VKMZ: Identifying and Visualizing Metabolites through Galaxy

Mark Esler[1], Stephen Brockman[1], Arthur Eschenlauer[1,2], Timothy Griffin[2], Adrian Hegeman[1]

[1] Dept. of Horticultural Science and Microbial and Plant Genomics Institute, University of Minnesota—Twin Cities, [2] Dept. of Biochemistry, Molecular Biology, and Biophysics, University of Minnesota—Twin Cities

Overview

- Predicts the molecular formulas of features in mass spectrometry data
- Plots predicted molecules on a van Krevelen diagrams (VKD): a 2D scatter chart plotting molecules based on elemental ratios [0]
- Creates and interactive VKD web page
- VKMZ is a Galaxy tool [1]
- XCMS [2][3] or tabular files can be used as input

Introduction

Liquid chromatography-mass spectrometry (LC-MS) data is often visualized as a mass chromatogram. Mass chromatograms can convey the constituents of a complex mixture, such as untargeted plant metabolomics, but requires close examination. To briefly convey the constituent metabolites in high resolution LC-MS data researchers have created van Kreveln Diagram's from MS data [4]. A van Krevelen diagram (VKD) [0] plots molecules on a 2D scatter plot based on each molecule's hydrogen to carbon ratio (H:C) against it's oxygen to carbon ratio (O:C). Originally the VKD was developed to plot the burning potential of petroleum oils as oils with similar properties have similar elemental ratios and cluster together. Classes of metabolites also cluster together on a VKD, sunce they share molecular structures and elemental ratios. By identifying metabolites in LC-MS data and plotting them to a VKD the molecular composition of the data can be briefly conveyed.

This tool is a branch of OpenVanKrevelen [4]. VKMZ is intended to run on The Galaxy Project web platform [1]. Galaxy enables researchers to create shareable and reproducible data-processing workflows using command line driven software. Metabolomics researchers may use XCMS as their feature processing tool [2][3] which is incorporated in the Worflow4Metabolomics variant of Galaxy [6]. VKMZ can use XCMS as input data and can be integrated into a XCMS Galaxy workflows. Alternatively, VKMZ can be ran as a standalone Python tool or use tabular data as input.

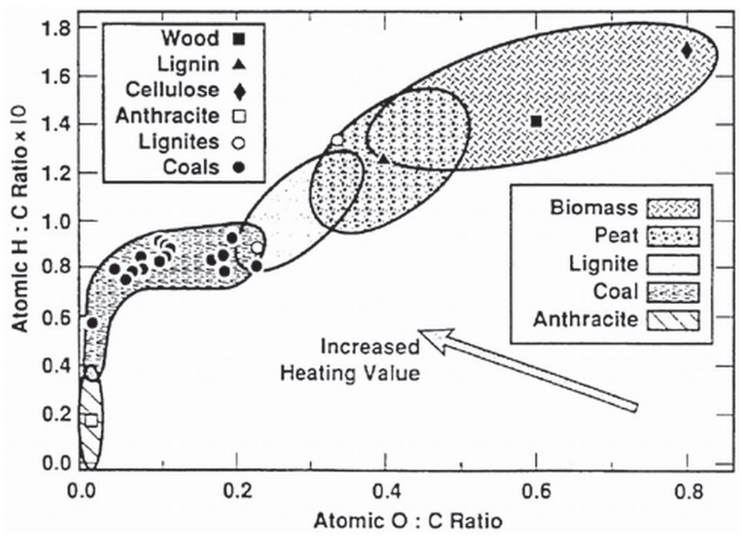

Figure 1. VKD of solid fuels [5]

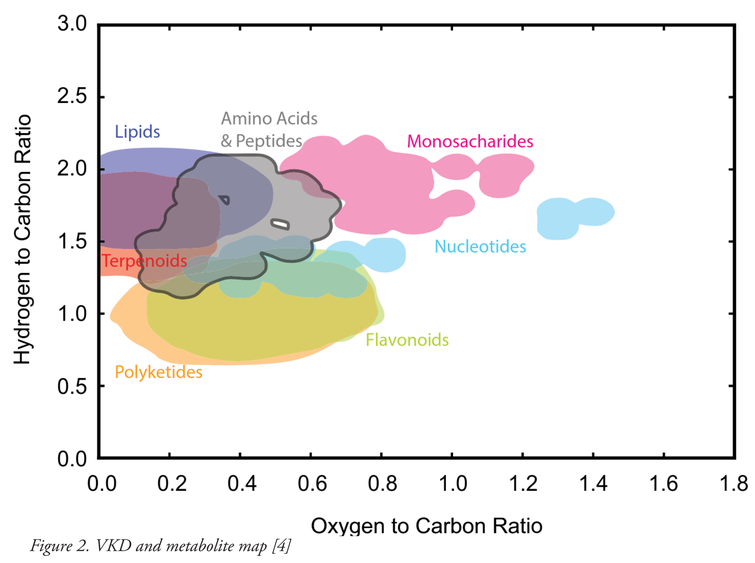

Figure 2. VKD and metabolite map [4]

Methods

Data Input

VKMZ can read either XCMS data or tabular data as input. If using XCMS data, VKMZ expects to read the data matrix, sample metadata, and variable metadata tabular files generated by XCMS. Tabular input files require a specific format. For each feature in the data, VKMZ stores metadata on sample data, polarity, observed mass, retention time, and intensity.

Prediction Generation

Each feature has an observed charged mass and a polarity. By default, charged masses are converted to a neutral mass by adding or removing the mass of a proton based on the feature's polarity. Input polarity information can be globally overridden if necessary. VKMZ compares each feature's neutral mass to a dictionary* of known mass-formulas. A prediction is made when a mass in the dictionary is within a mass error of a feature's neutral mass. The dictionary is sorted by mass so that a binary search can make the first prediction. If no prediction is made the feature is removed from the dataset. If a prediction is made, dictionary indexes adjacent to the initial prediction are checked. Predictions are made until adjacent indexes mass-keys are beyond the mass error range. All predictions are then sorted by lowest absolute delta (mass difference between the predicted and neutral mass). Optionally, feature's with multiple predictions can be removed from the output. All predictions for a feautre are saved as a single list in the feature's metadata. For each prediction the prediction's neutral mass, formula, and delta are saved. Elemental ratios between H:C, O:C. and nitrogen to carbon (N:C) are calculated for the prediction with the smallest absolute delta and saved to the formula's metadata.

Included with VKMZ are four dictionaries from the Biological Magnetic Resonance Bank [7]. Each dictionary contains heuristically generated molecular mass-formula pairs. These four dictionaries differ in their isotopic labeling: monoisotopic, carbon labeled, nitrogen labeled, and carbon & nitrogen labeled. Alternatively, a custom dictionary can be used.

** VKMZ generates a dictionary from two lists since, Python dictionaries are non-indexable.*

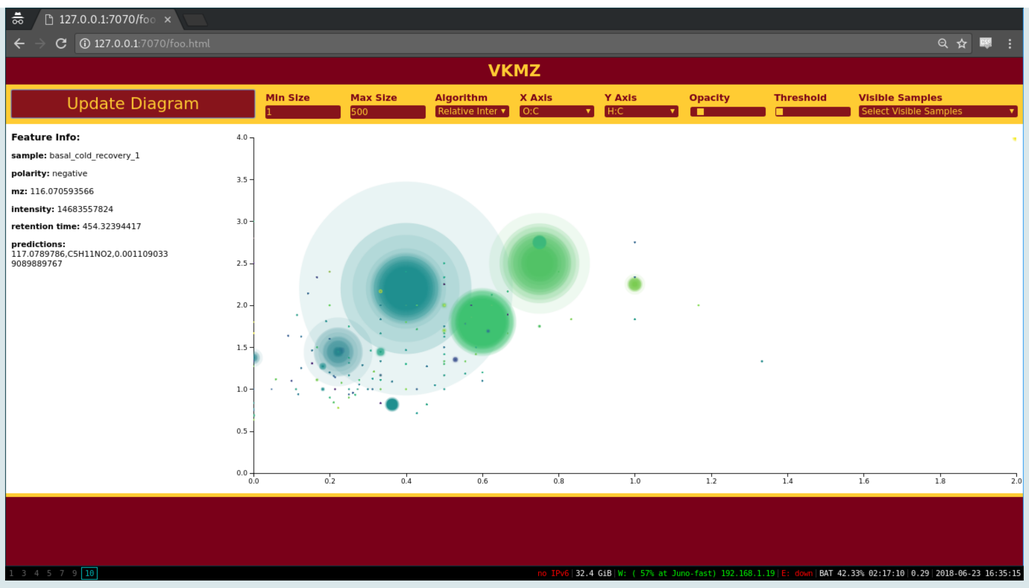

Figure 3. An HTML web page generated by VKMZ

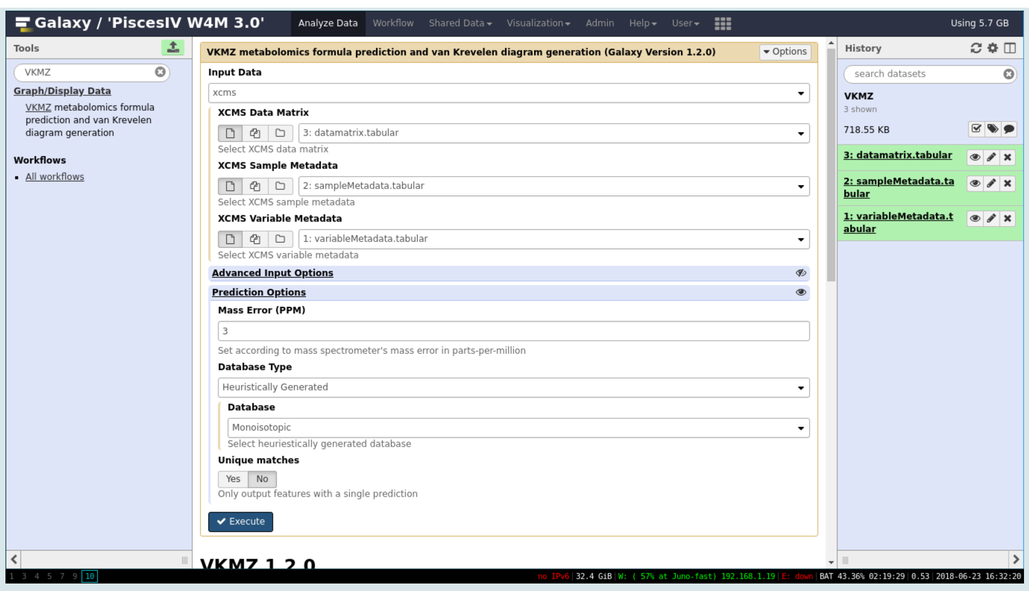

Figure 4. VKMZ Galaxy Tool

Results

Tabular Output

Each row of tabular data represents a single feature. The columns represent types of metadata. In order, metadata types and column names are: sample (sample_id), polarity (polarity), observed mass (mz), retention time (rt), intensity (intensity), list of predictions (predictions), H:C (H:C), O:C (O:C), and N:C (N:C). Predictions are sorted by intensity. Tabular output can be use to make graphs for publication.

HTML Output

Predictions and all metadata are saved as JSON in the output web page and graphed with the D3 [8] JavaScript plotting library. D3 creates a SVG graph and overlays two Canvas layers for plotting features. Canvas is used for plotting instead of SVG since, SVG can be demanding on a browser when plotting many points. One canvas layer contains visible circle symbols representing features and uses elemental ratios as coordinates. The second layer is a copy of the first, but each circle is a random, unique, color which is non-visible. The RGB color values of the invisble canvas layer are used as dictionary keys to map corresponding features on mouse click [9]. When a user select a feature's symbol metadata is displayed in a sidebar. The web page uses responsive design through Basscss [10].

The web page is designed to explore data. By default, a traditional VKD is created with O:C on the x-axis and O:H on the y-axis. Users can set the axis to any combination of O:C, H:C, and N:C. Symbol's color is dependent on the feature's retention time and mapped to the Viridis color palette. An opacity slider sets the opacity of symbols. Minimum and maximum symbol size can be set. Users can choose the symbol sizing algorithms. VKMZ includes three symbol sizing algorithms: uniform where all prediction symbols are set to the maximum size, relative intensity where each prediction's symbol is relative to the feature's intensity and the maximum intensity in the dataset multiplied by the maximum size, and log relative intensity which is identical to relative intensity except, log value of feature intensity and maximum intensity are used. An intensity threshold slider removes predictions which are below a given intensity. A dropdown of checkboxes for each samples sets which samples are plotted.

Acknowlegement

This work was made possible by two NSF grants: A unified Galaxy-based platform for multi-omic data analysis and informatics (award ID 1458524) and Improving Dynamic Metabolic Flux Analysis for the Discovery of Molecular Determinants of Plant Phenotypes (award ID 1238812).

Further Information

https://github.com/HegemanLab/VKMZ

Citations:

[0] van Krevelen, Fuel Vol. 29 1950 pg 269-228
[1] The Galaxy Project, https://galaxyproject.org/
[2] xcms for Galaxy, https://github.com/workflow4metabolomics/xcms
[3] Smith et. al, 10.1021/ac051437y
[4] Brockman et. al, 10.1007/s11306-018-1343-y
[5] Loo and Koppejay, 978-1-84407-249-1
[6] Workflow4Metabolomics, http://workflow4metabolomics.org
[7] Biological Magnetic Resonance Data Bank, http://www.bmrb.wisc.edu/
[8] D3, https://d3js.org/
[9] https://bl.ocks.org/veltman/f539d97e922b918d47e2b2d1a8bcd2dd
[10] Basscss, http://basscss.com/