NTA WebApp user guide v0.1 (Note: This is an incomplete draft version)

Changelog:

V0.1 – 11/3/2023 – Basic framework of user guide created to inclusion on WebApp for testing purposes.

1. Overview

The WebApp is an automated QA/QC and ID annotation algorithm intended for use on mass spectral data produced for non-targeted analysis (NTA). The WebApp represents a standardized methodology for processing NTA data that is both transparent and reproducible.

2. Table of Contents

3. Background/Introduction

The exposome has emerged as a new paradigm in public health, encompassing the sum of exposures an individual experiences over their lifetime. To that end, targeted analyses of known compounds is an insufficient tool for exposomic studies. Emerging tools such as high-resolution mass spectrometry (HRMS) with Orbitrap and quadrupole time-of-flight (Q-TOF) mass analyzers generate mass spectral data with sufficient resolution and sensitivity to enable compound identification in downstream analyses independent of chemical target lists. These analytical advancements have allowed researchers to pursue non-targeted analysis (NTA) methods that are far more suited to exposomics research where previous targeted methods fall short.

4. MS1 Module

4.1 Input Data

4.2 Input Parameters

4.3 Methods

*4.3.1 Ionization Mode Check*

The Web App uses the ionization mode for data collection to correct feature masses (e.g., proton loss or gain during ionization) and determine possible adducts (see Table 1). Ionization mode is no longer a required column in the user’s data, as the Web App now assigns an ‘Ionization\_Mode’ value to the user’s data based on the input field used (see Input Figure in Section 4.1).

*4.3.2 Filter Void Volume*

Void volume is the volume of solvent contained within your chromatography system. In a chromatogram, void volume is manifested as features that occur in the chromatogram prior to the injected sample making it to the detector. These features (if any) aren’t present in the injected sample – they are coming from the void volume. The simplest way to remove features introduced into a chromatogram by void volume is to identify how long it takes for the sample to reach the detector (typically a solvent peak is used as the indicator) and remove any features from the data that have a retention time less than that value.

The Web App handles void volume in precisely this manner (Figure 2). A user-defined retention time is applied to the input data, and all features with a retention time less than that value are removed. This protects the user from falsely identifying a feature as present in a sample. Features that are removed from the “Cleaned\_feature\_results\_full” and “Cleaned\_feature\_results\_reduced” pages in this way are noted in the “Filter\_documentation” page with a ‘V’.

*4.3.3 Drop Duplicates*

One of the benefits of high-resolution mass spectrometry is the resolving power, or ability to differentiate features with very small mass differences into distinct peaks. This allows for researchers to better identify features in a complex matrix (e.g., wastewater or soil) or quantify isotopic variants of a chemical compound. However, in the latter case, high resolution creates individual peaks for chemical compounds that are structurally the same. Many vendor softwares have functionality for removing isotopic duplicates, but there are also cases where artifacts are produced during peak selection. It is important to remove duplicate features in order to get an accurate count of features in a sample.

In order to drop duplicates the Web App focuses on two feature characteristics, the mass and retention time. The Web App calculates matrices from the difference between each features’ masses and retention times, and then identifies features where both the mass difference and retention time difference are below thresholds set by the user (see Figure 3). In this way, features that are duplicates of one another are flagged and removed from the data. Features that are removed from the “Cleaned\_feature\_results\_full” and “Cleaned\_feature\_results\_reduced” output pages during the de-duplication step are noted in the “Filter\_documentation” page with a ‘D’.

*4.3.4 Calculate Statistics*

Each feature that passes the “Void Volume” and “Duplicate” filters has the following statistics calculated for each sample group: mean (

x−�-

), median, standard deviation (σ), coefficient of variation (CV; calculated as

x−/σ�-/�

), number of abundances (per sample replicate group), and replicate percent. Sample groups are identified in the Web App by comparing adjacent column names and grouping columns that are at most one character different – for sample replicates to be grouped together they must be adjacent, and their names must be off by one character or less (e.g., ‘My\_Sample\_a’ and ‘My\_Sample\_b’).

Following the calculation of feature statistics, features tested for two possible cases: another feature is an adduct of a given feature, or a given feature is an adduct of another feature. The Web App determines a list of possible adducts based on the ionization mode of the data (see Table 1). For each possible adduct, the mass difference (given in Table 1) is added to (i.e., is an adduct) and subtracted from (i.e., has an adduct) the feature mass difference matrix – this is identical to the “Drop Duplicates” calculation with the added step of addition/subtraction of an adduct mass. Two features that have a mass difference of ~0 (a user-defined cutoff value is used; the default cutoff is 0.005 Da) after the adduct mass is considered are determined to have or be that given adduct. The Web App identifies which features are related by the adduct mass difference and annotates those features with the adduct name, and feature ID of the related feature.

The output includes the statistics columns for each sample group, as well as three columns with adduct information, including ‘Has\_Adduct\_or\_Loss’, ‘Is\_Adduct\_or\_Loss’, ‘Adduct\_or\_Loss\_Info’. The first two of the columns contain either a 0 or a 1, indicating ‘no’ or ‘yes’, in the respective column. The third column contains an identifying number of the related feature and the name of the adduct the feature is related by (e.g., Na). No features are removed during the statistics calculation and adduct identification steps.

|  |  |  |
| --- | --- | --- |
| **Possible Adduct** | **Possible Adduct Mass Difference (Da)** | **Ionization Mode** |
| Na | 22.989218 | positive |
| K | 38.963158 | positive |
| NH4 | 18.033823 | positive |
| Cl | 34.969402 | negative |
| Br | 78.918885 | negative |
| HCO2 | 44.998201 | negative |
| CH3CO2 | 59.013851 | negative |
| CF3CO2 | 112.985586 | negative |
| H2O | 18.010565 | neutral loss |
| CO2 | 43.989829 | neutral loss |

*4.3.5 Check Tracers*

Tracers are known compounds, spiked into samples by the user to evaluate method/instrument response (e.g., retention time, response factor, and/or coefficient of variation). The Web App provides functionality for optionally identifying tracers if the user chooses to input a tracers file alongside their data. The Web App provides visualizations that help the user to evaluate proper QC thresholds (e.g., CV value) that are described in greater detail below.

If the user uploads a tracer file, the Web App looks for features in the user input data that match a compound listed in the tracer file. Possible tracers are identified by comparing the ‘Mass’ and ‘Retention\_Time’ of features with the known values given in the tracer file. Features that are within a user-defined mass and retention time cutoff of a known tracer are saved into a separate table and formatted for export as the ‘Tracer\_Sample\_Results’ and ‘Tracers\_Summary’ sheets. This provides useful information for the user (e.g., determining matrix effects on feature CV values), and stores the possible tracer features separate from the rest of the data prior to blank subtraction during the ‘Clean Features’ step.

*4.3.6 Clean Features*

Following the calculation of feature statistics and optional identification of tracers, the data undergo multiple QC filters that remove features from the final output pages, “Cleaned\_feature\_results\_full” and “Cleaned\_feature\_results\_reduced”. These filters include user-determined values for minimum number of sample replicates and CV threshold, as well as a minimum detection limit (MDL) calculated as the mean abundance of a feature in the blank replicates plus three times the standard deviation. For each of these filters, sample occurrences that are flagged and/or removed are documented in the ‘Filter\_documentation’ output sheet with filter-specific symbology (see Table 2).

|  |  |
| --- | --- |
| **Filter/Threshold** | **Documentation Code** |
| Void Volume | V |
| Duplicate | D |
| Minimum # Replicates | R |
| Maximum CV Threshold | CV |
| Minimum Detection Limit (MDL) | ND |

*4.3.7 Create Flags*

*4.3.8 Combine Modes*

*4.3.9 Search Dashboard*

*4.3.10 Store Data to MongoDB*