Pacific Northwest
NATIONAL LABORATORY

PNNL-29388

Full Integration of Lipidomics Data into Multi-OMIC Functional Enrichment

November 2019

Hugh D Mitchell
Jennifer E Kyle

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
email: orders@ntis.gov <https://www.ntis.gov/about>
Online ordering: http://www.ntis.gov

PNNL-29388

Full Integration of Lipidomics Data into Multi-OMIC Functional Enrichment

November 2019

Hugh D Mitchell
Jennifer E Kyle

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Goals

The goals of this project were to 1) use a text-mining approach to greatly expand the database of lipid-protein links, and include directional metabolism (i.e., consumption versus production) annotation to those links, and 2) incorporate the new database into a useful tool capable of true integration of proteomics and lipidomics datasets.

Background

Lipids have many roles critical to living systems including acting as membrane structural components, signaling molecules, and energy sources; however, many of these vital functions are tightly linked to regulated proteins. The ability to integrate lipidomics data with other omics data is essential for fully understanding the lipidome and expanding our knowledge of the mechanisms by which lipids participate in biological responses. To date, the integration of lipidomics data with other complex omics data types at PNNL has been mostly manual and conducted by a single staff scientist. Although the links between many lipids and enzymes are understood and documented, they exist in large part as text (e.g., in publications) instead of in databases. The process of protein-lipid integration from PNNL omic's data sets is therefore highly manual and involves visually inspecting the changed lipids generated from the lipidomics platform, searching for expression of potentially lipid related proteins generated from the proteomics platform against LipidMaps Protein Database and mining the literature, and assessing whether the direction of change of the linked lipid and protein are in sensible agreement. For this project we proposed automating the identification of lipid-protein linkages and creating a tool that integrates lipid and protein data based on these relationships.

Strategies utilized

The following steps were taken to generate the lipid/protein linkage database.

1. Identified lipid-related proteins by downloading list from LipidMaps, with their associated Uniprot description fields. This included proteins from mammals, plants, fungi and bacteria.
2. Identified all verbs (and their related forms) in the description fields, and retained those most likely related to lipid metabolism (e.g. "synthesize", "hydrolyze", "convert", "elongate", etc).
3. Expected sentence context structures were mapped around each of these action term, so that consumed and produced lipid terms could be extracted from the sentence structure surrounding them (Example: <enzymeX> **converts** <lipidterm1> **and** <lipidterm2> **to** <lipidterm3>.) A unique structure was crafted for each term; however, this was not a particularly time-consuming step since many terms had analogous context structures. Some positions (lipidterm1 and -2) are stored as consumed lipids, while others (lipidterm3) are stored as produced.
4. Built scanning function to iterate through protein description fields, scanning for action terms and extracting lipid terms located in target positions.
5. Identified terms that indicated specificity (e.g. "specific", "preference", "especially") and built expected context structure around these as in step 3.
6. Used similar scanning function to extract specificity information to be added to information acquired in step 4.
7. Built lipid term conversion table, allowing lipid terms to be converted to searchable regular expressions (Example: "palmitic acid" -> 16:[0-6])

The following steps were incorporated into the integration tool.

1. Identify all perturbed proteins in the input proteomics dataset that are present as lipid-related proteins in the link database.
2. For all converted lipid terms associated with a detected lipid-related protein, search the dataset lipid for matches. Only search lipids in the appropriate list, i.e. for up-regulated proteins, only search for matches to consumed lipids among down-regulated lipid species, and search for produced lipids among up-

regulated lipid species. The opposite procedure is followed for down-regulated proteins, since diminishing their abundance (rather than increasing it) is expected to reverse the effect of its activity.

3. Produce results by outputting each protein found to have corresponding lipids in the dataset, its direction of change, then a list of detected consumed and/or produced lipids.

Results

We have built the **P**rotein **A**nd **L**ipid **L**inkage for **I**ntegration with **D**irectionality (PALLID) tool, which encompasses the protein/lipid links database and the associated linkage tool. The PALLID database currently encompasses 1330 lipid-related proteins linked to 302 lipid-related terms. The linkage tool is currently limited to text-based output which is sent to a text file organized with a multi-level indent structure. Applying PALLID to a dataset of lipidomics and proteomics from infection with the Ebola virus allowed linkage of 215 specific lipids to simultaneously expressed and directionally coherent lipids, out of 275 of the total lipids detected as changing in the experiment (78% linkage rate). Independent manual spot checks of linkages verified that the linkage tool works according to design expectations. By comparison, linking a separate dataset of 374 lipids to associated proteins using a current state-of-the-art tool (Omicnet.ca) forced collapsing lipids to their respective 7 classes, then linking classes to proteins, with a resulting linkage rate of 7/374 or <2%.

Impact

We plan to utilize ongoing funding efforts (see below) to 1) incorporate PALLID into an R package, 2) publish an Applications Note in the journal Bioinformatics, and 3) publish a biology-focused paper which fully showcases PALLID's integrative advantages.

We have submitted an EMSL Capability Development proposal (PI Lisa Bramer) which integrates PALLID into a larger proteomics, lipidomics and metabolomics processing, integration and visualization tool in a user-friendly platform. Furthermore, we are preparing an NIH R21 proposal (RFA PA-19-068, PI Jennifer Kyle) with a submission date of Oct. 16, 2019.

Acknowledgement

This research was supported by the Earth and Biological Sciences Directorate (EBSD) Mission Seed, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354
1-888-375-PNNL (7665)

www.pnnl.gov