Database and Ontologies

Automated classification of lipids

Ryan Taylor[1,*], Ryan H. Miller[2], Michael Porter[2], Ryan D. Miller[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: While a majority of the lipid fragmentation software currently available relies upon 'divide-and-conquer' principles to partition the complexity of fragmentations; no automated lipid classification exists.

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

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. The roles of lipids in a cellular context include: controlling the transport in and out of the cell, offering structural stability and flexibility to the cell, providing and storing energy, regulating electron transport and energy metabolism, transmitting information across the membrane, signaling systemic conditions, and delivering signal transduction information as secondary messengers. 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 functionally 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 in nearly every disease of industrialized societies including cardiovascular diseases (WHO, 2011), diabetes (Facts and Diabetes, 2011), obesity (Masters *et al.*, 2013), Alzheimer's (Alzheimer's Association, 2013), arthritis (CDC, 2010), asthma (Akinbami *et al.*, 2011), and cancer (Facts, 2013). In addition to many structural lipid abnormalities (Pietiläinen *et al.*, 2007, 2011; Han *et al.*, 2002; Jia *et al.*, 2007; Guerrera *et al.*, 2009; Fujiwaki *et al.*, 2002; Wright *et al.*, 2000; Fuchs *et al.*, 2005), low abundance lipid hormones such as eicosanoids, steroids and sterols are well-known for the dramatic functions exhibited in a variety of diseases (Han and Cheng, 2005). 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

Proteomics has succeeded in large part due to the relative ease of searching a protein database for matches to experimental fragmentation spectra. This approach works for proteins because the rules for basic protein digestion and fragmentation within a mass spectrometer are understood. Lipidomics will benefit from tools which can perform the analogous task for lipids. However, the lipid alphabet is more structurally diverse than that of proteins, and the fragmentation spectra are far less consistent(Hsu and Turk, 2010; Hsu *et al.*, 2007). Most fragmentation software treat lipids on a class-by-class basis, or in a 'divide-and-conquer' approach (Herzog *et al.*, 2012; Kind *et al.*, 2013; Kangas *et al.*, 2012; Song *et al.*, 2007). Thus, fragmentation prediction relies upon appropriate classification of lipids.

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 pregeneration of structural libraries using a few proto-typical lipids (Kind *et al.*, 2013). Flexible, automated fragmentation requires that lipids be classifiable 'on-the-fly', but no such tool currently exists. Computational classification tools are necessary to efficiently enable extension of current ontologies to novel lipid species.

We present a computational approach to automated classification of any lipid into the existing LIPID MAPS ontology as represented by the LIPID MAPS structural database (LMSD).

2 METHODS

Despite the major role of synthetic pathways within the LIPID MAPS ontology, each lipid species is structurally distinct. The structural characteristics of lipids contain the identifying characteristics required for complete classification.

2.1 Chemical Language

Rubabel, a cheminformatics software suite built upon the OpenBabel library, formed the basis for this analysis by providing a programmatic representation of chemical structures.

Commented [RT1]: Establish: Why do we want to extend the ontology?

Commented [RT2]: This was the argument by Eion Fahy via email…

*To whom correspondence should be addressed.

selected the Simplified Molecular-Input Line-Entry System (Weininger, 1988) (SMILES) as a basis for chemical structure analysis. The SMILES Arbitrary Target Specification, or SMARTS, provides a simple, cross-software method for searching chemical structures for chemically 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, from simple functional group recognitions and atom counts to peptide and glycosidic bond recognitions, to characterize the structural motifs sufficient to classify any given lipid. These are provided as an open-source resource.

2.3 Classification by Machine Learning

~~We employed the premier data mining and machine learning software, WEKA, which is capable of classification, regression, clustering, association rule mining, and attribute selection.~~ WEKA demonstrated the best performance under the J48 decision tree algorithm with (PARAMETERS) and optimized by 10-fold cross-validation. Selection of algorithm attributes was optimized upon the category level ontology and adopted for all subsequent analyses. The product decision trees represent a rule-by-rule basis for determining lipid classification based upon the chemically-identifying structural characteristics previously described. ~~Manual classification is possible from these product decision trees.~~

2.3.1 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 of any format supported by Rubabel.

2.3.2 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 level4 class levels.

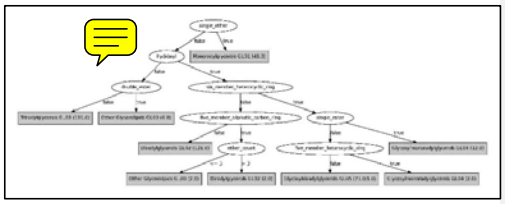
2.3.3 Improvements to existing ontology By virtue of producing an automated classification scheme, we explored the LMSD for areas where the current ontology failed to fully differentiate lipids or for cases where the assigned LMSD classification was erroneous.

3 RESULTS

Analysis of the LMSD provides less than 1% error across all classification levels. At the category level, we reach 99.9% accuracy when suggested improvements to the existing ontology are followed, as outlined in the supplemental materials. The provided software supplies a reference implementation written in the Ruby language suitable for analysis of any chemical structure. The software also includes the reference chemically identifying characteristics and classifications for implementation in other frameworks.

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.

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



Future work can expand the classification system to classify non-lipids into general categories, and improve upon some existing limitations of extensive sugar nomenclature within some lipid subclasses. We hope to collaborate with other research groups in further improving the LIPID MAPS ontology.

Table 2. Classifier performance for selection of ontology.

Level	Name	Size	Error
	Overall	36785	0.13%
Category	Fatty Acyl [FA]	5763	0.3%
Category	GL	7538	0.2%
Category	ST	2561	1.9%
Category	PR	1193	0.0%
Category	SL	1293	0.0%
Category	PK	6744	0.0%
Category	SP	3934	0.0%
Category	GP	7759	0.0%

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

ACKNOWLEDGEMENTS

We…

Funding: BYU Institutional Funds, BYU Undergraduate Research Awards

REFERENCES

Akinbami,L.J. *et al.* (2011) Asthma prevalence, health care use, and mortality: United States, 2005-2009. *Natl. Health Stat. Report.*, 1–14.

Alzheimer's Association (2013) Alzheimer's disease: facts and figures.

CDC (2010) Prevalence of doctor-diagnosed arthritis and arthritis-attributable activity limitation: United States, 2007-2009. *MMWR Morb. Mortal. Wkly. Rep.*, **59**, 1261–1265.

Facts,C. (2013) Cancer Facts & Figures.

Facts,F. and Diabetes,O.N. (2011) National Diabetes Fact Sheet , 2011.

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.

Fujiwaki,T. *et al.* (2002) Application of delayed extraction matrix-assisted laser desorption ionization time-of-flight mass spectrometry for analysis of sphingolipids in cultured skin fibroblasts from sphingolipidosis patients. *Brain Dev.*, **24**, 170–3.

Guerrera,I.C. *et al.* (2009) A novel lipidomic strategy reveals plasma phospholipid signatures associated with respiratory disease severity in cystic fibrosis patients. *PLoS One*, **4**, e7735.

Han,X. *et al.* (2002) Substantial sulfatide deficiency and ceramide elevation in very early Alzheimer's disease: Potential role in disease pathogenesis. *J. Neurochem.*, **82**, 809–818.
Han,X. and Cheng,H. (2005) Characterization and direct quantitation of cerebroside molecular species from lipid extracts by shotgun lipidomics. *J. Lipid Res.*, **46**, 163–175.
Herzog,R. *et al.* (2012) LipidXplorer: a software for consensual cross-platform lipidomics. *PLoS One*, **7**, e29851.
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.
Jia,L. *et al.* (2007) Metabolomic identification of potential phospholipid biomarkers for chronic glomerulonephritis by using high performance liquid chromatography–mass spectrometry. *J. Chromatogr. B*, **860**, 134–140.
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.* (2013) LipidBlast in silico tandem mass spectrometry database for lipid identification. *Nat. Methods*, **10**, 755–8.
Masters,R.K. *et al.* (2013) The impact of obesity on US mortality levels: the importance of age and cohort factors in population estimates. *Am. J. Public Health*, **103**, 1895–901.
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.
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.
Weininger,D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.*, **28**, 31–36.
WHO (2011) Global status report on noncommunicable diseases 2010.
Wright,S.M. *et al.* (2000) Altered airway surfactant phospholipid composition and reduced lung function in asthma. *J. Appl. Physiol.*, **89**, 1283–92.