Automated lipid classification

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

[1]Department of XXXXXXX, Address XXXX etc.

[2]Department of XXXXXXX, Address XXXX etc.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Lipid fragmentation software currently available relies upon 'divide-and-conquer' to partition complex fragmentations; no automated lipid classification exists to facilitate this partitioning.

Results: We introduce a 99.9% accurate automated classification tool of any given lipid, available for immediate use, as well as all materials required for alternative implementations, based upon simple chemical characteristics.

Availability: Source code is available under open-source license at https://www.github.com/princelab/lipid_classifier

Contact: jtprince@chem.byu.edu

1 INTRODUCTION

1.1 Lipids are relevant to human health

Lipids are a fundamental component of biological systems and perform diverse roles in vital cellular pathways. Estimates claim that cellular lipids comprise several thousands of structurally distinct component species and that this diversity is preserved by dedicated cellular pathways and systems (Subramaniam *et al.*, 2011). The lipid composition of a cell is linked to its cellular function; lipids are excellent subjects for understanding cellular function and they are highly predictive of related abnormalities (Sone *et al.*, 2012). Lipids play a major role is diverse diseases afflicting millions, including: obesity (Pietiläinen *et al.*, 2007, 2011; Yetukuri *et al.*, 2007), diabetes (Han *et al.*, 2007; Gross and Han, 2009), asthma (Wright *et al.*, 2000; Heeley *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 field of rapidly increasing importance.

1.2 Lipids are complex and require classification

Lipidomics studies are largely dependent upon mass spectrometer based analysis, which is complicated by the difficulty in obtaining identities from the data. Two predominant methods are used, one of which relies upon chromatographic separation and sample preparation to limit the number of lipid species delivered to the mass spectrometer at a given instant; while the other, shotgun lipidomics by direction injection, is relatively simple in sample preparation, but requires robust data analysis to determine lipid identity. While some lipid identification can be done from precursor mass, many lipids are undistinguishable without fragmentation data. Lipid fragmentation is extremely complex (Hsu and Turk, 2010; Hsu *et al.*, 2007), and most of the existing fragmentation software treat lipids on a class-by-class basis, also known as a 'divide-and-conquer' approach (Herzog *et al.*, 2012; Kind *et al.*, 2013a; Kangas *et al.*, 2012; Song *et al.*, 2007). Thus, fragmentation prediction relies upon appropriate classification of lipids.

Commented [RT1]: Make it more compelling, understandable, and substantial

Proteomics has succeeded in large part due to the relative ease of identifying protein from experimental fragmentation spectra. This approach works for proteins because the rules for fragmentation within a mass spectrometer are understood. Lipidomics will benefit from tools which can provide quality fragmentation spectra useful for lipid identification.

Commented [RT2]: Tie this in or get rid of it!

The most accepted classification scheme is provided by the LIPID MAPS consortium (Fahy *et al.*, 2005, 2009). Currently, classification is performed by hand or is executed during the pre-generation of structural libraries using a few proto-typical lipids (Kind *et al.*, 2013a). Despite the major role of synthetic pathways in the derivation of the LIPID MAPS ontology, each lipid species is structurally distinct. The structural characteristics of lipids contain the identifying characteristics required for complete classification.

Commented [RT3]: Same reference as the next paragraph, where it is definitely required. Do we need to discuss this type of classification? (NO) but how do I restructure this.

And covers some 38000 lipid species. (But that would make this paragraph very hodge-podge and lacking in focus.

Flexible, automated fragmentation requires that lipids be classifiable 'on-the-fly', but no such tool currently exists.

Lipid models suggest a lipidome size in excess of 120,000 species (Kind *et al.*, 2013b). Manual classification lacks both scaling and accuracy (Danziger *et al.*, 2011) when dealing with such a large problem domain. Computational classification tools are necessary to apply current ontologies to novel lipid species.

We present an approach to generate a classifier trained on the LIPID MAPS ontology and structural database (LMSD) which can be used to classify novel lipids.

2 METHODS

2.1 Chemical Language

[*]To whom correspondence should be addressed.

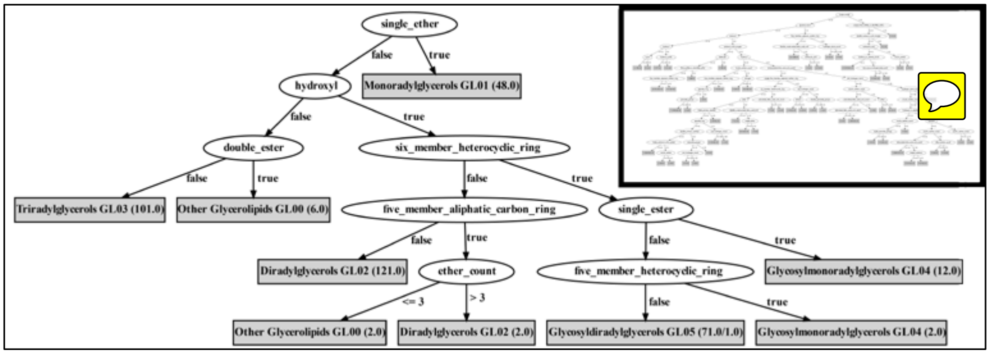

single_ether
false
true
hydroxyl
Monoradylglycerols GL01 (48.0)
false
true
double_ester
six_member_heterocyclic_ring
false
true
Triradylglycerols GL03 (101.0)
Other Glycerolipids GL00 (6.0)
false
true
five_member_aliphatic_carbon_ring
single_ester
false
true
Diradylglycerols GL02 (121.0)
ether_count
false
true
five_member_heterocyclic_ring
Glycosylmonoradylglycerols GL04 (12.0)
<= 3
> 3
Other Glycerolipids GL00 (2.0)
Diradylglycerols GL02 (2.0)
false
true
Glycosyldiradylglycerols GL05 (71.0/1.0)
Glycosylmonoradylglycerols GL04 (2.0)

Fig. 1. Representative decision tree for LMSD classification of the Glycerolipid category (GL) into 6 class levels, GL00-GL05, based upon chemical features.

Rubabel, a cheminformatics software suite built upon the OpenBabel library, provided a programmatic representation of chemical structures.

The SMILES Arbitrary Target Specification, or SMARTS, provides a simple, cross-software method for searching chemical structures for identifying structural characteristics. SMARTS searching is implemented in all major chemistry modeling libraries and provides unambiguous recognition of diverse identifying characteristics.

2.2 Identifying Features

We generated a list of some 300 chemically identifying structural characteristics including: simple functional group recognitions, atom counts, peptide and glycosidic bond recognitions (see further examples in Supplemental Table 1). Each identifying structural characteristic was used to form binary answer questions (Quinlan, 1993).

2.3 Classification by Machine Learning

2.3.1 Optimization of machine learning parameters We examined machine learning and classification algorithms from WEKA (Hall *et al.*, 2009) and elected to use the decision tree style algorithms. J48 optimization was performed at the category level and adopted for all subsequent analyses, setting the pruning confidence (-C) to 0.25 and setting the minimum number of instances (-M) to 1. The WEKA produced decision trees represent a rule-by-rule basis for determining lipid classification based upon the chemically-identifying structural characteristics previously described.

2.3.2 Adaptation of distinct decision trees into comprehensive classification WEKA produced decision trees were parsed into a programmatic classification system written in Ruby. This reference implementation provides complete classification for any given structure specified in 125 chemical input formats supported by OpenBabel 2.3.1.

2.3.3 Evaluation of classification system The reference implementation performance was tested by a comprehensive examination of the entire downloadable LMSD. A classification was considered a miss if it failed to classify properly at the category, class, subclass, or level 4 class levels.

2.3.4 Improvements to existing ontology By virtue of producing an automated classification scheme, we explored the LMSD for areas where the current ontology failed to fully differentiate lipids or for cases where the assigned LMSD classification was erroneous.

2.3.5 Iterate to complete classification Classifier performance was iteratively improved by classification of the entire LMSD using the current chemical features in the decision trees generated by the machine learning program. Any lipids for which the proposed classification showed discrepancies with the LIPID MAPS classification were examined manually to find the source of the discrepancy. In cases where the classifier did not account for necessary chemical features needed to correctly classify the lipid, necessary changes and/or additions were made to the chemical features used for classification. When further investigation showed that the classification proposed by the classifier was superior to the LIPID MAPS classification, those lipids were marked for ontology changes in subsequent analyses. This process was performed several times on the whole LMSD as well as on subsets of lipids to facilitate quick, concentrated testing and refinement.

2.3.6 Evaluation on novel lipids Evaluation of the trained classifier was performed upon an extracted subset of the LipidBank database (Watanabe *et al.*, 2000), consisting of some 1195 molecules, many of which are representative of molecules which are contained in the LMSD. Manual analysis was conducted to determine if these lipids were 1) properly categorized into, and 2) fit within the LMSD ontology.

3 RESULTS

We created software to accurately classify novel lipids according to the LIPID MAPS ontology. Comparative performance upon a subset of the LMSD among all decision trees demonstrated superior performance of the J48 classifier. By careful choice of structural characteristics and machine learning algorithms, we ensured that the resulting decision tree is human understandable and can be used and interpreted manually. The resulting comprehensive classification provides a quality classification by machine learning that retains relevance as a human interpretable toolkit.

3.1 Classifier performance

We generated a classification by 10-fold cross-validation analysis of the LMSD. The comprehensive classification implementation in the Ruby language provides less than 1.2% error across all classification levels. At the category level, we reach 99.98% accuracy when suggested improvements to the existing ontology are followed as outlined in the supplemental materials, as described in Table 1.

There remains a section of higher error, in a set of subclasses which depend upon sugar oligomer length as the differentiator. Currently,

Commented [RT4]: Unfortunately, Vector graphics aren't cooperating with this… I'm instead trying really high DPI rasters.

the SMARTS dependent characteristics are unable to address this complication.

Table 1. Classifier performance for entire, and category slices of the LMSD.

	Number of lipids	Category Level Error counts (%)		Within Category Error counts (%)	
Entire LMSD	36785	6	(0.02%)	429	(1.17%)
Fatty Acyl [FA]	5763	1	(0.02%)	3	(0.05%)
Glycerolipids [GL]	7538	1	(-0.01%)	2	(0.03%)
Sterol Lipids [ST]	2561	2	(-0.08%)	18	(0.70%)
Prenol Lipids [PR]	1193	1	(0.08%)	0	(0.00%)
Sphingolipids [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%)

Representative errors of classifier performance overall, for categories and for representative classes.

3.2 Novel lipid analysis

Evaluation of classified lipids from the LipidBank database demonstrates the capability of this classification to handle novel lipid classes. For

3.3 Ontology modifications

Evaluation of the LIPID MAPS classification exposed ~700 misclassified lipids. These misclassifications are represented within the supplemental materials and demonstrate the difficulty of hand-curating a database the size of the LMSD and the ability of automated classification tools to assist in the classification process.

Many of the misclassifications are due to small structural differences. Lipid LMGP04040006 is classified as a dialkylglycerophosphoglycerol. Upon inspection of the structure, however, it becomes apparent that it contains an acyl group in place 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, which is currently classified as a straight chain fatty acid, is clearly branched, corresponding to the reclassification suggested by our analysis.

Our classifier excelled in correctly 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. From the data provided, our classifier was able to follow the established ontology that branching takes precedence over unsaturation and assign these lipids the correct classification.

New groups in Neutral Glycosphingolipids (LMSP05) are necessary to improve identification and subsequent classification of current and future lipids. In accordance with IUPAC guidelines(REFERENCE?), neutral glycosphingolipids were assigned a group based on their root sugar chain. The root is the first four sugars and their linkages attached to a Ceramide. The IUPAC and LIPID MAPS suggested nine groups (or series, LMSP0501-09). However, there are two sub-subgroups within the Neolacto subgroup that do not fit into it nor any other group. These sub-subgroups (LMSP0505DC-F and LMSP0505DM-N) contain 32 and 16 lipids with two unique roots. We suggest naming these new groups 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 illustrates the relationship to the Gala series (LMSP0509) in that it contains repeated galactose monomers and highlights the terminal glucose monomer. These new ontologies appropriately represent the incorrectly classified lipids in their own ontology.

Commented [RT6]: Galacto-gal-gal-glu ? What is the right name for these?

It is for lipids which contain 3 galactoses and then a terminal glucose.

Commented [RT7R6]: I don't like Galacto-lacto since that emphasizes the gal-gal not the gal-gal-gal-GLUCOSE which seems to be the important part.

Commented [RT5]: This same pattern continues into the DOXX lipids… also, the DPXX lipids are GLCGALGALGAL which doesn't seem to fit the GALGLCNAcGALGLC pattern either! I think we've significantly under 'fixed' this category (LMSP0505)…

3.4 Future Directions

Future work will expand the classification system to classify non-lipids into general categories, and improve upon some existing limitations of the extensive sugar nomenclature within the sphingolipid category. Future work will further evaluate a need for an alternative ontology which enables multiple classifications for a given lipid which would eliminate some of the precedence issues we observed.

ACKNOWLEDGEMENTS

We…

Funding: BYU Institutional Funds, BYU Undergraduate Research Awards

REFERENCES

Danziger,S. et al. (2011) Extraneous factors in judicial decisions. Proc. Natl. Acad. Sci. U. S. A., 108, 6889–92.

Fahy,E. et al. (2005) A comprehensive classification system for lipids. J. Lipid Res., 46, 839–61.

Fahy,E. et al. (2009) Update of the LIPID MAPS comprehensive classification system for lipids. J. Lipid Res., 50 Suppl, S9–14.

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, 925–33.

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, E297–303.

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, L305–11.

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, 3236–45.

Hsu,F.-F. et al. (2007) Structural Characterization of Phosphatidyl-myo-inositol Mannosides from Mycobacterium bovis Bacillus Calmette Guérin by Multiple-Stage Quadrupole Ion-Trap Mass Spectrometry with Electrospray Ionization. I. PIMs and Lyso-PIMs. J. Am. Soc. Mass Spectrom., 18, 466–478.

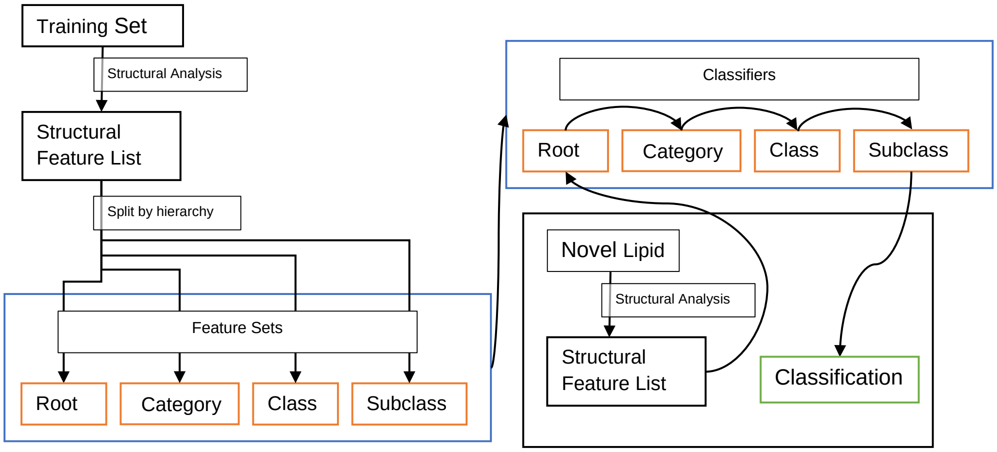
Hsu,F.-F. and Turk,J. (2010) Toward total structural analysis of cardiolipins: multiple-stage linear ion-trap mass spectrometry on the [M - 2H + 3Li]+ ions. J. Am. Soc. Mass Spectrom., 21, 1863–9.

Kangas,L.J. et al. (2012) In silico identification software (ISIS): a machine learning approach to tandem mass spectral identification of lipids. Bioinformatics, 28, 1705–13.

Kind,T. et al. (2013a) LipidBlast in silico tandem mass spectrometry database for lipid identification. Nat. Methods, 10, 755–8.

Kind,T. et al. (2013b) LipidBlast in silico tandem mass spectrometry database for lipid identification. Nat. Methods, 10, 755–8.

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.
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, 1150–7.
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, 1848–58.
Subramaniam,S. et al. (2011) Bioinformatics and systems biology of the lipidome. Chem. Rev., 111, 6452–90.
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.
Wright,S.M. et al. (2000) Altered airway surfactant phospholipid composition and reduced lung function in asthma. J. Appl. Physiol., 89, 1283–92.
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, 302–13.
Yetukuri,L. et al. (2007) Bioinformatics strategies for lipidomics analysis: characterization of obesity related hepatic steatosis. *BMC Syst. Biol.*, **1**, 12.


Training Set
Structural Analysis
Structural Feature List
Split by hierarchy
Feature Sets
Root
Category
Class
Subclass
Classifiers
Root
Category
Class
Subclass
Novel Lipid
Structural Analysis
Structural Feature List
Classification