REPRODUCING THE MOUSE & MM14 OPTIMIZATION AND FIGURES

CHRISTOPHER CONLEY, JOHN PRINCE

1. Introduction

This document primarily describes the optimization scripts for Massifquant & CentWave IT detection applied to the MOUSE and MM14 datasets as described in Massifquant's publication. It provides a commentary and basic example on how to run and verify the reported results and figures.

2. Software Requirements

The following section is easily reproducible for those running a Unix-based OS with a recent version of R installed. Be sure to install CRAN packages: ggplot2, reshape2, gridExtra; and Bioconductor packages: xcms, faahKO. Some of the evaluation was written in Matlab and reproduction will require a software license (greater than 2012). For relevant code and/or difficulty in following tutorial, please contact cjconley 'at' ucdavis 'dot' edu.

3. Reproducing the Greedy Search Optimization

3.1. **optimization routine and corresponding scripts.** The greedy-search optimization has the following general outline:

(1) Run either Massifquant or CentWave varying at most 2 parameters at a given time.
(2) Write the reported ITs to file in their respective parameter setting directory.
(3) Evaluate the reported ITs by cross-referencing them to the respective annotation ITs and computing various performance metrics (e.g. f-score, sample specificity and specificity).
(4) Update the current parameters to those that gave the best performance through the f-score. Go back to (1) until satisfied or can no longer improve the f-score.

The file "search-best-params.r" contains various R wrapper functions that encode steps (1,2). The IT detection results are written to a specific directory such as "mouse/cw-params/prefilter10", which is an indexed directory based on a prefilter variation. An example of these functions being called with different parameter variations is in the script "reproduce-cw-mouse.r", which applies parameter variations of CentWave on the MOUSE subsample.

The file "optimizeCentWave.m" is the core function behind steps (3,4). It is a bit of a misnomer because it can optimize any IT detection output from XCMS, including

Massifquant. It calls another function "storeCentWaveFeats.m" to read in the IT detection results (step 3); and calls another function "evalAreaFinder.m", which computes the f-score and sample sensitivity/specificity for a given parameter setting. An example of steps (3,4) in action is found in the Matlab script "mouse/optmousecw.m".

3.2. **A step by step guide to running the optimization.** First inspect that the following directories are empty and no IT detection results have yet been generated (steps 1,2).

```
#MOUSE
gate-reproducible/mouse/cw-params
gate-reproducible/mouse/mq-params
gate-reproducible/mouse/opt-results
#MM14
gate-reproducible/mm14/cw-params
gate-reproducible/mm14/mq-params
gate-reproducible/mm14/opt-results
```

The optimization results may be reproduced by the following commands on the command line. This will likely take about 5 hours.

```
$tar -xzf gate-reproduce.tgz  #unpack the file in prefered directory
$cd gate-reproducible/  #move to the unpacked directory
$ pwd #this is your working directory
/home/great-scott/gate-reproducible
$nohup ./reproduce-optimization.sh > reproduce.out & #reproduce the results
```

Check the status of the optimization by this command:

```
$cat reproduce.out
```

Every parameter set evaluated is recorded in a corresponding directory and the optimization values are recorded in the base directory of that optimization step. For example, to see the best results for *centWave* under the peakwidth category for MM14, simply open the file "gate-reproducible/mm14/cw-mm14-opt.out". You should see within the file the following reports on best F-score (fmax), recall (rmax), precision (pmax), and sensitivity (snmax) and specificity (spmax), with the corresponding directory of parameters used to generate that result. In this case, the best peak width values are recorded in "gate-reproducible/mm14/cw-params/pkwidth1/parameters.txt". The output of calling the function "optimizeCentWave.m" appears below for a particular variation set such as peak width. You may consult the file "out-optmm14cq.out' for more details.

```
 pkwidth-cw-mm14.mat
fmax =           #fscore
    0.9438
fmax_idx =   #the index of the directory with best parameters
     1
rmax =          #IT recall or sensitivity
```

```
    0.9130
pmax =          #IT precision
    0.9767
snmax =         #sample sensitivity
    0.8995
spmax =         #sample specificity
    0.9970
```

.

Within the matlab console type

```
>>load  /mm14/opt-results/pkwidth-cw-mm14.mat
```

to view the fscores, recall, precision, and proportional area correct per feature for all the peakwidth values evaluated (see Figure **??**). Alternatively, visit the recorded F-score results in

```
gate-reproducible/mm14/cw-params/pkwidth/_fscore_fine.txt
```

.

In keeping with this example, all other optimization may be verified under different conditions. To navigate other conditions and directories, files or procedures containing "cw" correspond to *centWave* results; those containing "mq" correspond to the direct output of massifquant, the c++ standalone implementation of Kalman Filter approach. Associated with "mq" files are "kx" files, which are results produced from massifquant and reported through *XCMS*. All values reported in the paper corresponding to Massifquant come from "kx" files.

4. Evaluation of Performance by Feature Type

It is possible to look for potential confounding variables like feature intensity, ppm (feature variance), and length by running this script.

```
$pwd
path-to-gate-reproducible/
$nohup matlab <runSimpsons.m > simpsons.out &
```

The corresponding figure of interest will be called quantiles-metrics-updated.eps and is generated in the next section.

5. Reproducing Figures for Massifquant Publication

Check that the figures/ directory is empty. Once the optimization results have been generated, please now recreate the figures.

```
$nohup ./reproduce-figures.sh  > figures.out &
$#wait 10 seconds for it to finish
$cd figures/
$ls
```