - 2               | 5294                       | 3902 - 4086        | 1 - 5               |
| 21                 | 311               | 3722 - 3906        | 1 - 2               | 4639                       | 4102 - 4134        | 1 - 4               |
| 22                 | 315               | 3906 - 4090        | 1 - 2               | 5242                       | 4314 - 4346        | 1 - 5               |
| 23                 | 618               | 4074 - 4259        | 1 - 2               | 4830                       | 4511 - 4559        | 1 - 4               |
| 24                 | 685               | 4259 - 4443        | 1 - 2               | 4061                       | 4679 - 4711        | 1 - 5               |
| 25                 | 681               | 4443 - 4627        | 1 - 2               | 8251                       | 4891 - 4923        | 1 - 5               |

Structure Visualization in 2D and 3D representations

Lignin structures included in the LGS dataset can be visualized as 2D and 3D representations. Figure C shows 2D and 3D visualizations of G and SG type structures using CDK and Avogadro, respectively.
Figure C: Lignin model. These models, derived from 20 monomeric units, represents MWL structures of (a) softwood (predominantly G units) and (b) hardwood (ratio of S and G unit as 1.8) using LGS tool.
TMAP visualization of SG type structures
Figure D shows the SG type structural dataset using Tree MAP algorithm.

Data format specifications:
SMILES notation
The Simplified Molecular-Input Line-Entry System (SMILES)\[2, 3\] is a line notation for describing chemical structures using short ASCII strings. Lignin structures are stored as SMILES string representations of the generated molecules, as it is a key asset in cheminformatics and is becoming increasingly relevant to the general chemical community, due to the steadily growing impact of Big Data and Machine Learning. Example of the SMILES notation is shown in Table B.
Table B: Example of SMILES representation for structure with degree of polymerization as 3

| Lignin Oligomer | Molecular structure | SMILES notation |
|-----------------|--------------------|-----------------|
| S1 – βO4 – S2   | ![Molecular Structure](image1.png) | OC=1C(CC1)C=CC3C(OCC)C4C=CC(OC)=CC=CC(OC)(O)C5C=CC(OC)=CC=C5C=OC=C4 |
| S2 – ββ – G3    | ![Molecular Structure](image2.png) | |

JSON format

We used JSON (JavaScript Object Notation) to integrate structural features of all possible permutations for a given degree of polymerization. JSON is widely used data-interchange format. A common data format with diverse uses and stores data as key/value pairs. JSON data definition used in this study is presented in Figure E: JSON File definition for G type structures with DP as 4. The JSON data file can be directly imported into database such as MongoDB for easy data analysis. This file provides the catalog of structural information with specific DP. Lignin id (lig_id) in the JSON object is the unique id to locate the properties of specific structure in MOL and CSV file respectively.

![JSON Object Diagram](image3.png)

Figure E: JSON File definition for G type structures with DP as 4

Definition of generated structures
- `monotype` - represents 'g' or 'qg'
- `name` - polymer - constant for this paper
- `dp` - requested degree of polymerization
- `lig_id` - unique identifier for the structure generated. Name of (*.csv) and (*.mol) files.
- `smilingstring` - SMILES representing molecular structure
- `f uncgrp_count` - Document type representing functional group counts that includes
  - Free phenolic hydroxyl group
  - Methoxy group
  - Primary alcoholic hydroxyl group
  - Secondary alcoholic hydroxyl group (benzyl-OH)
  - Cinnamyl alcohol and group
- `bnbCount` - Document type number of bonds present in the generated structure
- Other structural features such as
  - issingleChain - true (single oligomer) or false (more than one oligomers)
  - branchingfactor - average outdegree in the structure
  - numOligomers - Number of fragmented structures present
  - evaluatedDP - number of monomers in the generated structure

Summary and averaged values from all generated structures
Matrices:
Machine readable form of molecular structure representing linkages between monomers given as adjacency and connectivity matrices. It represents the presence of linkage between the monomers (adjacency) and type of linkage between monomers (connectivity) respectively. Example of a CSV file and its fields definition is explained in Figure F.

Figure F: Matrices (*.csv) file definition and corresponding 2D molecular structure representation for G type structure with DP of 4

Molecular data file:
MOL files are generally classified as data files that contain molecular data information, atom, bonds, coordinates, and connectivity information in plain text format. It was developed and published by Molecular Design Limited (MDL). MDL molfile version V3000 is generated for lignin structures using CDK (Chemistry Development Toolkit). V3000[4] is used for representing proteins and polymers structures. The Avogadro software can be used in Microsoft Windows based systems, Linux and Mac OS to access and view MOL files. Example of a MOL file is provided in Figure G.
Figure G: Example MDL MOL file V3000 generated for G type structure with DP as 3

References

1. Willighagen, E.L., et al., The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. Journal of Cheminformatics, 2017. 9.

2. Weininger, D., Smiles, a Chemical Language and Information-System .1. Introduction to Methodology and Encoding Rules. Journal of Chemical Information and Computer Sciences, 1988. 28(1): p. 31-36.

3. Weininger, D., A. Weininger, and J.L. Weininger, Smiles .2. Algorithm for Generation of Unique Smiles Notation. Journal of Chemical Information and Computer Sciences, 1989. 29(2): p. 97-101.

4. Chemical Table files, in Chapter 10: The Extended Connection Table (V3000). 2005, Elsevier MDL.