Data Article

Data on generation of Kekulé structures for graphenes, graphynes, nanotubes and fullerenes and their aza-analogs

Sergey Trepalin a,*, Sasha Gurke b, Mikhail Akhukov c, Andrey Knizhnik c, Boris Potapkin c

a Institute of Physiologically Active Compounds, Russian Academy of Sciences, Chernogolovka, Moscow Region 142432, Russia
b Elsevier Inc., 230 Park Ave., 8th Floor, New York, NY 10169, USA
c Kintech Lab, 12 3rd Khoroshevskaya Str., Moscow 123298, Russia

ARTICLE INFO

Article history:
Received 12 October 2018
Received in revised form 25 October 2018
Accepted 25 October 2018
Available online 1 November 2018

ABSTRACT

Two new features are added to existing algorithms for kekulization of chemical structures, i.e., handling of triple and cumulene bonds in cycles and use of random atom sorting to remove unmatched atoms. Handling of triple and cumulene bonds enables kekulization of graphynes and graphdiynes. Random sorting speeds up the calculation time, i.e., kekulization of large chemical structures containing about $10^7$ atoms takes $\leq 1$ min on a typical PC. Source codes (Pascal, GNU GPL license) are included as a compiled application (Windows 64). Calculation times and unmatched atom statistics are provided for graphenes, graphynes, nanotubes, graphyne nanotubes and fullerenes. Benchmark comparisons are made for some data.

© 2018 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
Specifications table

| Subject area       | Chemistry (General)                                      |
|--------------------|----------------------------------------------------------|
| More specific subject area | Kekulization of large chemical structures              |
| Type of data       | Tables, text file, figures, computer-readable chemical structures |
| How data was acquired | Calculation using Notebook MSI GE 60 2PG Apache     |
| Data format        | *.mol, *.xyz and *.cc1 files for chemical structures   |
| Experimental factors | 2.8MHz i7 processor, 16G RAM, Windows 2012 server   |
| Experimental features | Single-thread application; Win API method GetTickCount() used for precise time measurements |
| Data source location | Article text for source code and calculation statistics, Attachment1.zip for compiled application, Attachment2.zip for chemical structures |
| Data accessibility | Downloadable .zip                                      |
| Related research article | Trepalin S., Gurke S., Akhukov M., Knizhnik A., Potapkin B., A fast approximate algorithm for determining bond orders in large polycyclic structures, Journal of Molecular Graphics and Modelling, vol. 86 (2019), pp. 52-65 |

Value of the data

- This data proves the viability of a fast algorithm for the modeling of state-of-the-art materials such as graphenes in scientific and commercial applications
- The data is provided for very large chemical structures (10^7-10^8 atoms) and is obtained using ordinary hardware with short calculation times
- This data has been benchmarked against existing algorithms and can be used for benchmarking for the future improvements to the algorithm

1. Data

Detailed description of an algorithm for fast generation of Kekulé structures from a list of atomic valences and connectivity matrices is given in [1]. The source code that was based on this algorithm and used for calculations is provided in this publication.

Model calculations were performed for graphenes, nanotubes, fullerenes and their aza-analogs, polycyclopentadienes and porphine, using a Windows 2012 server with 2.8 GHz i7 processor and 16 GB RAM. The procedures were not run in parallel but rather in a single thread. The Win API method GetTickCount() was used for precise time measurements. This method allows the measurement of the elapsed time with a millisecond precision. For the compounds containing less than 1,000,000 atoms, the computing time was determined as an average of 1000 calculations, less than 10,000,000 atoms – as an average of 100 calculations and for more than 10,000,000 atoms – as the time of one calculation.

For repeated calculations, the sorting order of the sequence of chemical bonds was random. For an algorithm of aromatic bond alternation, this is equivalent to the random selection of a double bond in the node. For a large number of calculations, this procedure allows estimation of the statistical distribution of unmatched atoms after alternation of aromatic bonds.

Compiled application (Windows 64) used for calculations is in the Attachment1.zip. Computer-readable chemical structures are in the Attachment2.zip.
2. Experimental design, materials and methods

2.1. Nanotubes

Carbon nanotubes of various lengths were generated using the pattern shown in Fig. 1. Generated nanotubes contain two acyclic methylene groups at the beginning and at the end of a nanotube.

In Fig. 1, the arrows show the points of attachment for generation of a polymer molecule. The points of attachment for the outer rings connect with the points of attachment for the inner rings. The order of an aromatic bond resulting from connecting a pair of attachment points is unknown. For the end moieties of the polymer, the points of attachment were replaced with single bonds to hydrogen atoms. The results of calculation are given in Table 1 and discussed below.

2.2. Graphene

The pattern for generation of graphene is shown in Fig. 2.

To generate polymers, the points of attachment on the left were connected with the points of attachment on the right of another graphene block in the way similar to that used for nanotubes. The points of attachment of the first and the last blocks were capped with hydrogen atoms via single bonds. For acyclic carbon atoms, required for the generation of repeating aromatic cycles, two hydrogen atoms were added to the point of attachment, resulting in a methylene group. The results of calculations are given in Table 1.

Kekulization of various compounds, including nanotubes and graphenes, is described [2,3] and computing times are provided. For comparison, we performed calculations for some of the same compounds. The results are in Table 1. Computing times were dramatically shorter than those reported in [2].

Detailed discussion of this data can be found in [1].
Table 1
Time required for generation of Kekulé structures and unmatched atom statistics for nanotubes and graphenes.

| Chain length | Formula | Time (ms) | Unmatched atoms statistics |
|--------------|---------|-----------|----------------------------|
|               |         |           | 2  | 4  | 6  | 8  | 10 | 12 | 14 | 16 | 18 | 20 | 22 | 24 | 26 | 28 | 30 | 32 | 34 | 36 | 38 | 40 | 42 |
| Nanotubes     |         |           |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 1             | C_{144}H_{48} | 0.33     | 32 | 351 | 510 | 107 |
| 10            | C_{1440}H_{48} | 2.42     | 1  | 28  | 93  | 111 | 146 | 232 | 215 | 113 | 9  | 3  |
| 100           | C_{14400}H_{48} | 28.1     | 5  | 12  | 15  | 36  | 119 | 270 |
| 1000          | C_{144000}H_{48} | 407.7    | 1  | 1   | 1   | 12  | 91  | 281 | 332 | 227 | 51 | 3  |
| 10000         | C_{1440000}H_{48} | 8123.3   | 6  | 30  | 34  | 23  | 7   |
| 100000        | C_{14400000}H_{48} | 108437   | 1  |
| Graphenes     |         |           |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| 1             | C_{144}H_{38} | 0.16     | 31 | 340 | 442 | 182 | 5  |
| 10            | C_{1440}H_{146} | 2.09     | 9  | 42  | 97  | 140 | 166 | 199 | 167 | 107 | 58 | 15 |
| 100           | C_{14400}H_{1226} | 28.5     | 6  | 14  | 12  | 20  | 24  | 77  | 165 | 255 | 248 | 138 | 37 | 4  |
| 1000          | C_{144000}H_{12026} | 451.8    | 2  | 2   | 2   | 1   | 5   | 7   | 36  | 168 | 421 | 288 | 65 | 3  |
| 10000         | C_{1440000}H_{120026} | 8497.8   | 1  |
| 100000        | C_{14400000}H_{1200026} | 118297   | 1  |
|               |         |           |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
| Data for some compounds from [2] obtained with our algorithm |
| tube-980      | C_{980}H_{26} | 1.36     | 17 | 75  | 164 | 312 | 246 | 143 | 39  | 4  |
| sheet-1800    | C_{1800}H_{178} | 2.19     | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0  |
| sheet-19602   | C_{19602}H_{592} | 28.9     | 0  | 0   | 0   | 0   | 0   | 0   | 0   | 0  |

Where:
Chain length – number of monomeric units in a polymer,
Formula – molecular formula,
Time (ms) – time required for kekulization of a single chemical structure in milliseconds,
Unmatched atoms statistic - number of calculations for a given number of unmatched atoms (2-42).
2.3. Fullerenes and porphine

Alternation of aromatic bonds was studied for fullerenes C20, C60, C70, C80, C82, their random aza-analogs, as well as for porphine (Fig. 3).

Polymeric analogs of these compounds do not exist. Consequently, all the calculations were performed for monomers. To validate the efficiency of the algorithm for five-member cycles, model calculations were performed for azafullerenes by randomly replacing 2, 4 or 8 carbon atoms with nitrogen. Nitrogen has a valence of 3 and three single converging bonds in each node. This substitution can be done for the even number of atoms only. Otherwise, bonds cannot be alternated, and the number of unmatched atoms is odd. The results of calculations are given in Table 2.

Detailed discussion of this data can be found in [1].

In addition, model calculations were performed for 2488 fullerenes from a library by Yoshida [4]. This data is provided in Table S3.

Fig. 2. The pattern for generation of graphenes.
The legend for the column headers in this table is the same as for Table 2, except the column No. non-existent is not provided because every compound from the set of 1000 had a Kekulé structure.

2.4. Graphynes and graphyne nanotubes

We studied graphynes GY1 and GY7 (Fig. 4) and graphyne nanotubes of various degrees of polymerization. Graphyne nanotubes were generated by replacing hydrogen atoms in GY7 with carbon atoms and adding a bond between these atoms in a vertical position.

The times required for kekulization of graphynes and graphyne nanotubes of various degrees of polymerization are shown in Table 3.

Fig. 3. Fullerenes and porphine used for model calculations.
Table 2

Time required for generation of Kekulé structures and unmatched atoms statistics before their removal for fullerenes, their aza-analogs and porphine. The data is for 1000 transactions.

| Compound | Formula | Time (ms) | No. non-existent | Avg. no. iterations | Max. no. iterations | Unmatched atoms statistic |
|----------|---------|-----------|------------------|---------------------|---------------------|--------------------------|
|          |         |           |                  |                     |                     |                          |
| C20      | C20     | 0.078     | 0                | 1.00                | 1                   | 699 301                  |
| Diaza-C20| C18N2   | 0.062     | 0                | 1.03                | 3                   | 409 583 8                |
| Tetraaza-C20| C16N4 | 0.438     | 33 + 59          | 14.2                | 2                   | 499 493 8                |
| Octaaza-C20| C12N8 | 0.953     | 541 + 218        | 70.6                | 1                   | 769 222 9                |
| C60      | C60     | 0.156     | 0                | 1.00                | 1                   | 329 600 68 3             |
| Diaza-C60| C58N2   | 0.171     | 0                | 1.01                | 2                   | 124 681 182 11 1        |
| Tetraaza-C60| C56N4 | 0.234     | 4 + 6            | 3.72                | 3                   | 128 556 337 28 1        |
| Octaaza-C60| C52N8 | 1.250     | 63 + 83          | 24.4                | 3                   | 146 492 306 54 1 1      |
| C70      | C70     | 0.187     | 0                | 1.003               | 2                   | 264 656 76 4             |
| Diaza-C70| C68N2   | 0.188     | 0                | 1.01                | 3                   | 102 269 24               |
| Tetraaza-C70| C66N4 | 0.407     | 2 + 3            | 2.54                | 3                   | 61 498 382 57 2         |
| Octaaza-C70| C62N8 | 0.890     | 49 + 52          | 14.86               | 3                   | 79 419 390 103 9        |
| C80      | C80     | 0.187     | 0                | 1.093               | 3                   | 291 615 94               |
| Diaza-C80| C78N2   | 0.203     | 0                | 1.065               | 3                   | 90 577 312 21           |
| Tetraaza-C80| C76N4 | 0.344     | 1 + 5            | 1.69                | 4                   | 50 421 440 84 5         |
| Octaaza-C80| C72N8 | 0.860     | 31 + 42          | 12.8                | 4                   | 55 303 461 162 18 1     |
| C82      | C82     | 0.203     | 0                | 1.014               | 2                   | 223 622 149 5 1         |
| Diaza-C82| C80N2   | 0.203     | 0                | 1.020               | 2                   | 74 547 347 30 2         |
| Tetraaza-C82| C78N4 | 0.328     | 2 + 4            | 2.84                | 2                   | 42 450 425 78 5         |
| Octaaza-C82| C74N8 | 0.984     | 28 + 37          | 13.97               | 3                   | 46 288 468 175 21 2     |
| Porphine | C20H14N4| 0.093     | 0                | 1.00                | 1                   | 441 559                  |

Where:
- Compound – chemical structure in Fig. 2,
- Formula – molecular formula,
- Time (ms) – time required for kekulization of a single chemical structure in milliseconds,
- No. non-existent – number of compounds from a 1000 set for which Kekulé structure was not generated,
- Avg. no. iterations – average number of shuffles. This number includes the maximum number of shuffles (300) after which a decision is made that a Kekulé structure does not exist,
- Max. no. iterations – maximum number of shuffles required for successful kekulization,
- Unmatched atom statistic - number of calculations for a given number of unmatched atoms (0–> 8).
Polycyclopentadienes (Fig. 5) are remarkable because they contain odd-sized cycles and can be easily generated as long-chain polymers.

![Diagram of monomer units of graphynes GY1 (A) and G7 (B).](image)

Fig. 4. Monomer units of graphynes GY1 (A) and G7 (B). Upper half of GY7 monomer unit is shown only. The entire monomer could be visualized by connecting the upper part with its reflection relative to the zig-zag line, using two aromatic bonds.

2.5. Polycyclopentadienes

Polycyclopentadienes (Fig. 5) are remarkable because they contain odd-sized cycles and can be easily generated as long-chain polymers.
Table 3
Kekulization of graphynes and graphyne nanotubes of various degrees of polymerization.

| Chain length | Formula | Time (ms) | Formula | Time (ms) | Formula | Time (ms) | Formula | Time (ms) | Formula | Time (ms) |
|--------------|---------|-----------|---------|-----------|---------|-----------|---------|-----------|---------|-----------|
|              |         |           | Tube    | C_{192}H_{72} | 0.5      | C_{1920}H_{72} | 2.06    | C_{19200}H_{72} | 29.8    | C_{192000}H_{72} | 332     |
|              |         |           | GY1     | C_{136}H_{74} | 0.45     | C_{1360}H_{146} | 1.36    | C_{13600}H_{666} | 13.7    | C_{136000}H_{60066} | 176     |
|              |         |           | GY7     | C_{188}H_{66} | 0.39     | C_{1880}H_{122} | 2.44    | C_{18800}H_{882} | 25.3    | C_{188000}H_{40082} | 358     |
|              |         |           |         |            |          |           |         |            |          |           |          |
|              |         |           | Tube    | C_{1200}H_{72} | 332     | C_{12000}H_{72} | 7201    | C_{120000}H_{72} | 59,515  |           |          |
|              |         |           | GY1     | C_{1360}H_{666} | 176     | C_{136000}H_{60066} | 4811    | C_{1360000}H_{600066} | 75,938  |           |          |
|              |         |           | GY7     | C_{1880}H_{40082} | 6414    | C_{18800}H_{400082} | 82,032  |           |          |           |          |
We studied three types of polycyclopentadienes. In the first type, after generation of the structure, free valences of the end-group carbon atoms with arrows were replaced with hydrogens. That resulted in a methylene end group. In the second type, free valences in the left-hand end group were combined with these in the right-hand end group to form a cycle. In the third type, end groups were combined in a crisscross fashion for form a Moebius loop. The results of calculations are shown in Table 4.

The legend for the column headers is the same as for Table 2. Parenthetical values in the column No. non-existent are the counts of structures for which Kekulé representations were found using a backtrack algorithm [5].

The calculation statistics for polycyclopentadienes differ from those for the rest of studied compounds. Specifically, 100 calculations were performed for the number of atoms in the $10^6$–$10^7$ range and 10 calculations – for the number of atoms $> 10^7$. The increase in the number of calculations was due to the probabilistic nature of the algorithm for polycyclopentadienes, requiring multiple initial approximations for the generation of Kekulé structures.
| Chain length | Formula | Time (ms) | No. non-existent | Avg. no. iterations | Max. no. iterations | Unmatched atoms statistic |
|--------------|---------|-----------|------------------|---------------------|---------------------|---------------------------|
|              |         |           |                  |                     |                     |                           |
| Linear       |         |           |                  |                     |                     |                           |
| 1            | C_{150}H_{54} | 0.53       | 0 + 0            | 1.93                | 13                  | 372 607 21 0              |
| 10           | C_{1500}H_{504} | 4.08       | 0 + 0            | 4.63                | 25                  | 747 250 3 0               |
| 10           | C_{1400}H_{504}N_{2} | 35.2       | 0 + 10(1)        | 22.9                | 264                 | 477 431 81 1             |
| 10           | C_{1400}H_{504}N_{4} | 65.75      | 4 + 91(71)       | 68.9                | 300                 | 262 393 210 31            |
| 10           | C_{1400}H_{504}N_{8} | 189        | 87 + 348(261)   | 163                | 299                 | 50 183 269 247            |
| 100          | C_{1500}H_{504} | 67.7       | 0 + 0            | 5.60                | 36                  | 852 146 2 0               |
| 1000         | C_{1500}H_{504} | 1192       | 0 + 0            | 5.61                | 47                  | 850 150 0 0               |
| 10000        | C_{15000}H_{504} | 47354      | 0 + 0            | 7.57                | 33                  | 89 11 0 0                 |
| 100000       | C_{15000}H_{504} | 505401     | 0 + 0            | 6.5                 | 22                  | 7 3 0 0                   |
| Cyclic       |         |           |                  |                     |                     |                           |
| 1            | C_{150}H_{50} | 0.34       | 0 + 0            | 1.23                | 5                   | 289 681 30 0              |
| 10           | C_{1500}H_{50} | 2.77       | 0 + 0            | 2.89                | 17                  | 624 332 44 0              |
| 100          | C_{15000}H_{50} | 33.6       | 0 + 0            | 2.90                | 28                  | 622 331 47 0              |
| 1000         | C_{150000}H_{50} | 585        | 0 + 0            | 2.80                | 14                  | 634 335 31 0              |
| 10000        | C_{1500000}H_{50} | 16368     | 0 + 0            | 3.09                | 12                  | 69 29 2 0                 |
| 100000       | C_{15000000}H_{50} | 193859    | 0 + 0            | 2.2                 | 6                   | 7 2 1 0                   |
| Möbius       |         |           |                  |                     |                     |                           |
| 1            | C_{150}H_{50} | 0.36       | 0 + 0            | 1.26                | 4                   | 189 687 121 3             |
| 10           | C_{1500}H_{50} | 3.67       | 0 + 0            | 3.66                | 29                  | 544 382 73 1              |
| 100          | C_{15000}H_{50} | 52.2       | 0 + 0            | 4.41                | 27                  | 616 332 51 1              |
| 1000         | C_{150000}H_{50} | 1381       | 0 + 0            | 4.42                | 29                  | 602 350 48 0              |
| 10000        | C_{1500000}H_{50} | 24147      | 0 + 0            | 3.75                | 26                  | 52 42 6 0                 |
| 100000       | C_{15000000}H_{50} | 604359     | 0 + 0            | 6.9                 | 28                  | 8 0 2 0                   |
Acknowledgements

The authors wish to thank Prof. Geoffrey Hutchison and Dr. Noel O’Boyle for their valuable advice and useful references, as well as to an article reviewer for examples of structures and references to these examples.

Transparency document. Supporting information

Transparency data associated with this article can be found in the online version at https://doi.org/10.1016/j.dib.2018.10.128.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.dib.2018.10.128.

References

[1] S. Trepalin, S. Gurke, M. Akhukov, A. Knizhnik, B. Potapkin, A fast approximate algorithm for determining bond orders in large polycyclic structures, J. Mol. Graph. Model. 86 (2019) 52–65.
[2] (https://github.com/NarainKrishnamurthy/chemposer) (Accessed 12 May 2018).
[3] (http://sharangdhar.github.io/finalReport.html#) (Accessed 12 May 2018).
[4] (http://www.jcrystal.com/steffenweber/gallery/Fullerenes/Fullerenes.html) (Accessed 8 March 2018).
[5] S.V. Trepalin, A.V. Yarkov, I.V. Pletnev, A.A. Gakh, A java chemical structure editor supporting the modular chemical descriptor language (MCDL), Molecules 11 (2006) 129–141. https://doi.org/10.3390/11040219.