Introduction

Lipids are beside DNA/RNA, proteins, and metabolites the most frequently occurring biomolecules in organic cells. Several lipid classes have a similar molecular structure. Therefore, lipid names were designed to represent their systematic structure in a clear and concise way to help classify and therefore distinguish lipids by name. (1, 2) Within the past two decades, a systematic shorthand notation for lipids has evolved, spurred by initial standardization approaches within the LipidMaps consortium. (3, 4) With the advent of high-resolution mass spectrometry (MS), as the key tool to investigate lipids, the hierarchical representation of lipids was extended to be able to report multiple levels of structural knowledge. (5, 6) Since mass spectrometry is currently the dominant technology for the identification and quantification of lipids, (7, 8) these notations are primarily designed to satisfy the requirements in that application area. Structural knowledge about lipids identified with MS can be represented as a hierarchical tree or table (see Table 1), as introduced by LipidHome and SwissLipids. Goslin was designed to address the following pressing issues in the lipidomics field: 1) to simplify the implementation of lipid name handling for developers of mass spectrometry-based lipidomics tools; 2) to offer a tool that unifies and normalizes the main existing lipid name dialects enabling a lipidomics analysis in a high-throughput fashion. We provide implementations of Goslin in four major programming languages, namely C++, Java, Python 3, and R to kick-start adoption and integration. Further, we set up a web service for users to work with Goslin directly. All implementations are available free of charge under a permissive open source license.

Grammar, Lipid Nomenclature, Computational Lipidomics

Correspondence: robert.ahrens@univie.ac.at

Results

To overcome these challenges, we developed the 'Grammar of Succinct Lipid Nomenclature' (Goslin). It is designed to act as a library for the development of lipidomics tools providing a standardized data structure for storing structural lipid information. The parsing of lipid names as well as the lipid name generation are the main functions of Goslin. We therefore defined a context-free grammar (11) that defines rules and productions for all structural properties of the lipid nomenclature, including mass spectrometry specific information about unlabeled and heavy isotope labeled species, as well as fragments. A short extract of that grammar is illustrated in Figure 1. Currently, the grammar covers 289 lipid names, which includes for example Cer(d18:1/16:0), Cer(d18:1/16:0), d18:1/16:0 Cer, Cer 18:1/2/16:0, or CER[N(16)S(18)].

Although some of these lipid names may look very similar or differ perhaps only in one character, a simple string matching fails when these names are not exactly equal. As a consequence, long and error-prone manual curation is necessary in order to streamline lists of lipid names for their processing in follow-up analysis scripts, workflows, or tools, or for their submission to research data repositories.
Table 1. Hierarchical classification of PE(16:1(6Z)/16:0). Higher levels represent the lipid with less detail and more summarized information, e.g. the top most category level reports that PE(32:1) as a species is a Glycerophospholipid. The species level expresses only the head group (HG), and the sum of # carbon atoms and # double bonds in both fatty acyls (FAs). Intermediate levels report additional details about FA composition (molecular subspecies) and SN position with respect to the head group (structural subspecies). At the lowest level, the FA composition (# carbon atoms, # double bonds), their SN position (first is SN1, second is SN2), and the position of the double bond are expressed (6th carbon in SN1 FA, in Z configuration). The ‘Level’ column details the mapping of Goslin’s levels to the respective levels in the LipidMaps (LM) and SwissLipids hierarchies (SL). Hydroxyl groups have been omitted from this example for brevity, but they are also fully supported by Goslin.

| Level                        | Name                                      | Description               |
|------------------------------|-------------------------------------------|---------------------------|
| Category (LM)                | Glycerophospholipids                      | Lipid category            |
| Class (LM)                   | Glycerophosphoethanolamine                | Lipid class               |
| Species (SL, LM Subclass)    | Phosphatidylethanolamine (32:1), PE(32:1)| HG, FA summary            |
| Molecular Subspecies (SL)    | PE(16:0_16:1)                             | HG, two FAs               |
| Structural Subspecies (SL)   | PE(16:1_16:0)                             | HG, SN pos. for FA1 and FA2|
| Isomeric Subspecies (SL, LM) | PE(16:1(6Z)/16:0)                         | HG, SN pos., Stereo       |

Figure 1. Snippet of the Goslin grammar: on the left hand side of each colon, one non-terminal (NT) rule defines a set of production rules which are either terminal or non-terminal rules. For instance, the production rule `lipid_eof` defines a lipid name by identifying either a pure lipid name or a pure lipid name followed by an identified scheme of an adduct. The complete grammars are provided in our repositories (see https://github.com/lifs-tools/goslin).

This feature simplifies several use-cases, such as the computation of the distribution of lipids among the lipid classes or lipid categories. The polyglot approach already covers several existing dialects of the lipid nomenclature and can be extended arbitrarily to cover new lipid categories and classes, head groups and FA modifications. Tailored event handlers (one for each grammar) transfer all recognized elements of the structural lipid name into a common data structure that reflects the hierarchy similarly to that used in LipidHome(9) and SwissLipids.(10) Our implementations are written in C++ (version 11 and higher), Java (version 11 and higher), Python (version 3), and R (version 3.6 and higher) and can easily be included into existing lipidomics tools. Detailed descriptions on the integration and usage of the programming libraries are provided in the Supplementary Information. Additionally, we provide several instructive code snippets on the usage of Goslin for input and output of lipid names. We benchmarked our implementations with respect to their execution time for parsing. As a benchmark device, we used an Intel(R) Xeon(R) 2.80 GHz quad core desktop computer with 16 GB RAM. The results are listed in Table 2. All implementations are fast enough to parse lipid name lists of regular experiment sizes (about 1000 lipid names) within a second. To provide a user-friendly alternative to the libraries and command line interface (CLI), we developed a Goslin web application (see https://apps.lifs.isas.de/goslin) that offers form-based lipid name translation, as well as a REST API for language agnostic, programmatic translation.
Lipid names can be directly copied and pasted into the submission form of the web application or can be uploaded as a file, with one lipid name per line. Upon submission, the application parses all lipid names and provides a result list with all identified lipids and additional structural information, as well as cross-links to LipidMaps and SwissLipids, where applicable. For lipid names that cannot be parsed, it reports specific errors within the submission form to help the user fix the name. The result list can also be downloaded in a spreadsheet friendly tab-separated value format. For programmatic access, the user can send a web request (HTTPS) to translate their lipid list using the provided REST application programming interface (API) by sending a POST request with a JavaScript Object Notation (JSON) body (JSON list of lipid names). The response is a JSON list with embedded objects (for usage and tutorials, see Supplementary Information). The REST API is meant to be used by automated workflows, when neither of our offline implementations is suitable or applicable. We also offer instructions to build a Docker image of the web-application for local and customized deployments, as well as a Bioconda recipe for the jgoslin CLI for inclusion in automated bioinformatics workflows.

Conclusions

With Goslin, we provide another building block towards a more harmonized and inter-operable bioinformatics tool landscape for lipidomics. Equipped with multiple grammars, Goslin is able to interpret several lipid name dialects, such as the LipidMaps nomenclature, the shorthand nomenclature by Liebisch et al., the SwissLipids nomenclature, and the HMDB nomenclature. Goslin translates them into a normalized, up-to-date nomenclature and structured representation. Currently, we are working on several improvements of the library, 1) to offer an implementation of Goslin in additional programming languages (e.g., C#); 2) to improve the performance of the existing implementations; and 3) to add more structural lipid classes and categories into the grammar for a higher coverage of lipids. We already applied Goslin in other recent projects(16, 17) and achieved a high performance boost especially when exchanging our lipid lists with our collaboration partners. The Goslin implementations are freely available under https://github.com/lifs-tools/goslin under the terms of liberal open source licenses.

Author contributions

D.K., N.H. designed the grammars, N.H. implemented the Java and R libraries and the Goslin web-application. D.K. implemented the C++ and Python versions. B.P. and R.A. contributed their structural and chemical knowledge of lipids. R.A. supervised the project. All authors wrote and approved of the manuscript.

Acknowledgements

The supports by the Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen and the Regierende Bürgermeister von Berlin, Senatskanzlei Wissenschaft und Forschung - inkl. Wissenschaft und Forschung, and the Bundesministerium für Bildung und Forschung (BMBF) to DK (Code 031L0108A) and NA are gratefully acknowledged. We thank Harald Köfeler and Gerhard Liebisch for the close collaboration within the International Lipidomics Society on lipidomics standardization.

Bibliography

1. IUPAC-IUB Commission on Biochemical Nomenclature. The nomenclature of lipids. Journal of Lipid Research, 8:523–528, 1967.
2. MA Chester. IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN) Nomenclature of glycolipids. European Journal of Biochemistry, 257(2):293–298, 1998.
3. Eoin Fahy, Shankar Subramaniam, H. Alex Brown, Christopher K. Glass, Althof H. Merrill, Robert C. Murphy, Christian R. H. Raatz, David W. Russell, Yusuke Seya, Walter Shaw, Takah Shimizu, Friedrich Spener, Gerrit van Meer, Michael S. VanNieuwenhze, Stephen H. White, Joseph L. Witztum, and Edward A. Dennis. A comprehensive classification system for lipids. Journal of Lipid Research, 46(5):839–862, 2005.
4. Eoin Fahy, Shankar Subramaniam, Robert C. Murphy, Masanori Nishijima, Christian R. H. Raatz, Takah Shimizu, Friedrich Spener, Gerrit van Meer, Michael J. O. Wakelam, and Edward A. Dennis. Update of the LIPID MAPS classification system for lipids. Journal of Lipid Research, 50(Supplement):S9–S14, 2009.
5. Gerhard Liebisch, Juan Antonio Viscaino, Harald Köfeler, Martin Titzmüller, William J. Griffin, Gerd Schmitz, Friedrich Spener, and Michael J. O. Wakelam. Shorthand notation for lipid structures derived from mass spectrometry. Journal of Lipid Research, 54(6):1523–1530, 2013.
6. Joch K. Pauling, Martin Hermansson, Jürgen Hartler, Klaus Christiansen, Sandra F. Gallego, Bing Peng, Robert Ahrends, and Christer S. Ejsing. Proposal for a common nomenclature for fragment ions in mass spectra of lipids. PLOS ONE, 12(11):1–21, 2017.
7. Stephen J. Blanksby and Todd W. Mitchell. Advances in mass spectrometry for lipidomics. Annual Review of Analytical Chemistry, 3(1):433–465, 2010.
8. Markus R. Wenk. Lipidomics: New tools and applications. Cell, 143(6):888 – 895, 2010.
9. Joseph Foster, Pablo Moreno, Antonio Fabregat, Henning Hermjakob, Christoph Steinbeck, Rolf Apweiler, Michael Wakelam, and Juan Viscaino. LipidHome: A Database of Theoretical Lipid Structures Optimized for High Throughput Mass Spectrometry Lipidomics. PloS one, 8:e46951, 05 2013.
10. Lucila Aimo, Robin Liechti, Nevila Hya Kouysipkel, Anne Nikrjad, Anne Gleizes, Lou Gitz, Dmitry Kuznetsova, Fabrice PA, David, F. Giusu van der Goot, Howard Riemann, Lydia Bougueleret, Ioannis Xenarios, and Alan Bridge. The SwissLipids knowledgebase for lipid biology. Bioinformatics, 31(17):2860–2866, 2015.
11. N. Chomsky. Three models for the description of language. IRE Transactions on Information Theory, 2(3):113–124, 1956.
12. Stephen Cole Kleene. Representation of events in nerve nets and finite automata. Technical Report, RAND PROJECT AIR FORCE SANTA MONICA CA, 1951.
13. Terence Parr. The Definitive ANTLR 4 Reference. O’Reilly UK Ltd, 2013. ISBN 1934356999.
14. David S. Wishart, Dan Tour, Craig Knox, Roman Eisner, An Chi Guo, Nelson Young, Dean Cheng, Kevin Jewell, David Arndt, Sumit Sawhney, Chris Fung, Lisa Nikolai, Mike Lewis, Marie-Aude Coutouly, Ian Forsythe, Peter Tang, Savita shrivastava, Kevin Jeroncic, Paul Stothard, Godwin Amogbey, David Blod, David, D. Hau, James Wagner, Jessica Miniaci, Melissa Clements, Mulu Gbretmedhin, Natalie Guo, Ying Zhang, Gavin E. Duggan, Glen D. MacInnis, Alm M. Wejle, Reza Dowlatabadi, Fiona Barnholt, Derrick Cleve, Russ Greiner, Liang Li, Tom Marrie, David, D. Sykes, Hans J. Vogel, and Lori Querengesser. HMDB: the Human Metabolome Database. Nucleic Acids Research, 35(suppl_1):D521–D526, 2007.
15. Björn Gröning, Ryan Dale, Andreas Sjödin, Brad A. Chapman, Jillian Rowe, Christopher H. Tomkins-Tinch, Renan Valerius, Johannes Kötter, and The Bioconda Team. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nature Methods, 15(7):475–476, 2018.
16. Bing Peng, Ben Weintrub, Cristina Coman, Srigayatri Ponnaiyan, Rakesh Sharma, Björn Tews, Dominic Winter, and Robert Ahrends. A comprehensive high-resolution targeted workflow for the deep profiling of sphingolipids. Analytical Chemistry, 89(22):12480–12487, 2017.
17. Bing Peng, Sascha Geue, Cristina Coman, Patrick Münzer, Dominik Kopczynski, Canan Has, Nils Hoffmann, Mailin-Christin Marke, Florian Lang, Albert Sickmann, Meinrad Gawaz, Oliver Borst, and Robert Ahrends. Identification of key lipids critical for platelet activation by comprehensive analysis of the platelet lipidome. Blood, 132(5):e1–e12, 08 2018.