Bioinformatics, 31(5), 2015, 621–625
doi: 10.1093/bioinformatics/btu723
Advance Access Publication Date: 29 October 2014
Discovery Note

OXFORD

Databases and Ontologies

Automated structural classification of lipids by machine learning

Ryan Taylor[1,*], Ryan H. Miller[1], Ryan D. Miller[1], Michael Porter[1], James Dalgleish[2] and John T. Prince[1]

[1]Department of Chemistry and Biochemistry and [2]Department of Microbiology and Molecular Biology, Brigham Young University, Provo, UT 84602, USA

*To whom correspondence should be addressed.
Associate Editor: Igor Jurisica

Received on May 17, 2014; revised on October 23, 2014; accepted on October 26, 2014

Abstract

Motivation: Modern lipidomics is largely dependent upon structural ontologies because of the great diversity exhibited in the lipidome, but no automated lipid classification exists to facilitate this partitioning. The size of the putative lipidome far exceeds the number currently classified, despite a decade of work. Automated classification would benefit ongoing classification efforts by decreasing the time needed and increasing the accuracy of classification while providing classifications for mass spectral identification algorithms.

Results: We introduce a tool that automates classification into the LIPID MAPS ontology of known lipids with >95% accuracy and novel lipids with 63% accuracy. The classification is based upon simple chemical characteristics and modern machine learning algorithms. The decision trees produced are intelligible and can be used to clarify implicit assumptions about the current LIPID MAPS classification scheme. These characteristics and decision trees are made available to facilitate alternative implementations. We also discovered many hundreds of lipids that are currently misclassified in the LIPID MAPS database, strongly underscoring the need for automated classification.

Availability and implementation: Source code and chemical characteristic lists as SMARTS search strings are available under an open-source license at https://www.github.com/princelab/lipid_classifier.

Contact: ryanmt@byu.net

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Lipids are a fundamental component of biological systems and perform diverse roles in many cellular pathways. They comprise several thousands of structurally distinct species whose diversity is preserved by dedicated cellular systems (Subramaniam *et al.*, 2011). The lipid composition of a cell is linked to its function; hence lipids are excellent subjects for gaining insight into biological systems and predicting abnormalities (Sone *et al.*, 2012). Indeed, lipids are known to play a major role in diverse diseases afflicting millions, including obesity (Pietiläinen *et al.*, 2007, 2011; Yetukuri *et al.*, 2007), diabetes (Gross and Han, 2009; Han *et al.*, 2007), asthma (Heeley *et al.*, 2000; Wright *et al.*, 2000), hypertension (Graessler *et al.*, 2009), arthritis (Fuchs *et al.*, 2005) and cancers (Hilvo *et al.*, 2011; Xiao *et al.*, 2001). Lipidomics—the analysis of the lipid composition, localization and activity of a cellular or physiological system—is a burgeoning field of research (Wenk, 2010).

One major difficulty in studying lipids is dealing with their great structural diversity. To help address this challenge, the LIPID MAPS Consortium has created and is refining the LIPID MAPS Lipid Classification System (LMLCS), which has become the de facto ontology used in lipid research. With this ontology, the lipid research community is able to discuss predicted lipid properties and cellular functions in ways that would otherwise be impossible.

Indeed, classification of biomolecules is a prerequisite to any systems biology approach (Fahy *et al.*, 2005). This is particularly true in the area of mass spectrometry identification, where theoretical fragmentation spectra (used for matching with each experimental spectrum) are generated by different sets of rules based on a biomolecule's type (protein, metabolite, etc.). The principle holds true for lipids: nearly all existing identification approaches require that a lipid be classified (although sometimes implicitly) to generate a theoretical spectrum from a lipid's structure (Herzog *et al.*, 2012; Kangas *et al.*, 2012; Kind *et al.*, 2013; Song *et al.*, 2007). Classification makes possible restricted search space structural comparisons and even fundamental tasks such as representing lipids in a systematic fashion (Fahy *et al.*, 2009).

The benefits resulting from classification are definite, but these benefits are currently inaccessible to lipids, which have not been previously classified. In an enormous feat, LIPID MAPS has classified over 38 000 lipids over the last ~15 years. Still, there are more than 120 000 lipid species (Kind *et al.*, 2013) and probably more when considering oxidative modifications, yet undiscovered natural products and unanticipated future synthetic modifications. And, although automatic classification tools have been alluded to (Fahy *et al.*, 2009), currently there is no publicly available software for the automated classification of lipids. Although classification can be performed manually, manual classification cannot be used in any automated software pipelines, does not scale well and may not always be accurate (Danziger *et al.*, 2011).

We present an approach to generate a classifier trained on the LMLCS, which can be used to classify novel lipids automatically and assist in manual classification workflows.

2 Methods

More extensive methodology is given in the methods supplement. All models were classified upon the LIPID MAPS structural database (LMSD) as downloaded from lipidmaps.org.

2.1 Chemical language and identifying features

We used Rubabel, a cheminformatics software suite built upon the OpenBabel library, to provide programmatic representation of chemical structures (Smith *et al.*, 2013), which were searched by SMILES Arbitrary Target Specification (SMARTS) search strings to produce a list of chemically identifying structural characteristics, as detailed in Supplementary Table S1. Each identifying structural characteristic formed a binary variable indicating presence of a feature (Quinlan, 1993) or representing a numerical count of feature matches.

2.2 Classification by machine learning

WEKA (version 3.7.11) machine learning algorithms (Hall *et al.*, 2009) were compared by several numerical performance measures (see Supplementary Algorithm Selection). The J48 decision tree algorithm was selected based on performance, speed and interpretability. Classification accuracy was optimized upon a 15% subset of the LMSD. The WEKA-produced decision trees contain a rule-by-rule determination of lipid classification based on the identifying structural characteristics and were trained upon a randomized 90% sampling of the entire training dataset.

Each lipid of the training dataset, which consisted of the entire LMSD, was structurally analyzed to produce a structure feature list, which was then split into the hierarchal levels of the LMLCS (see Fig. 1, panel A). Each feature list was then analyzed by WEKA to produce decision trees representing the classification steps at every hierarchical level.

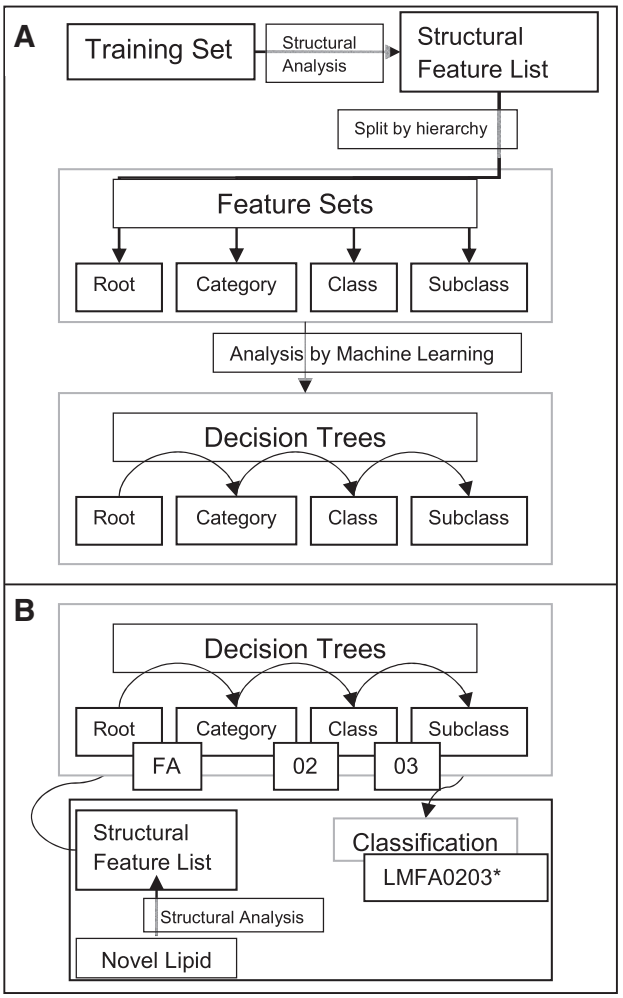

Fig. 1. Lipid classification workflow schematic. (**A**) A lipid classification is constructed from a training set of lipids with known classifications. Each lipid in the training set is analyzed structurally to produce a list of structural characteristics which can be used for machine learning analysis. These feature sets are split into hierarchal groups according to their classification. Machine learning analysis builds a distributed set of decision trees at each hierarchal level. (**B**) A novel lipid is analyzed structurally and then compared with the WEKA produced decision trees at each hierarchal level to generate a complete classification. Overlaid boxes highlight the contributions of each hierarchical layer to a complete classification

This generated hierarchy of classification trees were loaded into a programmatic classification system (see Fig. 1, panel B) implemented in Ruby, which classified each given lipid structure by (i) generation of a structural feature list and (ii) application of each hierarchical decision tree.

The Ruby classification system was evaluated for accuracy across all hierarchical levels by examination of the entire LMSD. Each classification was considered a miss or hit in two categories, category classification and category-internal classification. These two scores present scores for both the complete classification and the relative importance of category selection.

Classification model accuracy was confirmed by manual evaluation of all misclassified lipids and annotated any potential ontology changes. When the classification models assigned lipids to categories that were not indicated by their structure, we identified relevant structural features which could provide correct classification assignment.

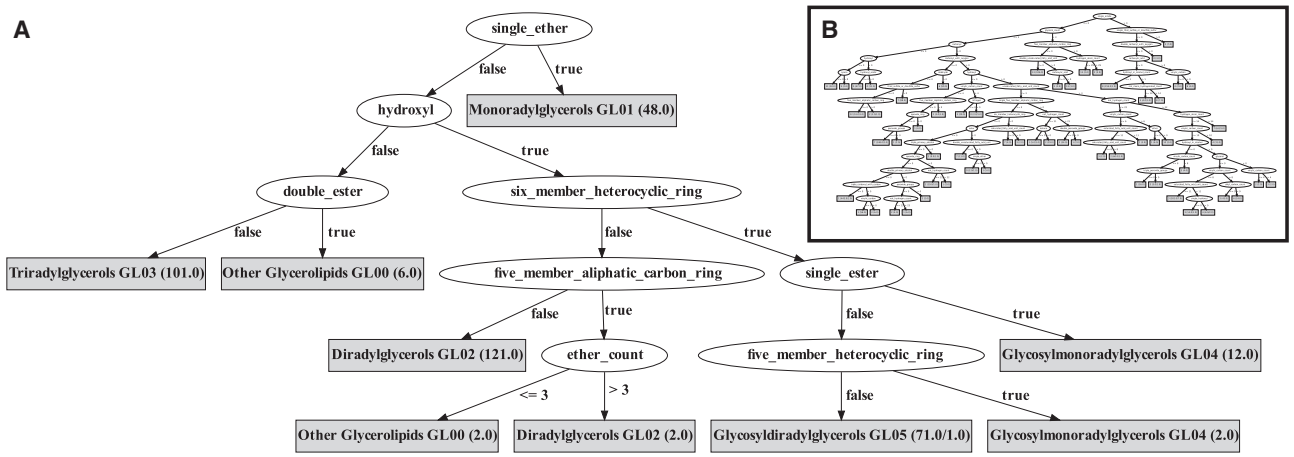

Fig. 2. Representative decision trees for LMSD classifications. (**A**) Glycerolipid category (GL) into six class levels, GL00-GL05, based upon chemical features. (**B**) Fatty acyl (FA) category demonstrates increasing complexity

2.3 Evaluation of novel lipids

Evaluation of the trained classifier was performed upon an extracted subset of the LipidBank database (Watanabe *et al.*, 2000), which consisted of some 1195 molecules, many of which are similar to molecules which are contained in the LMSD. Accuracy was measured by manual evaluation to determine if these lipids were (i) properly categorized into and (ii) fit within the LMLCS. Hits and misses were only counted when the similar lipids were found in the LMSD.

2.4 Statistical considerations and model validation

Algorithm selection was performed by split-percentage validation at 66%. WEKA derived accuracy scores for these data were compared with the J48 model based on 90% of the training dataset with cross-validation.

3 Results

The classification model for each dataset produces decision tree output for both manual interpretation and programmatic analysis as in Figure 2.

3.1 Classifier model performance

The comprehensive reference implementation in the Ruby language provides <1.2% error across all classification levels (see Supplementary Misclassifieds). An equivalent implementation trained on a 66% split percentage yields 3.0% error. At the category level, we reach 99.98% accuracy (as described in Table 1) when suggested improvements to the existing ontology are followed as outlined in the Supplementary Tables S2 and S3.

3.2 Novel lipid analysis

Evaluation of classified lipids from the LipidBank database demonstrates the capability of this classification to handle novel lipid classes. In 780 novel lipids, we correctly classified 63% of the lipids, which were manually determined to be within the current LIPID MAPS ontology (see Supplementary Novel Lipid Analysis).

3.3 Ontology modifications

We carefully inspected the results of the model built upon cross-validation training and discovered ~150 lipids that are misclassified in the LMSD (as detailed in the Supplementary Misclassified).

Table 1. Classifier performance for entire LMSD and categories slices of the LMLCS

	Number of lipids	Category level error counts (%)	Within category error counts (%)
Entire LMSD	36 785	3 (0.01)	426 (1.16)
Fatty acyl (FA)	5763	1 (0.02)	3 (0.05)
Glycerolipids (GL)	7538	0 (0.00)	1 (0.01)
Sterol lipids (ST)	2561	0 (0.00)	16 (0.62)
Prenol lipids (PR)	1193	1 (0.08)	0 (0.00)
Saccharolipids (SL)	1293	0 (0.00)	0 (0.00)
Polyketides (PK)	6744	0 (0.00)	11 (0.16)
Sphingolipids (SP)	3934	0 (0.00)	385 (9.79)
Glycerophospholipids (GP)	7759	1 (0.01)	10 (0.13)

Category level errors represent misclassifications, which put a lipid into the wrong category and are excluded from 'within category' error counts. Within category, errors represent any misclassification of a lipid other than a category level error.

The number of misclassified lipids indicates the difficulty of hand curating a database the size and complexity of the LMLCS. In addition, the ability of our classifier to find these misclassified lipids strongly supports the utility of our automated approach.

4 Discussion

4.1 Classifier design and performance

The high degree of accuracy we achieved suggests that the LMCLS is a structurally coherent representation of lipid structural diversity. Despite an origin in synthetic pathways, the 2009 revision of the ontology is largely structurally derived, making an analysis like this successful. The chemical characteristics chosen provided robust machine learning attributes across a majority of potential classifier algorithms (see the Supplementary Algorithm Selection). The J48 algorithm was chosen because it provided nearly identical performance to the LMT algorithm but at 1/200th of the analysis time. Taken collectively, the 90% and 66% training set analyses and algorithm selection analysis suggest that the chemical features selected are robust attributes for classification.

The neutral glycolipid class of the sphingolipid (SP) category remains a source of error as the current ontology fails to encompass

the diversity of these ~3000 glycolipids. The classification of these lipids is dependent upon sugar oligomer length to differentiate nomenclature precedence. Thus, resolution of this issue will require (i) consideration of ontology changes to better reflect the diversity of neutral glycolipid structures and (ii) consideration of structural characteristics beyond the SMARTS systems currently employed, such as a longest-path finding algorithm. The current diversity of glycolipids within the LMSD fails to encompass possible diversity, as there are only a few sugar monomers contained therein, fucose, mannose, galactose and glucose. We suggest that future efforts investigate this diversity more fully and propose more sweeping ontology changes, such as subclasses, which allow for classification of unrepresented sugar monomers or subclasses, which only contain lipids with a specified sugar root structure.

4.2 Novel lipid analysis

Evaluation of a novel lipid library demonstrates the capacity of our classifier to streamline novel lipid analysis. It further demonstrates a need for further refinements within the established ontology. Many missed lipids are likely missed due to the limited size of some subclasses. Additionally, many lipids which were not currently contained within the LMLCS would be very appropriately grouped into ontology not found in the LMLCS. The differential in performance between the novel analysis and the curated database was expected given that the current ontology does not endeavor to deal with edge-case lipids. Many of the tested lipids were much shorter than the chain lengths represented in the LMLCS and thus beyond the capability of our models to differentiate between them. This analysis provides an understanding of the limitations of the LMLCS in its current form and can guide future efforts in accommodating novel lipids.

4.3 Ontology modifications

Many of the misclassifications are due to small structural differences. Lipid LMGP04040006 is classified as a dialkylglycerophosphoglycerol. The structure contains an acyl group instead of an alkyl group, corresponding to our classifier's assignment for this lipid as a 1-acyl, 2-alkylglycerophosphoglycerol or LMGP0411. The fatty acid LMFA01010053, in the straight chain fatty acid class, is clearly branched, as classified by our analysis.

The model excelled at assigning lipids that contain multiple structural features. Several fatty acids are both branched and unsaturated. These fatty acids are distributed among both the unsaturated fatty acids and the branched fatty acids even though they are structurally similar. The model followed the established ontology that branching takes precedence over unsaturation and assigned these lipids to the correct classification.

In accordance with IUPAC guidelines (Chester, 1997, 1999; Horton, 1999; JCBN, 1998a, 1998b, 1999), neutral glycosphingolipids were assigned a group based on their root sugar chain, the first four sugars, and their linkage to the SP. LIPID MAPS suggested nine groups (or series, LMSP0501-09). There are several sets of distinct lipids within the Neolacto subgroup that do not fit into it nor any other group. These sets (LMSP0505DC-F and LMSP0505DM-N) contain 32 and 16 lipids with two unique roots. We suggest implementation of two new ontology groups for these distinct roots: gluco-globo (LMSP0510) and galacto-lacto (LMSP0511). Gluco-globo highlights the similarity to the isoglobo (LMSP0506) series, excepting the terminal *N*-acetyl glucosamine. Galacto-gluco associates with the Gala series (LMSP0509) in their shared repeated galactose monomers while highlighting the terminal glucose monomer. These new ontologies would represent the incorrectly classified lipids in their own ontology and improve classification of all neutral glycosphingolipids.

4.4 Future directions

Future work should expand the classification system to classify non-lipids into general categories and improve upon some existing limitations of the extensive sugar nomenclature within the SP category. Future efforts will shorten analysis time per lipid. Future work should evaluate whether the need for an alternative ontology which enables multiple classifications for a given lipid exists, such as the aforementioned branching and unsaturation precedence. An alternative ontology derived from machine learning clustering analysis might provide a more natural fit and reduce the prevalence of 'other' classification categories.

Funding

This work was supported by BYU Institutional Funds, BYU Undergraduate Research Awards and BYU CHIRP Grant.

Conflict of interest: none declared.

References

Chester,M.A. (1997) Nomenclature of glycolipids. *Pure Appl. Chem.*, **69**, 2475–2487.

Chester,M.A. (1999) IUPAC-IUB joint commission on biochemical nomenclature (JCBN) nomenclature of glycolipids. *J. Mol. Biol.*, **286**, 963–970.

Danziger,S. *et al.* (2011) Extraneous factors in judicial decisions. *PNAS*, **108**, 6889–6892.

Fahy,E. *et al.* (2005) A comprehensive classification system for lipids. *J. Lipid Res.*, **46**, 839–861.

Fahy,E. *et al.* (2009) Update of the LIPID MAPS comprehensive classification system for lipids. *J. Lipid Res.*, **50** (Suppl), S9–S14.

Fuchs,B. *et al.* (2005) The phosphatidylcholine/lysophosphatidylcholine ratio in human plasma is an indicator of the severity of rheumatoid arthritis: investigations by 31P NMR and MALDI-TOF MS. *Clin. Biochem.*, **38**, 925–933.

Graessler,J. *et al.* (2009) Top-down lipidomics reveals ether lipid deficiency in blood plasma of hypertensive patients. *PLoS One*, **4**, e6261.

Gross,R.W. and Han,X. (2009) Shotgun lipidomics of neutral lipids as an enabling technology for elucidation of lipid-related diseases. *Am. J. Physiol. Endocrinol. Metab.*, **297**, E297–E303.

Hall,M. *et al.* (2009) The WEKA data mining software. *ACM SIGKDD Explor. Newsl.*, **11**, 10.

Han,X. *et al.* (2007) Alterations in myocardial cardiolipin content and composition occur at the very earliest stages of diabetes: a shotgun lipidomics study. *Biochemistry*, **46**, 6417–6428.

Heeley,E.L. *et al.* (2000) Phospholipid molecular species of bronchoalveolar lavage fluid after local allergen challenge in asthma. *Am. J. Physiol. Lung Cell. Mol. Physiol.*, **278**, L305–L311.

Herzog,R. *et al.* (2012) LipidXplorer: a software for consensual cross-platform lipidomics. *PLoS One*, **7**, e29851.

Hilvo,M. *et al.* (2011) Novel theranostic opportunities offered by characterization of altered membrane lipid metabolism in breast cancer progression. *Cancer Res.*, **71**, 3236–3245.

Horton,D. (1999) *Advances in Carbohydrate Chemistry and Biochemistry*, 55th ed, ISBN 9780124080928.

JCBN. (1998a) IUPAC-IUB joint commission on biochemical nomenclature (JCBN) nomenclature of glycolipids: recommendations 1997. *Eur. J. Biochem.*, **257**, 293–298.

JCBN. (1998b) Nomenclature of glycolipids. *Carbohydr. Res.*, **312**, 167–175.

JCBN. (1999) IUPAC-IUB joint commission on biochemical nomenclature (JCBN) nomenclature of glycolipids: recommendations 1997. *Glycoconj. J.*, **16**, 1–6.

Kangas,L.J. *et al.* (2012) In silico identification software (ISIS): a machine learning approach to tandem mass spectral identification of lipids. *Bioinformatics*, **28**, 1705–1713.

Kind,T. *et al.* (2013) LipidBlast in silico tandem mass spectrometry database for lipid identification. *Nat. Methods*, **10**, 755–758.

Pietiläinen,K.H. *et al.* (2007) Acquired obesity is associated with changes in the serum lipidomic profile independent of genetic effects–a monozygotic twin study. *PLoS One*, **2**, e218.

Pietiläinen,K.H. *et al.* (2011) Association of lipidome remodeling in the adipocyte membrane with acquired obesity in humans. *PLoS Biol.*, **9**, e1000623.

Quinlan,R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Smith,R. *et al.* (2013) Rubabel: wrapping open Babel with Ruby. *J. Cheminform.*, **5**, 35.

Sone,H. *et al.* (2012) Comparison of various lipid variables as predictors of coronary heart disease in Japanese men and women with type 2 diabetes: subanalysis of the Japan diabetes complications study. *Diabetes Care*, **35**, 1150–1157.

Song,H. *et al.* (2007) Algorithm for processing raw mass spectrometric data to identify and quantitate complex lipid molecular species in mixtures by data-dependent scanning and fragment ion database searching. *J. Am. Soc. Mass Spectrom.*, **18**, 1848–1858.

Subramaniam,S. *et al.* (2011) Bioinformatics and systems biology of the lipidome. *Chem. Rev.*, **111**, 6452–6490.

Watanabe,K. *et al.* (2000) How to search the glycolipid data in "LIPIDBANK for Web" the newly developed lipid database in Japan. *Trends Glycosci. Glycotechnol.*, **12**, 175–184.

Wenk,M.R. (2010) Lipidomics: new tools and applications. *Cell*, **143**, 888–895.

Wright,S.M. *et al.* (2000) Altered airway surfactant phospholipid composition and reduced lung function in asthma. *J. Appl. Physiol.*, **89**, 1283–1292.

Xiao,Y.J. *et al.* (2001) Electrospray ionization mass spectrometry analysis of lysophospholipids in human ascitic fluids: comparison of the lysophospholipid contents in malignant vs nonmalignant ascitic fluids. *Anal. Biochem.*, **290**, 302–313.

Yetukuri,L. *et al.* (2007) Bioinformatics strategies for lipidomics analysis: characterization of obesity related hepatic steatosis. *BMC Syst. Biol.*, **1**, 12.