ProbMetab: an R package for Bayesian probabilistic annotation of LC-MS based metabolomics

Ricardo R. Silva[1*] *et al.* (2013)

[1]LabPIB, DCM-FFCLRP-USP, Universidade de São Paulo, Brazil

* rsilvabioinfo@gmail.com

Abstract

This document illustrates the usage of R package *ProbMetab* in two real metabolomics datasets. The aim is to perform a complete analysis work-flow, from spectra preprocessing to network visualization, to show the package integration capabilities with upstream tools (CAMERA and mzmatch.R in this example) and with downstream tools (Cytoscape and DiffCor). In order to demonstrate the analysis flow usefulness we used publicly available data from *Trypanosoma brucei*, causative agent of sleeping sickness and data from *Sacarum oficinarum* (sugarcane), an important source of 1st generation biofuels. The *Trypanosoma brucei* dataset was chosen to illustrate the annotation procedure since the published experiment has a set of compounds identified with the aid of standard compounds, being specially suited for validation purposes. The sugarcane original dataset was chosen to show how results provided by *ProbMetab* can be used to study metabolism changes.

External files refereed in the text:

filter_comp.xls - Table showing the comparison of xcms – CAMERA/mzMatch to previously published as identified compounds (level 1 identification [1]) IDEOM table.

inf_incorporation.xls – Table showing how each model component ranks the candidate compounds for previously identified compounds.

classByReactions.xls – Table showing that the subset of reactions that were overlaid with reactions agrees with probability ranking.

mzMatch_outputPOS.peakml – PeakML file for positive acquisition mode for mzMatch integration example.

mzMatch_outputNEG.peakml – PeakML file for negative acquisition mode for mzMatch integration example.

probmetab-case1-box00.Rdata – xcms pre-processing objects for *T. brucei* dataset.

probmetab-case1-box01.Rdata – CAMERA pre-processing objects for *T. brucei* dataset.

probmetab-case1-box02.Rdata – *ProbMetab* objects necessary for graph drawing for *T. brucei* dataset.

probmetab-case2-box00.Rdata - xcms and CAMERA pre-processing objects for *S. oficinarum* dataset.

probmetab-case2-box01.Rdata - *ProbMetab* objects necessary for graph drawing for *S. oficinarum* dataset.

ProbMetab Case Study 1: Metabolomics dataset with internal standard compounds illustrates annotation performance

Motivation

In order to demonstrate the analysis flow usefulness we used the publicly available data from *Trypanosoma brucei*, causative agent of sleeping sickness, to illustrate the annotation procedure. The published experiment has a set of compounds identified with the aid of standard compounds, being specially suited for validation purposes. *T. brucei* cell cultures extracts were analyzed with RSLC3000 UHPLC (Thermo Scientific) chromatography platform, with a ZIC-HILIC (SeQuant) column coupled to a Exactive Orbitrap mass spectrometer (Thermo Scientific), operating in positive and negative acquisition modes. The description of these experiments, acquisition and data availability can be found in [2], and the data at:

http://sourceforge.net/projects/mzmatch/files/Ideom/IDEOM_demo_mzXMLfiles.zip/download

The set of routines implemented in our ProbMetab package can be divided in three steps: 1) Ion annotation extraction and data base matching; 2) Probability modeling and estimation and 3) Comprehensive output representation. The three following sections show how ProbMetab approaches these points. It is worth notice that step 1 (pre-processing) and step 3 (post-processing) are performed using third party packages and ProbMetab's role is restricted to integration only in those steps.

Data Analysis

Ion annotation extraction and database matching

ProbMetab assumes that the most basic preprocessing steps to transform raw data into open interchangeable format (e.g. .mzXML, .cdf, etc) were already achieved. Useful guidance to perform that can be found elsewhere [3].

ProbMetab assumes peak detection, retention time correction and peak grouping [4, 5] in order to perform mass peak to compound assignment. Starting from xcms compatible raw data files, that can be downloaded from the mzMatch [6] project page as cited above, one starts an R section as:

```
# load required libraries
library(ProbMetab)
library(xcms)
library(CAMERA)

# set nslaves for the number of available cores of your machine
# Follow xcms vignette to understand data structure
#
http://www.bioconductor.org/packages/release/bioc/vignettes/xcms/inst/doc/xcmsPreprocess.pdf

nslaves <- 4

# positive acquisition mode, directory: 'POS/'
xset <- xcmsSet(
         "POS/", method='centWave', ppm=2, peakwidth=c(10,50),
        snthresh=3, prefilter=c(3,100), integrate=1, mzdiff=-0.00005,
           verbose.columns=TRUE,fitgauss=FALSE, nSlaves=nslaves
        )

# negative mode
xset2 <- xcmsSet(
         "NEG/", method='centWave', ppm=2, peakwidth=c(10,50),
          snthresh=3, prefilter=c(3,100), integrate=1, mzdiff=-0.00005,
          verbose.columns=TRUE,fitgauss=FALSE, nSlaves=nslaves
         )

# load("probmetab-case1-box00.RData") # run to avoid deal with raw data and go directly to
examples


# align retention times across samples, grouping and integration
xsetP <- retcor(xset, method='obiwarp', plottype="none", profStep=1) #positive mode
xsetPnofill <- group(xsetP, bw=5,  mzwid=0.015)
xsetP <- fillPeaks(xsetPnofill)

xsetN <- retcor(xset2, method='obiwarp', plottype="none", profStep=1) #negative mode
xsetNnofill <- group(xsetN, bw=5,  mzwid=0.015)
xsetN <- fillPeaks(xsetNnofill)

#save(list=ls(all=TRUE), file="probmetab-case1-box01.RData")
 #run to save intermediary steps
```

The parameters used above were set for data obtained from orbitrap mass spectrometer [7]:

http://www.nature.com/nprot/journal/v7/n3/fig_tab/nprot.2011.454_T1.html

Once a complete list of mass peaks is selected, we now perform a complexity reduction step, in order to filter each ion redundant forms. This step can be accomplished in different ways within R environment, among them through *Astream* [8], *CAMERA* and *mzMatch.R* libraries. In the present illustration we use CAMERA and mzMatch:

```
# standard CAMERA processing
an <- xsAnnotate(xsetP)
```

```
an <- groupFWHM(an, perfwhm = 0.6)
an <- findIsotopes(an, mzabs = 0.01)
an <- groupCorr(an, cor_eic_th = 0.75)
anP <- findAdducts(an, polarity="positive")

an <- xsAnnotate(xsetN)
an <- groupFWHM(an, perfwhm = 0.6)
an <- findIsotopes(an, mzabs = 0.01)
an <- groupCorr(an, cor_eic_th = 0.75)
anN <- findAdducts(an, polarity="negative")

# load("probmetab-case1-box01.RData") # run to skip the previous block

# combine positive and negative acquisition modes, keeping track of individual modes.
# It is possible to combine the acquisition modes setting positive or negative
# as reference input table.
# comb 1
camAnot <- combinexsAnnos(anP, anN)
camAnot <- combineMolIon(peaklist=camAnot, cameraobj=anP, polarity="pos")

# Extract and format a set of non redundant putative molecular ions from CAMERA annotation
# with ProbMetab
# comb 2
ionAnnotP <- get.annot(anP)
ionAnnotN <- get.annot(anN, polarity="negative")

# number of isotopic peaks and non redundant putative molecules
sum(ionAnnotP$molIon[,3]==1)
sum(ionAnnotP$molIon[,3]==0)
```

Once the initial annotation for different forms of the same ion (adducts and isotopes), is defined, one can seek for a non-redundant set of putative molecules (after charge and possible adduct correction) for further inference of compound identity. The diversity of fragments and adducts formed during the ionization process adds high complexity to compound annotation [9]. Experience shows that standard mass rules for adduct search may lose peaks, and specific rule tables must be setup for a given experimental condition. In order to address this issue, a flexible workflow, which allows users to integrate different methods, would improve true molecular ions recovery.

A standard format definition for an ion annotation table would allow one to obtain it from different upstream tools. The ion annotation table has the following core information: exact mass of putative molecule with experimental error; isotopic pattern associated; adduct form associated, and the original reference to raw data. Our current implementation extracts the ion annotation from CAMERA objects. Following this format one can integrate datasets built with other tools into the proposed downstream analysis. As an example of integration we highlight how the mzMatch PeakML files can be added to the downstream workflow, building on mzMatch methods to write xcms objects.

```
# This is the file automatically generated by IDEOM [10]: http://mzmatch.sourceforge.net/
# that was used to produce the analysis of filter_comp.xls table.
#
```

```
# Below also follows how to integrate the mzMatch analysis to ProbMetab.
# The mzMatch package can be downloaded at
# http://mzmatch.sourceforge.net/tutorial.mzmatch.r.advanced.php
# Please set the working directory to where your files are

setwd("mzXMLfiles/POS")
rawfiles <- dir (full.names=TRUE,pattern="\\.mzXML*",recursive=TRUE)
outputfiles <- paste(sub(".mzXML*","",rawfiles),".peakml",sep="")

for (i in 1:length(rawfiles)){
    xset <- xcmsSet(rawfiles[i], method='centWave', ppm=2, peakwidth=c(5,100),
    snthresh=3, prefilter=c(3,1000), integrate=1, mzdiff=0.001,
    verbose.columns=TRUE, fitgauss=FALSE, nSlaves=2

    )
    PeakML.xcms.write.SingleMeasurement(xset=xset,filename=outputfiles[i],
        ionisation="negative",addscans=2,
        writeRejected=FALSE,ApodisationFilter=TRUE
    )
}

library (mzmatch.R)
mzmatch.init (4000)
MainClasses <- dir ()
dir.create ("combined_RSD_filtered")
dir.create ("combined_RSD_rejected")
dir.create ("combined")

for (i in 1:length(MainClasses)){
    FILESf <- dir (MainClasses[i],full.names=TRUE,pattern="\\.peakml$",recursive=TRUE)
    OUTPUTf <- paste ("combined/",MainClasses[i],".peakml",sep="")
    if(length(FILESf)>0){
        mzmatch.ipeak.Combine(i=paste(FILESf,collapse=","),
                v=T,rtwindow=30,o=OUTPUTf,combination="set",
                ppm=5,label=paste(MainClasses[i],sep="")
        )
        RSDf <- paste ("combined_RSD_filtered/",MainClasses[i],".peakml",sep="")
        REJf <- paste ("combined_RSD_rejected/",MainClasses[i],".peakml",sep="")

        if(length(FILESf)>1){
            mzmatch.ipeak.filter.RSDFilter(i=OUTPUTf,o=RSDf,rejected=REJf,
                rsd=0.8,v=T
            )
        else{
            file.copy(OUTPUTf,RSDf)
        }
    }
}


INPUTDIR <- "combined_RSD_filtered"
FILESf <- dir (INPUTDIR,full.names=TRUE,pattern="\\.peakml$")
mzmatch.ipeak.Combine(i=paste(FILESf,collapse=","),v=T,rtwindow=30,o="combined.peakml",combi
nation="set",ppm=5)
mzmatch.ipeak.filter.NoiseFilter(i="combined.peakml",o="combined_noisef.peakml",v=T,codadw=0
.8)
mzmatch.ipeak.filter.SimpleFilter(i="combined_noisef.peakml",     o="combined_sfdet.peakml",
mindetections=3, v=T)
mzmatch.ipeak.filter.SimpleFilter(i="combined_sfdet.peakml",o="combined_highintensity.peakml
", minintensity=1000, v=T)
PeakML.GapFiller(filename = "combined_highintensity.peakml", ionisation = "detect", Rawpath
= NULL, outputfile = "highintensity_gapfilled.peakml", ppm = 0, rtwin = 0)
mzmatch.ipeak.sort.RelatedPeak(i="highintensity_gapfilled.peakml",v=T,o="mzMatch_output.peak
ml",basepeaks="mzMatch_basepeaks.peakml",ppm=3,rtwindow=6)
annot <- paste("relation.id,relation.ship,codadw,charge")
```

```
mzmatch.ipeak.convert.ConvertToText(i="mzMatch_output.peakml",o="mzMATCHoutput.txt",v=T,anno
tations=annot)

# The PeakML importing function does not work with gap filled files
# so, we have to prepare a PeakML annotated file.

mzmatch.ipeak.sort.RelatedPeak(i="combined_highintensity.peakml",v=T,o="mzMatch_outputPOS.pe
akml",basepeaks="mzMatch_basepeaks.peakml",ppm=3,rtwindow=6)

# import PeakML file as xcmsSet object

mzAnnotP <- get.Mzmatch.annot("POS/mzMatch_outputPOS.peakml", onlyBP=FALSE)

# Repeat the process above to Negative mode to create a separate PeakML file
# this is how comb 3 (see below) was obtained

comb3 <- combineMolIon(mzAnnotP, mzAnnotN)
```

The combination of acquisition modes can be used as evidence to confirm a feature (peak associated to a retention time) as a true positive peak when it appears in both modes, and to increase the sampling power for molecules that ionize best in one mode. The package CAMERA provides ways to combine acquisition modes searching raw data for mass differences which obey user provided *ad hoc* rules (comb 1 in Table 1). We also provide an algorithm to combine individual ion annotations (comb 2 in Table 1) since it is hard to anticipate all possible ion rules in CAMERA's algorithm. Additionally to acquisition mode combination, one can chose to select only peaks with isotope/adduct evidence, or add non annotated peaks with simple heuristics as [M-/+H] (comb 2+ in Table 1), we refer to package manual pages for parameter details.

```
# include CAMERA non-annotated compounds, and snr retrieval
# comb 2+
ionAnnotP2plus <- get.annot(anP, allowMiss=TRUE, xset=xsetPnofill, toexclude=c("blank",
"medium", "QC"))
ionAnnotN2plus <- get.annot(anN, polarity="negative", allowMiss=TRUE, xset=xsetNnofill,
toexclude=c("blank", "medium", "QC"))

# Following with comb 2+
comb2plus <- combineMolIon(ionAnnotP2plus, ionAnnotN2plus)
sum(comb2plus$molIon[,3]==1)
sum(comb2plus$molIon[,3]==0)
```

For the present dataset we achieved a data reduction with combination strategies (comb), ranging from 74% (comb 2+, from peak groups to all possible putative molecular ions) to 86% (comb 2), as shown in Table 1.

Table 1 – Peak extraction and representation for xcms preprocessing steps.

	xcms			CAMERA			ProbMetab		
	peaks	peaks per sample	peak groups	groups	isotopes	adducts	all possible putative molecular ions	isotopes	adducts
pos	557071	24220	15270	8701	1595	2819	2546	1081	1465
neg	425602	18504	13226	7506	1039	2596	2111	737	1374
comb 1							4789/4096	1081/737	3708/3359
comb 2							4408	1718	2690
comb 2 +							8707	1718	6989
comb 3							6714	276	6438

pos/neg - acquisition mode; **comb1/2/3** - strategies to combine acquisition mode 1- CAMERA's combinexsAnnos **for positive/negative modes**, 2- ProbMetab's combMolIon function, with optional parameter to include non annotated ions (+), 3 – Integration to mzMatch, here using only “bp” and “potential bp” relationships.

Enhanced ion identification based on biological knowledge

In the following, we start from a reduced peak list and show how to combine spectral information and biological knowledge to improve metabolite annotation.

The main approach to search for candidate compounds with mass lists obtained by high resolution spectrometry, with soft ionization methods, is the search for exact mass in public compound databases. Public databases such as ChemSpider, PubChem and METLIN [11–13] provide extensive lists of compounds; however, these repositories do not have practical links to biological information, lacking associated pathways and reactions information. Moreover, these information sources bring synthetic compounds that are generally not present in biological matrices, and, therefore, add unnecessary complexity to the search space, thus hindering manual curation. Kind & Fiehn 2006 [14] concluded that the ideal would be to combine the use of databases aiming to be exhaustive with databases that have biological context, and this strategy has been used for making custom databases [6].

As a critical step, we build as biologically-driven as possible the database into which candidate peaks are searched on. First, we define a tabular format, inspired by mzMatch format. That format contains mandatory information: unique identifier, molecular formula and reactions that a compound is involved in. This information is strictly required for database matching and modeling, as explained below. Additional fields with links to external databases, pathways, structural information, etc, may be added.

As the public databases are constantly evolving, we provide online access (through dedicated API) to compound information of two main metabolic network databases with biological context to compounds, MetaCyc and KEGG [15, 16]. Additionally, as we believe that genome-scale metabolic reconstruction [17] can potentially provide the best representations of a specific organism metabolism, the ProbMetab package provides functionality to convert *SMBL* models [18] to the required tabular format, allowing integration to metabolism repositories such as MetExplore [19]. In the analyzes flow

presented here we used the KEGG (REACTION) database for exact mass matching.

```
# this mapping of compounds-to-reactions from KEGG is automatically loaded with ProbMetab
DB <- KEGGcpds
reactionM <- create.reactionM(DB, comb2plus, ppm.tol=8)
```

As illustrated in the ion extraction section, the choice of which subset of ion table will be used for downstream analysis can vary according to experimental setup. For now on, we assume that we have extracted a subset of non redundant putative molecule peaks, and have to deal with the uncertainty of assigning these peaks for known molecular formulas in a given database.

To illustrate the downstream analysis we choose the comb 2+ strategy where we have 8707 non redundant putative molecules from mass spectra, of which 1718 peaks (20%) have putative isotope peaks associated (Table 1). From all those putative molecules, 1386 (16%) have at least one candidate inside the mass search window (8 p.p.m) and 757 (54% of 1386) show two or more possible competing candidate compounds (matching formulas in the given mass window), highlighting the uncertainty in the assignment. It is known that there are still many compounds unknown in databases and that the number of possible different metabolites sampled by a single experimental technique is limited [20]. At this step we still have some mass peaks assigned to different compounds, since CAMERA (the peak summarizing tool chosen here) provides more than one possible annotation for some peaks. This is a desirable feature since, in the context of an exploratory analysis, we would investigate all possible annotations. It is important to note that a conservative mass window was used, which allows candidate overlap, choice which is justifiable in the analysis context of ranking and filtering, detailed in the following.

Probability modeling and estimation

With the list reduced to an observable subset of 757 mass to compound assignments with two or more candidates, we have now to try to rank these candidates with the information and knowledge available. For this we will use the model proposed below incorporating the likelihood of three components: observed isotopic ratios (for now we are considering only the proportion of molecules containing a single ^{13}C atom – $^{13}C^{12}C_{n-1}$ molecules [21]), the connection between compounds and retention time prediction error.

The relative isotopic abundance is very important to filter candidate formulas for a given mass [14]. However, there are few assessments on how accurate are the intensity measurements across

different mass spectrometry platforms, and practical ways to incorporate this information to mass annotation workflows, as exception the MZmine [22], that present challenges to automatically integrate with tools on R environment. One method to incorporate this information was presented by Weber et al. (2011) [21], the authors have shown that relative isotopic abundance have an *offset* in the prediction of the carbon number dependent of Signal to Noise Ratio (SNR), for the measured peak intensity. Taking the SNR in account the authors were able to correctly assign 44% of peaks to formulas.

In the present *Trypanosoma brucei* Exactive Orbitrap dataset, we were able to retrieve putative isotope $^{13}C^{12}C_{(n-1)}$ peaks for 30 compounds (among the 93 compounds with known identity), for which we know the true identity. For this set of compounds we can try to recover the SNR and estimate the carbon *offset* for defined intervals of SNR, and with that, build a formula filter. As shown in **Figure 1** for low values of Signal to Noise Ratio we have low confidence predicting the carbon number with the intensity ratio [23].

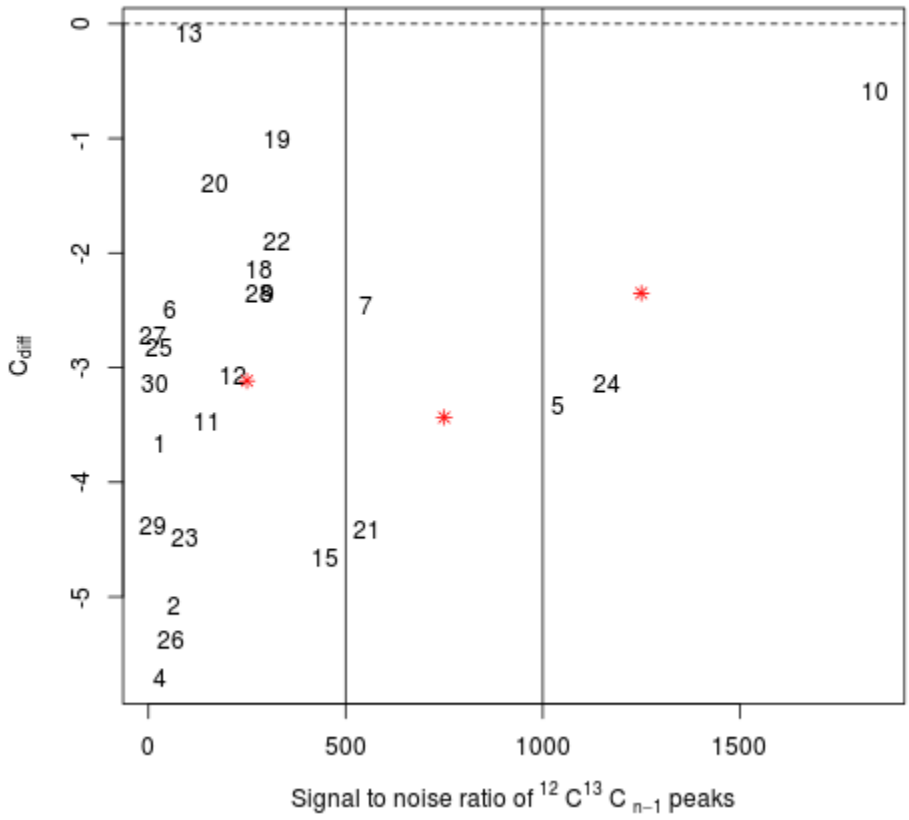


Figure 1 - Estimates of carbon number *offset* with Relative Isotopic Abundance for compounds of known identity. C_{diff} values (empirically calculated number of carbons minus actual number) against SNR of $^{13}C^{12}C_{(n-1)}$ peaks. Vertical lines represents the bins proposed by Weber et al. (2011) and red asterisks the C_{diff} mean for each bin.

Using the C_{diff} *offset* estimated by this approach we can implement a filter in the form (mean +*offset*) ± 3σ, proposed by [21], where σ is the standard deviation. In **Figure 2** it's shown, that the simple filter, mean ± 3σ, without the *offset*, misses the true carbon number for almost all compounds. The filter with the *offset* recovers all but two true compound carbon numbers. The only mistakes are compounds 19 and 22 (x-axis of **Figure 2**). This could be easily solved decreasing the bin size, which should reduce the *offset* for these compounds.

For the present subset of compounds only 11, 21 and 29 have candidate formulas (2, 4 and 2 different candidate formulas) with different number of carbons, and among them, only compound 21 had one formula (with 9 carbons) outside the filter range. For databases of compounds associated to biological knowledge the filter seems to have a narrow application. However, as shown below, the information of reaction can be codified from different sources, including the simulated possible formulas, case where the filter have been shown to be very useful.

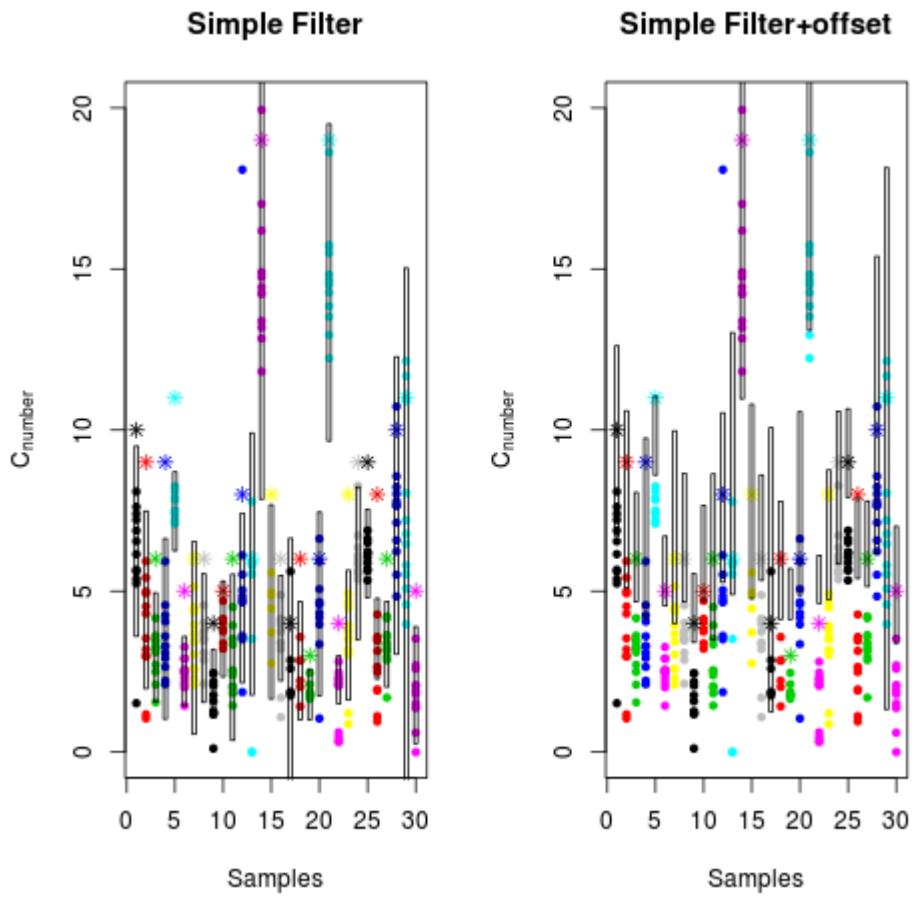


Figure 2 - Representation of a Relative Isotopic Abundance filter for carbon number. The filter on the right represents the simple filter, using raw sample predictions, and the left plot represents the use of the additional estimated carbon *offset*. The rectangles represent the filter, the points the estimates for repeated samples and the asterisks the true number of carbons in the molecule.

```
# number of masses with candidates inside the fixed mass window
# and masses with more than one candidate
length(unique(reactionM[reactionM[,"id"]!="unknown",1]))
sum(table(reactionM[reactionM[,"id"]!="unknown",1])>1)

# Calculate the ratio between observed and theoretical isotopic patterns.
# If you don't have an assessment of carbon offset to carbon number prediction
# skip this step and use the reactionM as input to weigthM function.
isoPatt <- incorporate.isotopes(comb2plus, reactionM, comb=1, samp=12:23, DB=DB)

# calculate the likelihood of each mass to compound assignment using mass accuracy, and
isotopic pattern, when present
wl <- weightM(isoPatt,intervals=seq(0,1000,by=500), offset=c(3.115712, 3.434146, 2.350798))

# codify the relation between compounds, given by reaction present in the biological
database DB
w <- design.connection(reactionM)
```

With the likelihood model at hand we have to provide a practical way to codify possible compound reactions. Previous works have shown that non random mass differences are correlated in replicated biological samples [24, 25]. These mass differences can be attributed to known metabolic reactions, and also be used to investigate new reactions. Although being very interesting in the context of an exploratory analysis, we chose to concentrate in the known metabolism universe, trying to focus in previous described metabolic reactions, and thus avoiding spurious connections of mass differences. There is clearly a tradeoff between the usage of known reactions or generic mass differences and the two are complementary. We show here the steps required to integrate mass differences to our approach. If one has a set of valid formulas with unique identifiers as in:

	id	name	formula	mass
1	c001	thymidine	C10H14N2O5	242.09
2	c002	thymidine (-H2O)	C10H12N2O4	224.08
3	c003	uracil	C4H4N2O2	112.03
4	c004	uracil (-H)	C4H3N2O2	111.02
5	c005	Glycerone phosphate	C3H7O6P	170.00
6	c006	Glycerone (-H2PO3)	C3H6O3	90.03

and wants to match these formulas against masses from a typical spectrometry experiment:

```
# Example of generic mass differences to ProbMetab modeling framework
exp_masses
rt massObs
[1,] 1035 242.09100
[2,] 500 224.07900
[3,] 711 90.03215

reac_matrix <- matrix(0, ncol=4)
ppm.tol <- 10

# match masses in a given mass window
for(i in 1:nrow(exp_masses)) {
    logical <- abs(((exp_masses[i,2]-db.mass)/db.mass)*10^6) < ppm.tol
    if(sum(logical)){
        reac_matrix0 <- cbind(matrix(exp_masses[i,], nrow=1),
        as.matrix(valid_formulas[logical,c("mass", "id")]))
    }
    reac_matrix <- rbind(reac_matrix, reac_matrix0)
}
reac_matrix <- reac_matrix[-1,]
```

Now, with the possible compound formula to masses attributions, one can search a list of generic reactions, with unique identifiers, and relate compounds to reactions. For a given list of generic reactions, as

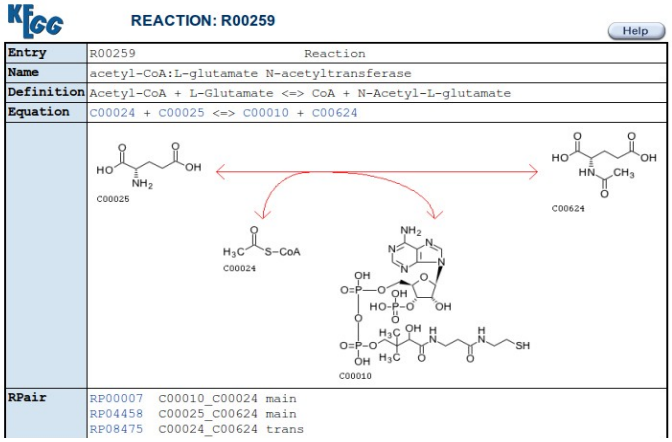
	reaction	reaction.name	mass.diff	reaction.id
1	-H2O	Loss of Water	18.01	r001
2	-H	Loss of Hidrogen	1.01	r002
3	+C2H2O	acetylation	42.01	r003
4	+CO2	Carboxylation	43.99	r004
5	-H2PO3	phosporilation	80.97	r005

we can use that list and build a matrix similar to reactionM matrix, shown above, allowing the integration to the analysis flux.

```
reac_matrix <- cbind(reac_matrix, rep("", nrow(reac_matrix)))
m_diff <- outer(as.numeric(reac_matrix[,3]),
          as.numeric(reac_matrix[,3]), "-"
       )
for(i in 1:nrow(m_diff)){
    for(j in 1:ncol(m_diff)){
        log <- abs(m_diff[i,j])> gen_reactions[,3]-0.01 &
            abs(m_diff[i,j]) < gen_reactions[,3]+0.01
        if(sum(log)){
            reac_matrix[i,5] <- paste(gen_reactions[log,4],
                            collapse=";")
        }
    }
}
cnames <- c("rt", "massObs","massDB", "id", "reactions")
colnames(reac_matrix) <- cnames
reac_matrix
rt massObs massDB id reactions
1 "1035" "242.091" "242.0903" "c001" "r001"
2 "500" "224.079" "224.0797" "c002" "r001"
6 "711" "90.03215" "90.0317" "c006" ""
```

With the analysis above we saw that the formulas c001 and c002 may be related by the reaction r001, and following this principle one can extend the list of generic reactions and use mass differences in an exploratory analysis context. The next step (Output representation) will allow one to cross putative reactions with (partial)correlations and export them in an user friendly visualization.

Instead of using a list of generic biochemical transformations we chose to use specific known reactions, using the reactions stored in our previous step. As discussed in [24], the usage of generic reactions can produce spurious connections, e.g., we observe a mass difference corresponding to a transformation, but the true compounds can't participate to this reaction. The **Figure 3** illustrates the basic approach to use reference manually inspected repositories of biochemical reactions.



	C00026	C00024	C00010	. . .	C00624
C00026	0	1	1	. . .	1
C00024	1	0	1	. . .	1
C00010	1	1	0	. . .	1
. . .			. . .		
C00624	1	1	1	. . .	0

Figure 3 – Example of KEGG reaction database to show how known reactions are codified as entry to probabilistic model. The same matrix can be obtained from alternative reaction sources, such as *SBML* models.

As preconceived by Rogers et al. (2011) every function of type $f(x_m, y_c)$ that increases as a compound y_c, of a vector with *C* candidate compounds, becomes a better candidate for a mass x_m, of a vector with *M* mass peaks, can be used to add information about compound classification (See the main text for a detailed explanation). In the present implementation, we should generate matrices of the same format of *wm*.

To implement the idea of information incorporation, we use of retention time prediction, using the idea presented in Creek et al. (2011) [26]. The authors proposed a model of quantitative relationship between the structure and retention time of a compound (QSRR - Quantitative Structure and Relationship Retention). A model of the form

$$\hat{t}_i = \sum_{j=0}^{p-1} \hat{\Psi}_j d_{ij}$$

where the dependent variable t_i (retention time) is predicted by a set of molecular descriptors for each compound d_i, which is weighted for its contribution by *j* regression coefficients, and *p* is the number of parameters in a model with intercept. To reproduce the result from Creek et al. (2012) we used the compound descriptors available at IDEOM's DB sheet, and the standard compound measured retention time in the RTcalculator sheet.

```
# Please install the suggested packages if you want to reproduce
# the retention time modeling
```

```
# install.packages(c( "bootstrap", "leaps", "mgcv"))

db <- read.csv("IDEOM_v18_DB.csv")

rtTab <- read.csv("standardRetentionTime.csv")
head(rtTab)
Compound.Name...MW..to.. standard.RT calculated.RT X..error logD..3.5.
1 Imidazole-4-acetate 17.98 10.80 -40% -1.4
2 N-Acetyl-D-glucosamine 12.09 12.55 4% -3.2
3 Melatonin 5.71 6.16 8% 1.2
4 Phenylhydrazine 5.94 9.86 66% -0.5
5 4-Aminobenzoate 6.02 7.09 18% 0.7
6 Nicotinate 7.93 8.58 8% -0.2
```

Now we have to format the data to run the regression model selection.

```
testM <- rtTab[,5:10]
predSet <- data.frame(id=reactionM[reactionM[,4]!="unknown", 4],
        data=as.numeric(reactionM[reactionM[,4]!="unknown", 1])/60
    )
l1 <- sapply(predSet$id, function (x)
                which(db[,"KEGGid."]==
                    sub("cpd:","",as.character(x))
                )
        )
l1[which(unlist(lapply(l1, length))==0)] <- 41639
db2 <- cbind(db[,"KEGGid."],db[,21:49])
db2 <- rbind(as.matrix(db2), rep(0,30))
sapply(which(unlist(lapply(l1, length))>1), function(x)
                            l1[[x]] <<- l1[[x]][1]
    )
v1 <- unlist(l1)
cols <- c(3,4,5,6,8,9)
predM <- db2[v1, cols]

# some compounds have missing descriptors, for those is not
# possible to predict the retention time
predM[which(is.na(predM), arr.ind=TRUE)[,1],] <- 0
predM <- apply(predM, 2, as.numeric)
sum(apply(predM, 1, sum)==0)
testSet <- data.frame(id=rtTab[,1], data=rtTab[,2])
descData <- list(predM=predM, testM=testM)
```

With the formatted test set (the measured retention time and structure descriptors to standard compounds) to estimate the model, and the prediction set (a matrix of structure descriptors to candidate compounds) to have the retention time estimated, we will use the *leaps* package [27] to select a model based on Mallows's Cp criteria.

```
# Reproducing Creek et al. (2012) to format the information to be added to the
# probabilistic model
```

```
myresult <- rt.predict(testSet, predSet, descData, voidTime=4.5)

# predicted retention factor
head(myresult$pred)
[,1]
[1,] 0.49353197
[2,] -0.28788740
 [3,] 0.41521135
[4,] -0.08651121
[5,] 0.88894204
[6,] 0.88194981
myresult$ans$rawRLm
[,1]
[1,] 0.7898401
```

The R^2 is the same that IDEOM estimated, now we can translate the retention time prediction error in a likelihood ranking of candidate compounds. We chose to design an exponential function in the form

$$f\left(e_o;\lambda\right)=\frac{1}{\lambda}e^{\frac{-e_o}{\lambda}}$$

where λ represents the scale parameter, which is set to the tolerated estimated error, and $\hat{e}_o$ is the measured error for each candidate compound.

```
myweight <- weightRT(myresult, reactionM)
```

With this likelihood function we can easily incorporate the information by multiplying the increasing functions as shown below.

```
# Example of Hadamard product (element-wise), which allows easy insertion
# of information on the model.


myweight$wm[1:5,1:5]
[,1] [,2] [,3] [,4] [,5]
[1,] 0.48483338 0.00000000 0 0 0
[2,] 0.03745601 0.00000000 0 0 0
[3,] 0.00000000 0.46685613 0 0 0
[4,] 0.00000000 0.03785027 0 0 0
[5,] 0.00000000 0.00000000 1 0 0

wm[1:5,1:5]
[,1] [,2] [,3] [,4] [,5]
[1,] 2.289289e-07 0.000000e+00 0.00000000 0 0
[2,] 2.289289e-07 0.000000e+00 0.00000000 0 0
[3,] 0.000000e+00 4.579618e-05 0.00000000 0 0
[4,] 0.000000e+00 4.579618e-05 0.00000000 0 0
```

```
[5,] 0.000000e+00 0.000000e+00 0.06752744 0 0


wm[1:5,1:5]*myweight$wm[1:5,1:5]
V1 V2 V3 V4 V5
1 1.109924e-07 0.000000e+00 0.00000000 0 0
2 8.574762e-09 0.000000e+00 0.00000000 0 0
3 0.000000e+00 2.138023e-05 0.00000000 0 0
4 0.000000e+00 1.733398e-06 0.00000000 0 0
5 0.000000e+00 0.000000e+00 0.06752744 0 0
```

Now, with all information, potentially coming from different sources, we can use the Gibbs Sampler as proposed by Rogers et al. 2009, and calculate the assignment of posterior probabilities:

```
# calculate the posterior probabilities according with the proposed model
x <- 1:ncol(wl$wm)
y <- 1:nrow(wl$wm) # These two will be wrapped in a the function in future

# should take about 30 minutes for a regular desktop computer
#(ubuntu 64bits, with 8gb of memory), with
# a problem of the same size (number of masses) of the presented in this document.
system.time(conn <- gibbs.samp(x, y, 5000, w, wl$wm))

# export the classification table, optionally as R matrix or .html file
system.time(ansConn <- export.class.table(conn, reactionM, comb2plus,
                        filename="AnalysisExample", html=TRUE, DB=DB))
```

The classification matrix provided by ProbMetab features the columns: experimental masses (Measured Masses); ranked candidate compound list (Most Probable Compound); probability of each candidate (Probability) - the correct way to interpret the probability of a mass to be assigned to a given compound is: among the set of candidates presented, this is the most likely ranking according to model assumptions. This interpretation is different from probability of a mass be a given compound, since the model did not restrict the search space to the true metabolome, nor guarantees that a mass peak refers directly to a metabolite's mass; entropy of probability distribution among the candidates (entropy – from information theory); optionally the p-value from the *t-test*/*anova* between samples; and a condensed ion annotation in the format: original mass# original retention time# isotopic pattern, if present# adduct pattern, if present. The user can choose to export a *html* table, which will be associated to extracted ions chromatogram (EIC - Extracted Ion Chromatogram) plots of all peaks, written to R working directory.

The *T. brucei* dataset gives a good opportunity to show how we want apply our approach to compound annotation. It was previously analyzed with IDEOM [10], and, according to IDEOM's documentation (http://mzmatch.sourceforge.net/ideom/Ideom_Documentation.pdf), “If metabolite

assignment remains ambiguous, the first matching metabolite in the DB is assigned", resulting in static ranking that may not be appropriated for different analysis scenarios.

As we show in the (external file filter_comp.xls, sheets "ideonRTnoRT" and "MzmatchProbMetab") we can recover the same amount of "TRUE peaks", but the ranking provided by associated information gives a dynamic ranking, that can be improved as one can model additional information. In the comb 2+ filtering strategy we have recovered 93 compounds with known identities (external file filter_comp.xls, sheet "cameraProbMetab"), out of 127 compounds previously identified [10].

The identification based on retention time standards presented by Creek et al., 2011 [26] is very interesting in the context of a targeted analysis, where a set of compounds is elected for tracking. However, such approach has the drawback of time and money expenses, the limited number of purified standards available, and the non-linear retention time deviation in Liquid Chromatography that compromises the retention time predictions, as previously reviewed [28].

Nevertheless, implementing the retention time prediction of [26] as additional source of information to the model by matrix multiplication, as shown above, we can rank candidate formulas based on retention time prediction error. The compound ranking for "rt Lik Only" column shown in Table 2 (see external file inf_incorporation.xls for a peak by peak ranking) represents the assessment of retention time prediction based on know compound identities. If we don't have (or don't trust) in a database with preferred identities ranked, as [10], the classification based on retention time prediction error can be misleading, with a low number of compounds being classified as the correct 19.2% and with a high number of incorrect rankings, 17%. If we look carefully to Table 3 we show examples of whole classification were the retention time prediction was the only model component able to distinguish the compounds N6-Acetyl-N6-hydroxy-L-lysine and N5-(L-1-Carboxyethyl)-L-ornithine, that have same molecular formula and were not differentiated by its reactions with the other compounds in the sample. This show how new information can be helpful, if the retention prediction could be improved, or even used for a subset of compounds with small prediction error.

Table 2 – Comparison among addition of different sources of information to the probabilistic model.

Class	rt Lik Only	Mass Lik Only	MetSamp MATLAB	ProbMetab R environment	All
No defined classification	44 (56%)	66 (84.6%)	0 (0%)	0 (0%)	0 (0%)
Correct identity between higher probabilities	2 (2.6%)	2 (2.6%)	0 (0%)	0 (0%)	0 (0%)
Correct identity is the higher probability	15 (19.2%)	6 (7.7%)	47 (60.2%)	50 (64.1%)	49 (62.8%)
Incorrect identity is the higher probability	17 (21.8%)	4 (5.1%)	31 (39.8%)	28 (35.9%)	29 (37.2%)
	78 (100%)	**78 (100%)**	**78 (100%)**	**78 (100%)**	**78 (100%)**

rt Lik Only – classification based on retention time error prediction likelihood, R implementation; **Mass Lik Only** – classification based only on isotopic pattern (when present) and mass accuracies likelihood, *erfc* function on R implementation; **Mass Lik+connections: MetSamp MATLAB implementation –** classification based on gaussian mass accuracy and KEGG reaction connections; **ProbMetab R implementation** - classification based on *erfc* mass accuracy and KEGG reaction connections; **All** - classification based on isotopic pattern (when present), mass accuracies, reaction connections and retention time error prediction.

Using only the mass accuracy (“Mass Lik Only”) has a poor classification definition 7.7%, mainly because of the isomers for a given formula. It can also be misleading for some mass windows where we have a higher error [29], and for suboptimal preprocessing parameters where we have higher errors in mass recovery.

We have used isotopic pattern carbon number prediction, as information associated to experimental accuracy, to filter possible formulas. In our database matching only 30 out of the 93 compounds had isotopic peaks recovered with CAMERA's standard search parameters. For those, only 3 had matching formulas with different number of carbons (external file inf_incorporation.xls), and the isotopic filter implemented was able to rule out only one formula, C9H15N4O9P (5-Amino-6-(5'-phosphoribosylamino)uracil), which had its probability decreased (fixed decrease of 10 times, still keeping the compound candidate in the list for manual inspection) to be assigned to mass 354.05, Table 3.

The Mass likelihood + connections (prior knowledge of metabolism) has the best response, considering the trade-off of correct identities 64,1% and low incorrect identities 35,9% assigned to masses (considering only the higher probability). The model using all probability components (“All” in Table 2) had a influence from incorrect assignments of retention time component, hindering a high number of global incorrect assignments.

If we look very carefully to peak by peak comparison (external file inf_incorporation.xls), we can see that most of so called “incorrect” or “correct” assignments have a small probability above the second candidate (around 1% higher), as the classifications of N6,N6,N6-Trimethyl-L-lysine in Table 3. In this scenarios it is most probable that we don't have enough information in the model to afford the

correct classification, additional information and manual curation should be used to inspect the classification, in some cases the ranking helps to differentiate the most probable identities ("Correct identity between higher probabilities" in Table 2**)** or to show to the experimenter that he has to further investigate that mass.

Table 3 – Examples of probability attributions for different model components, with the correct identity in bold.

		rt Lik Only	Mass Lik Only	MetSamp MATLAB	ProbMetab R environment	All
				Probability		
N5-(L-1-Carboxyethyl)-L-ornithine	204.110633	0.081	0.5	0.69	0.672	0.09
N6-Acetyl-N6-hydroxy-L-lysine		0.919	0.5	0.31	0.328	0.91
Phenolsulfonphthalein	354.0558718	0.25	0.839	0.654	0.926	0.905
5-Amino-6-(5'-phosphoribosylamino)uracil		0.25	0.115	0.255	0.023	0.029
WIN56291		0.25	0.046	0.064	0.045	0.06
2-Caffeoylisocitrate		0.25	0	0.026	0.006	0.006
7,8-Diaminononanoate	188.1521096	0.5	0.5	0.472	0.446	0.444
N6,N6,N6-Trimethyl-L-lysine		0.5	0.5	0.458	0.48	0.482
L-Histidine	155.0688293	0.2	0.2	0.792	0.781	0.798
3-(Pyrazol-1-yl)-L-alanine		0.2	0.2	0.082	0.106	0.092
D-Histidine		0.2	0.2	0.043	0.034	0.036
Kininogen		0.2	0.2	0.041	0.039	0.043
Histidine		0.2	0.2	0.041	0.04	0.031

With the examples discussed above we wanted to highlight that sources of information can contribute in different ways to compound classification and care must be taken interpreting the results. Here we wanted to provide a systematic way to combine information, once we know how to model that information through all metabolome sampling range. Taking into consideration the error rate of all model components we can benefit of summarized ranked view when we have efficient visual tools, as discussed in the next section. We have applied the MetSamp MATLAB implementation (the source code was kindly provided by Prof Dr Simon Rogers from University of Glasgow under personal request) with precision adjusted to 6.25 x 10^{12} (were two SD from mean is equivalent to 8 p.p.m), and using the unique masses (x) and compounds (y) from *reactionM* matrix and connections from *w* matrix. After the processing we manually exported the *out.allsampcomp* matrix from MATLAB and calculated the probabilities for the ranking presented in Table 3.

With our R implementation we want to highlight that, exporting the ranking provided by modeling information associated with yet not modeled information (ion adduct pattern, chromatograph shape, correlation) is essential to provide tools to allow the experimenter decided the true compound identity, and which ones have impact in the conditions under investigation. In Table 4 we see that, considering the

ranking, not only the correct classification as in Table 3, up to 90% of the compound have their correct identity among the first 3 top ranked probabilities, and with the help of our visualization strategies can led to high quality compound annotation. We have tested our implementation against MetSamp (http://www.dcs.gla.ac.uk/inference/metsamp/), in a regular desktop computer, running Ubuntu 12.10 64 bits, 8gb of memory, with a problem of the same size (number of masses) of the presented in this document. *ProbMetab* took 38 minutes to run the *gibbs sampler* algorithm, against 138 minutes of *MetSamp's* version.

Table 4 - Cumulative proportion of correct identity position of models presented on table 2.

	Cumulative Proportion of Correct Identity Position		
Correct Identity Position	**MetSamp MATLAB**	**ProbMetab R environment**	**All**
1	0.667	0.699	0.688
2	0.828	0.828	0.839
3	0.882	0.903	0.882
4	0.925	0.935	0.935
5	0.935	0.957	0.946
6	0.946	0.978	0.968

Comprehensive output representation

The main product of the probabilistic model is a list of ranked attributions, which depends on our knowledge of the biological model and experimental setup. The lists can be quite extensive and the user needs a proper representation to make sense of this data. The main output is a table that can be exported as an R matrix or *.html* file and contain the rank of candidates.

One way to put together a *post hoc* check of predicted connections, and a biologically inspired visualization is to cross reference the set of all possible reactions with (partial)correlation weighed networks. Assuming that a given mass could have as identity one or more compounds, and each of this compounds could have one or more connections to compounds candidates to other masses, the generation and inspection of all possible networks will be infeasible, just as the inference of its distribution for the probabilistic model. Instead of generating all possible networks one could treat each mass as a node, and all possible reactions between its candidates to other mass candidates as edges. This simple approach decreases the number of graphs to one, but a node still contains a high number of possible identities and connections as associated information to analyze. If we cross that information with correlations between masses in repeated samples, we reduce the possible connections and compound identities responsible for this connections. We provide an algorithm to cross all possibilities and automatically exports this networks to Cytoscape [30] (a biological network visualization

software).

```
# This script is intended to reproduce the Figure 4,
# as well as to provide a working example on the main features of ProbMetab
# graph functions
#
# Following this analysis we generate a graph with reactions overlaid with correlations
# and export use additional information to provide formatting to this graph.

# calculate the correlations and partial correlations and cross reference then with
reactions
# load("probmetab-case1-box02.RData") # load the necessary objects to draw the graph

mw <- which(w==1,arr.ind=TRUE)
corList <- reac2cor(mw, ansConn$classTable[,-c(8:18)], corprob=0)

gr.cor <- ftM2graphNEL(corList$cor.vs.reac)
classTable <- ansConn$classTable
node.names <- apply(classTable[classTable[,4]!="",1:7], 1, function(x) paste(x[6], "-",
            paste(strsplit(as.matrix(x[7]), "#")[[1]][1:2], collapse="\n"), sep="")
        )
node.names <- sub("^\\s+", "", node.names)
snode.names <- node.names[as.numeric(nodes(gr.cor))]

# Example of some edge and node attributes see export2graph man page to more details
form <- edgeNames(gr.cor)
form <- data.frame(form, form %in%
apply(corList$signif.cor[corList$signif.cor[,1]>0.75,2:3], 1, paste, collapse="~"))
form <- data.frame(form, form[,1] %in% apply(corList$signif.cor[corList$signif.cor[,1]<(-
                           0.75),2:3], 1, paste, collapse="~")
        )
# Format the edge attribute table with exadecimal color codes
# to export to Cytoscape. In this case red to positive correlations
# higher than 0.75, and green to negative correlations smaller than
# -0.75
cnames <- c("edge.name", "color.#FF0000", "color.#006400")
colnames(form) <- cnames

# index of known identity compounds from Supplementary File 4
csel <- "119 393 1106 661 1264 418 482 423 1114 413 459 62 1136 1067 362 767 656 92 618 109
555 1291 516 1123 1128 553 379 701 242 302 697 356 1155 47 184 896 3 40 182 4 1098 41 1031
782 45 246 560 416 1088 778 81 51 88 31 383 24 352 1121 730 744 425 1120 120 370 1086 782
757 451 52 58 498 769 376 337 551 69 477 548 543 677 526 581 766 933 90 870 317 332 212 396
594 713 748"

csel <- as.numeric(strsplit(csel, " ")[[1]])

form2 <- nodes(gr.cor)
form2 <- data.frame(form2, form2 %in% csel)
cnames2 <- c("node.name", "lcolor.#0000FF")
colnames(form2) <- cnames2

# The index of all nodes
sn <- as.numeric(sub("(^\\d+)-.+", "\\1", snode.names))

# Where the known compounds are in the vector of correlated nodes
coord <- sapply(csel, function(x) which(sn==x))
scoord <- coord[unlist(lapply(coord, length))!=0]
scoord <- unlist(scoord)

# names of known identity compounds from Supplementary File 4
cpdnames <- "AMP%Urocanate%Uracil%Pseudouridine 5'-phosphate%D-Ribose 5-phosphate%Guanine%D-
```

```
Sorbitol%Propanoyl phosphate%Maleamate%L-Methionine%sn-Glycerol 3-phosphate%N6,N6,N6-
Trimethyl-L-lysine%(S)-Malate%N6-Acetyl-L-lysine%L-Cysteine%Ascorbate%Glutathione
%Pseudouridine%Adenosine%5'-Methylthioadenosine%L-Cystathionine%Inosine%S-Sulfo-L-cysteine
%5-Amino-4-imidazole carboxylate%Mesaconate%N-Acetyl-D-glucosamine%4-Guanidinobutanal%S-
Adenosyl-L-methionine%alpha-D-Glucosamine 1-phosphate%Glycine%Riboflavin%L-2,4-
Diaminobutanoate%D-Xylonolactone%Xanthine%S-Adenosyl-L-homocysteine%D-Glucose%L-Arginine%L-
Lysine%L-Serine%Putrescine%(R)-3-Hydroxybutanoate%L-Glutamate%Urate%D-Gluconic acid%(S)-4-
Hydroxymandelonitrile%Hypoxanthine%Carnosine%L-Arabinose%L-Alanine%Citrate%Folate
%Isopyridoxal%L-Cystine%L-Asparagine%L-Glutamate 5-semialdehyde%Nicotinamide%L-Valine
%Taurine%2-Hydroxy-3-oxopropanoate%(S)-3-Methyl-2-oxopentanoic acid%L-Histidine%L-Threonine
%Phenolsulfonphthalein%Imidazole-4-acetate%Pyruvate%L-Gulonate%Allantoin%N(pi)-Methyl-L-
histidine%Pyridoxamine%L-Tyrosine%LL-2,6-Diaminoheptanedioate%3-(4-Hydroxyphenyl)pyruvate%L-
1-Pyrroline-3-hydroxy-5-carboxylate%Hypotaurine%Pantothenate%N6-Acetyl-N6-hydroxy-L-lysine
%D-Glucosamine%N2-(D-1-Carboxyethyl)-L-lysine%sn-glycero-3-Phosphoethanolamine%Maltose%L-
Kynurenine%Cytidine%D-Glucuronolactone%Succinate%Thymidine%N-Acetylneuraminate 9-phosphate
%Glycerol%Diethanolamine%D-Glucose 6-phosphate%Ethanolamine phosphate%L-Arginine phosphate
%CDP-ethanolamine%Deoxyribose"

cpdnames <- strsplit(cpdnames, "%")[[1]]

# replace the node label of compounds identified with the true compound name
scpdnames <- cpdnames[unlist(lapply(coord, length))!=0]
snode.names[scoord] <- scpdnames

cpdnames <- as.character(sapply(classTable[classTable[,2]!="unknown",2], function(x) D
            B$name[DB$id==as.character(x)])
                )
classTable <- as.matrix(classTable)
classTable[classTable[,2]!="unknown",2] <- cpdnames

# create an initial visualization without leaving R environament
createJSONToCytoscape(gr=gr.cor, node.label=snode.names)
openGraph("network.json", classTable=classTable, openBrowser=TRUE)

# This functions extracts pathway information from KEGG API,
# and needs the KEGG codes, so we have to load the original
# classification table again

cpdInfo <- create.pathway.node.attributes(ansConn$classTable, graph=gr.cor, DB=DB,
filename1="path1.noa", filename2="path2.noa", organismId="tbr")

create.reaction.edge.attributes(classTable, graph=gr.cor, w=w, reactionM=reactionM, DB=DB,
filename="reac.eda")

export2cytoscape(gr.cor, node.label=snode.names, cwName="test4",node.form=form2,
edge.form=form, cpdInfo=cpdInfo, classTable=classTable)
```

We are going to illustrate our graph representation with CAMERA filtering approach, in which we have recovered 93 correct identities, comb 2+ strategy (external file filter_comp.xls, sheet “cameraProbMetab”). If we look among these compounds, which ones have an absolute correlation higher than 0.75, and overlay with a set of possible reactions, we can export them colored in blue within the entire correlation network as in the **Figure 4**. For 32 identified compounds with significant correlations we had the identities confirmed by 20 (62.5%) (see external file classByReactions.xls for a detailed analysis). With our visualization strategy the user can see all the possible candidates, the probabilities of each one and the possible reactions for the correlation represented by the node. The

overlaying strategy was very efficient to show that, when there is a correlation between the node x and node y the known reactions between the possible identities of those nodes only led to few possible compounds, many times to only one compound (external file classByReactions.xls).

The automatic exporting from R to Cytoscape [31] allows the user to navigate through complex networks, with information on pathways and reactions associated as node and edge attributes, respectively, allowing search the pathways dynamically inside the network with Cytoscape filters. We can also export the graph format to a web server in a way that a user do not need to install Cytoscape, and use only R environment, sending the file to a web browser (see an example at http://labpib.fmrp.usp.br/methods/probmetab/).

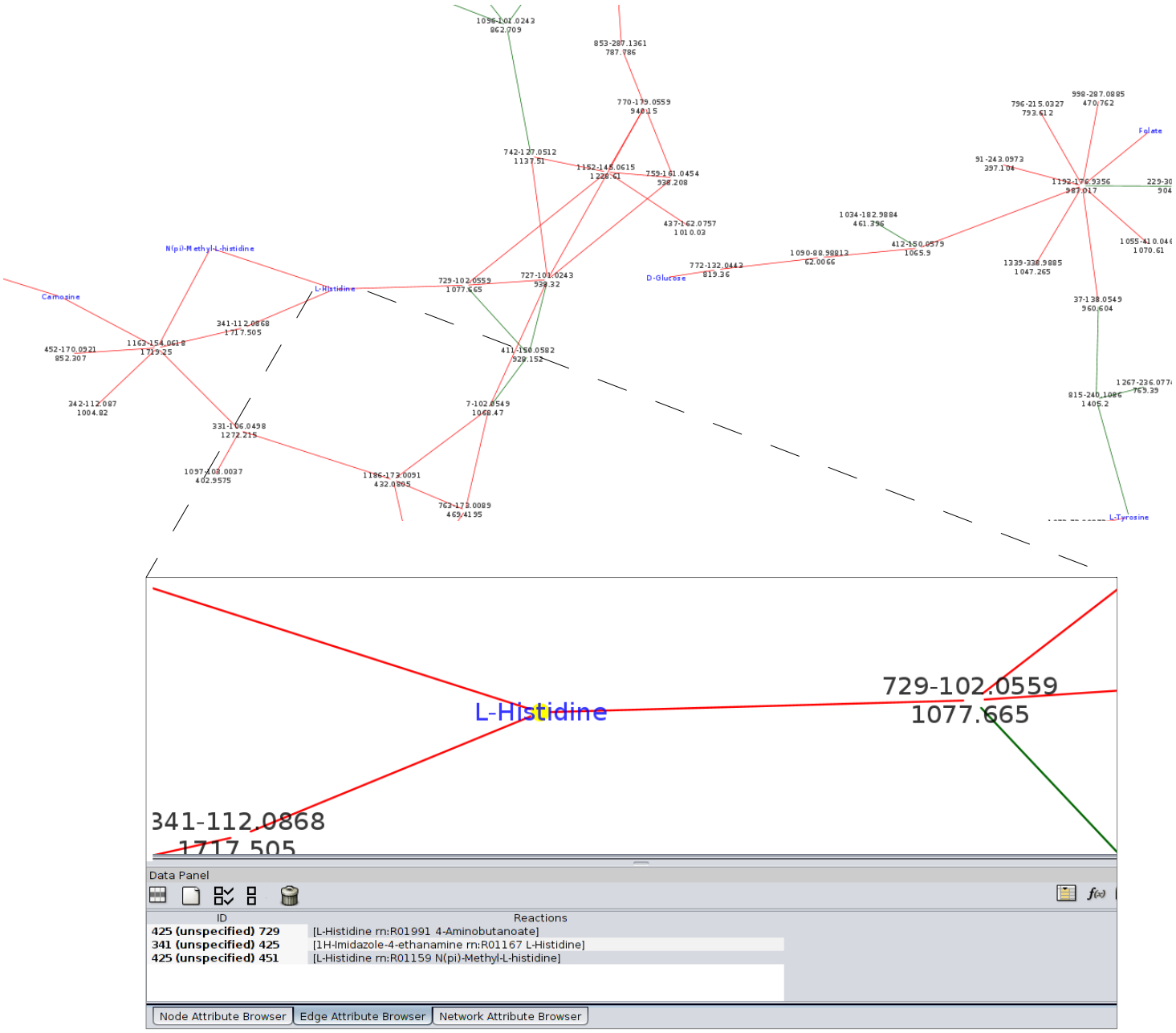


Figure 4 – Representation of peak table of extracted mass peaks exported as correlation weighted network overlaid with reaction network. The main information exported to Cytoscape window can be seen in Cytoscape's Data Panel. With the algorithm the node 425 (mass 155.06 in Table 3) had 6 possible identities, and each identity had a set of possible reactions with other mass candidates. However, the unique candidate reactions that overlaid with correlation are reactions that led to L-Histidine.

The present data set was used in [2] as Figure S3 to discuss the effect of the drug Nifurtimox on *T. brucei*, and the authors state "Nifurtimox (mass: 287.0577, RT: 5.25 minutes) was observed in all

treated samples, in addition to a mass (mass: 257.0834, RT: 13.5 minutes) consistent ..” . In the present analysis those peaks were found, mass peaks 105 and 599 (one can see the peaks creating the *ansConn$classTable* matrix as shown above), but no reaction linking this two peaks was present in our matching database. The graph inspection also shows many identities associated to aminoacid biosynthesis, the process of interest to investigate under drug treatment in the original work.

Conclusions

The restricted sampling nature of a specific analytical workflow (from sample preparation, to data processing), the complexity of ionization and the lack of knowledge on metabolome extension renders metabolomics to a far more limited capabilities than anticipated by their practitioners [32]. As time passes by and knowledge improves, we know better the gaps in sampling and the metabolite species subset that we can actually observe in each experimental setup, therefore we have to use an extensible analysis workflow to incorporate all knowledge we gain with this evolution. Dividing the processing in tree steps: 1) Ion annotation extraction and data base matching; 2) Probability modeling and estimation and 3) Comprehensive output representation, we hope to stress the importance of experimental information incorporation and biological context to capitalize the usage of available knowledge. Moreover, we supply an initial implementation in an environment that supports many auxiliary tools which can improve the annotation.

From the perspective of information incorporation we already have established sources of information that can be modeled and added to the present analysis, as sample preparation specificity [33], retention time [26], and MS/MS information [22]. As these ideas are developed all this information can start to be shared and better quality annotations will be provided. This information can be accessed from central repositories through API access, as functions provided to access biological context information. Databases such as MetaboLigths [34] should allow comprehensive information exchange giving opportunity to model different sources of information.

As the sampling nature of a given general untargeted LC-MS technique approach hamper us to observe all compounds of specific pathways, the best way to represent such sample would be superimposing it in the metabolic network. The correlation weighted networks have been used to map biochemical reactions to mass peaks [24, 25]. Partial-correlations may be biased because of hidden variables (e.g. enzymes) or non-linear relation between variables [35]. On the other hand, there is evidence showing that they can better represent relation between compounds, excluding the effect of other variables [36]. Therefore allowing the interchangeable use of the two approaches is important to spot specific metabolic changes.

We had implemented and extended a method to annotate compounds, in a framework that allows the introduction of prior knowledge and additional spectral information. With the R package *ProbMetab* we provide means to summarize the results of series of analysis needed to extract information from complex high dimensional mass spectrometry data, and help the experimenter to track metabolism changes in the process of interest.

ProbMetab Case Study 2: Original Metabolomics dataset from sugar cane leave, illustrates searching metabolism alterations during stress conditions

Motivation

The agricultural breeding traits, such as height, number of fruits, fruit size, dry weight, are controlled by the interaction and co-regulation of different metabolic pathways. The integration of information coming from varied molecular and genetics studies is the main challenge to elucidate the mechanism controlling these pathways. As an example, the mechanisms that regulate the synthesis, transport and accumulation of sucrose in sugar cane has been extensively studied since the middle of last century, revised [37, 38], however, conflicts revised in 1995 are still present in 2005, with the location of key enzymes and transporters still unknown.

Metabolomics is increasingly playing its role unraveling mechanistic metabolism changes, giving information of important pathways for phenotypes of interest [39, 40]. With the present study case we try to show how to spot metabolism changes that may be associated to an environment condition.

Data Analysis

Nine months old SP80-3280 hybrid sugar cane plants grown in the field (Experimental field Apta de Jaú), under irrigation (control) and area not irrigated (dry land, water stress treatment), had the +1 leaf samples collected from 5 plants (biological replicates). On the collection day, the soil water stress treatment had a relative humidity of 50% field capacity.

The extraction was performed according to the method described in [41] with some modifications. Control and water samples stress were extracted from 50 milligrams material macerated in liquid nitrogen. After addition of 1 mL extraction buffer (99.875% methanol and 0.125% formic acid) samples were vortexed for 10 seconds. The samples were then sonicated for 15 minutes at the maximum frequency (40KHz) under 20°C and then centrifuged for 20 minutes, at 14000 rpm and 20°C. The supernatant was filtered through a 0.22 μm filter and transferred to a new tube.

After metabolite extraction, the samples were analyzed with UPLC ACQUITY QTof – Ultima (Waters) system, in triplicate. For this analysis a reverse phase chromatography column (Acquity UPLC BEH C18 1.7 μm 2,1 x 100 mm) was used. Two buffers were used: Buffer A (H2O + 0.1 % formic acid) and buffer B (acetonitrile + 0.1 % formic acid). The elution gradient used was: 95 % A and 5% B (for 3 minutes), 75 % A and 25% B (for 3 minutes), 5% A and 95% B (for 3 minutes), and 95 % A and 5% B (for 4 minutes) to column wash and reconditioning. The flow gradient employed was 0.5 ml/min.

The source used was electrospray ionization (ESI). The mass spectra were acquired in positive mode, and mode V. The instrument was operated with capillary voltage of 3.0 kV and cone voltage of 35 KV. The temperatures of the source and desolvation were 150 °C and 450 °C. The desolvation gas flow was 550 L/h and 50 L/h for nebulizer gas (nitrogen). The mass spectrum was acquired in the ratio of mass/charge (m/z) range of 100-2000 Da

The peak picking, grouping and retention time correction were carried out as shown in the script below.

```
# ProbMetab suggested application
# initial parameters from
http://www.nature.com/nprot/journal/v7/n3/fig_tab/nprot.2011.454_T1.html
# load required packages
library(ProbMetab)
library(xcms)
library(CAMERA)

# Preprocessing
 xset <- xcmsSet(".", method='centWave', ppm=15,
            peakwidth=c(5,20), prefilter=c(0,0)
       )
 xset <- group(xset)
 xset2 <- retcor(xset,method="obiwarp",profStep=0.1)
 xset2 <- group(xset2, mzwid=0.015,minfrac=0.5,bw =2)
 xset3 <- fillPeaks(xset2)

an <- annotate(xset, perfwhm=0.6, cor_eic_th=0.75,
         mzabs = 0.01 , polarity="positive")

# load("probmetab-case2-box00.RData") # run to avoid to deal with raw data and go directly
to examples

# example of biocyc API alternative usage
# Chose an organism to download metabolites on its known metabolism
# ara is the organism code for Arabidopsis thaliana
vpth <- get.pathway.by.organism.biocyc("ara")
# optionally use parallel processing
library(doMC)
registerDoMC()

#  Retrieve all compounds associated to known Arabidopsis pathways
m <-   foreach(i=1:length(vpth)) %dopar% get.compounds.by.pathway.biocyc(vpth[i])
m2 <- do.call("rbind", m)
m3 <- unique(m2)
m4 <- m3[-grep("Error", m3[,3]),]

# Retrieve all single compound reactions for each compounds
rlist <- list()
rlist <- foreach(i=1:nrow(m4)) %dopar% get.reactions.by.compound.biocyc(m4[i,1])
# before attaching the reactions, verify if all compounds have at least one reaction
which(unlist(lapply(rlist, length))==0)
m4 <- cbind(m4, unlist(rlist))
colnames(m4)[4] <- "reactions"
m4[,3] <- gsub("\\s", "", m4[,3])

 m4[300:305,]
```

```
 id
 "ARA:PHOSPHORIBULOSYL-FORMIMINO-AICAR-P"
 "ARA:AICAR"
 "ARA:D-ERYTHRO-IMIDAZOLE-GLYCEROL-P"
 "ARA:IMIDAZOLE-ACETOL-P"
 "ARA:L-HISTIDINOL-P"
 "ARA:HISTIDINOL"
 name                                          formula
 "phosphoribulosylformimino-AICAR-P"           "C15H21N5O15P2"
 "aminoimidazole+carboxamide+ribonucleotide" "C9H13N4O8P1"
 "D-erythro-imidazole-glycerol-phosphate"      "C6H9N2O6P1"
 "imidazole+acetol-phosphate"                  "C6H7N2O5P1"
 "L-histidinol-phosphate"                      "C6H11N3O4P1"
 "histidinol"                                  "C6H12N3O1"
 reactions
 "ARA:GLUTAMIDOTRANS-RXN;ARA:PRIBFAICARPISOM-RXN"
 "ARA:AICARTRANSFORM-RXN;ARA:GLUTAMIDOTRANS-RXN;ARA:AICARSYN-RXN"
 "ARA:IMIDPHOSDEHYD-RXN;ARA:GLUTAMIDOTRANS-RXN"
 "ARA:HISTAMINOTRANS-RXN;ARA:IMIDPHOSDEHYD-RXN"
 "ARA:HISTIDPHOS-RXN;ARA:HISTAMINOTRANS-RXN"
 "ARA:HISTOLDEHYD-RXN;ARA:RXN-8001;ARA:HISTIDPHOS-RXN"

# In a similar way we can retrieve information from KEGG
keggdb <- read.table("http://rest.kegg.jp/link/compound/reaction")
head(keggdb)
         V1         V2
1 rn:R00001 cpd:C00001
2 rn:R00001 cpd:C00404
3 rn:R00001 cpd:C02174
4 rn:R00002 cpd:C00001
5 rn:R00002 cpd:C00002
6 rn:R00002 cpd:C00008
dim(keggdb)
[1] 38711     2
get.name(keggdb[1,2])
[1] "H2O"
get.formula.kegg(keggdb[1,2])
[1] "H2O"
get.name(keggdb[2,2])
[1] "Polyphosphate"
get.formula.kegg(keggdb[2,2])
[1] "H4P2O7(HPO3)n"
# Optionally the user can extract all compound information in the standard format
# with the KEGG organism code
# http://www.kegg.jp/kegg/catalog/org_list.html
system.time(ath <- build.database.kegg("ath"))


# We provide a formated KEGG database loaded with the package
# to perform the downstream analysis
# Database matching
DB <- KEGGcpds
ionAnnot <- get.annot(an, allowMiss=TRUE, minint=1000)
reactionM <- create.reactionM(DB, molIon=ionAnnot, ppm.tol=30)
wl <-weightM(reactionM, useIso=FALSE)
w <- design.connection(reactionM)

# Probability calculations
x <- 1:ncol(wl$wm)
y <- 1:nrow(wl$wm)
conn <- gibbs.samp(x, y, 5000, w, wl$wm)

# Output representation
system.time(ansConn <- export.class.table(conn, reactionM, ionAnnot, I
```

```
                filename="AnalysisExample",
                            html=TRUE,  m.test="t.test",
                            class1="F_I_Nature", class2="F_S_Nature",
                            DB=DB)
)

mw <- which(w==1,arr.ind=TRUE)
corList <- reac2cor(mw, ansConn$classTable, corths=0.7, corprob=0)
```

The analysis encompassed by *ProbMetab* package uses different sources of information to try to potentiate the understanding of dynamic changes in the metabolism. The **Figure 5** illustrates the partial view of a correlation weighted network, which summarizes information from peak ranking, metabolic pathway context and dynamic correlation changes, and allows the navigation and edition of the network in Cytoscape.

With the aid of such context visualization was possible to observe that 31 of detected mass peaks have between their possible identities retrieved from KEGG database, compounds that participate of Flavonoid biossynteses, a well known secondary metabolism pathway in plants. The alteration on Flavonoid levels is a known abiotic stress marker in plants, and was previously described to contain the generation of Reactive Oxygen Species (ROS), constituting a secondary ROS scavenging system in plants [42].

```
# creating and formatting a graph
# load("probmetab-case2-box01.RData") # load the necessary objects to draw the graph

classTable <- ansConn$classTable
gr.cor <- ftM2graphNEL(corList$cor.vs.reac)
node.names <- apply(classTable[classTable[,6]!="",1:7], 1, function(x) paste(x[6], "-",
paste(strsplit(as.matrix(x[7]), "#")[[1]][1:2], collapse="\n"), sep=""))
node.names <- sub("^\\s+", "", node.names)
node.names <- node.names[as.numeric(nodes(gr.cor))]

# color edges that represent correlations between mass peaks higher than
# 0.7 of red
form <- edgeNames(gr.cor)
form <- data.frame(form, form %in%
apply(corList$signif.cor[corList$signif.cor[,1]>0.70,2:3], 1, paste, collapse="~"))

# there is no negative correlation
form <- data.frame(form, form[,1] %in% apply(corList$signif.cor[corList$signif.cor[,1]<(-
0.70),2:3], 1, paste, collapse="~"))
cnames <- c("edge.name", "color.#FF0000", "color.#006400")
colnames(form) <- cnames

# color nodes representing differential representation of blue
pvec <- as.numeric(classTable[classTable[,5]!="",5])
form2 <- nodes(gr.cor)
form2 <- data.frame(form2, as.numeric(nodes(gr.cor)) %in% which(pvec<0.05))
cnames2 <- c("node.name", "lcolor.#FF0000")
colnames(form2) <- cnames2
```

```
# export the basic graph format do a quick web visualization
createJSONToCytoscape(gr=gr.cor, node.label=node.names)
openGraph("network.json", classTable=classTable, openBrowser=TRUE)

# format attribute tables to export to cytoscape visualization
cpdnames <- as.character(sapply(classTable[classTable[,2]!="unknown",2], function(x)
DB$name[DB$id==as.character(x)]))
classTable <- as.matrix(classTable)
classTable[classTable[,2]!="unknown",2] <- cpdnames

cpdInfo <- create.pathway.node.attributes(ansConn$classTable, graph=gr.cor, DB=DB,
    filename1="path1.noa", filename2="path2.noa", organismId="zma")

create.reaction.edge.attributes(classTable, graph=gr.cor, w=w, reactionM=reactionM, DB=DB,
filename="reac.eda")

# take care, there are not negative correlations, so the column 3 of form matrix is empty
export2cytoscape(gr.cor, node.label=node.names, cwName="test4", edge.form=form[,-3],
            node.form=form2, cpdInfo=cpdInfo, classTable=classTable)


# show correlations that changed in two groups of repeated samples

# water
corList1 <- reac2cor(mw, ansConn$classTable[,-c(25:39)], corths=0.7, corprob=0)

# drought
corList2 <- reac2cor(mw, ansConn$classTable[,-c(8:24)], corths=0.7, corprob=0)
gr.cor2 <- ftM2graphNEL(corList2$cor.vs.reac)
gr.cor1 <- ftM2graphNEL(corList1$cor.vs.reac)

mw1 <- t(sapply(edgeNames(gr.cor1), function(x) strsplit(x, "~")[[1]]))
mw2 <- t(sapply(edgeNames(gr.cor2), function(x) strsplit(x, "~")[[1]]))
mw3 <- unique(rbind(mw1, mw2))
gr.cor3 <- ftM2graphNEL(mw3)

inOne <- setdiff(edgeNames(gr.cor1), edgeNames(gr.cor2))
inTwo <- setdiff(edgeNames(gr.cor2), edgeNames(gr.cor1))

form <- edgeNames(gr.cor3)
form <- data.frame(form, form %in% inOne)
form <- data.frame(form, form[,1] %in% inTwo)

# Format and Normalize data do calculate fold change
metabData <- classTable[classTable[,6]!="",]
metabData2 <- apply(metabData[,8:39], 2, as.numeric)
rownames(metabData2) <- metabData[,6]
normalize.medFC <- function(mat) {
   # Perform median fold change normalisation
   #           X - data set [Variables & Samples]
   medSam <- apply(mat, 1, median)
   medSam[which(medSam==0)] <- 0.0001
   mat <- apply(mat, 2, function(mat, medSam){
      medFDiSmpl <- mat/medSam
      vec<-mat/median(medFDiSmpl)
      return(vec)
   }, medSam)
   return (mat)
}
metabData2 <- normalize.medFC(metabData2)

# Calculate differential correlations
# The DiffCorr package can be found at: http://diffcorr.sourceforge.net/
source("../DiffCorr_src/R/DiffCorr.R")
comp.2.cc.fdr(output.file="resM.txt", metabData2[,1:17], metabData2[,18:32], threshold=0.05)
```

```
res <- read.delim("resM.txt")

nres <- paste(res[,1], "~", res[,2], sep="")
form <- data.frame(form, form[,1] %in% nres)
# red for only in water
# green for only in drought
# width for differential correlation
cnames <- c("edge.name", "color.#FF0000", "color.#006400", "width.5")
colnames(form) <- cnames

node.names <- apply(classTable[classTable[,6]!="",1:7], 1, function(x) paste(x[6], "-",
paste(strsplit(as.matrix(x[7]), "#")[[1]][1:2], collapse="\n"), sep=""))
node.names <- sub("^\\s+", "", node.names)
node.names <- node.names[as.numeric(nodes(gr.cor3))]

pvec <- as.numeric(classTable[classTable[,5]!="",5])
foldChange <- apply(metabData2, 1, function(x) mean(x[18:32])/mean(x[1:17]))
colnames(classTable)[5] <- "Fold Change"
classTable[classTable[,5]!="",5] <- foldChange
form2 <- nodes(gr.cor3)
form2 <- data.frame(form2, as.numeric(nodes(gr.cor3)) %in% which(pvec<0.05))
cnames2 <- c("node.name", "lcolor.#9400D3")
colnames(form2) <- cnames2

cpdInfo <- create.pathway.node.attributes(ansConn$classTable, graph=gr.cor3, DB=DB,
filename1="path1Diff.noa", filename2="path2Diff.noa", organismId="zma")
create.reaction.edge.attributes(classTable, graph=gr.cor3, w=w, reactionM=reactionM, DB=DB,
filename="reacDiff.eda")

export2cytoscape(gr.cor3, node.label=node.names, cwName="test4", edge.form=form,
node.form=form2, cpdInfo=cpdInfo, classTable=classTable)
```

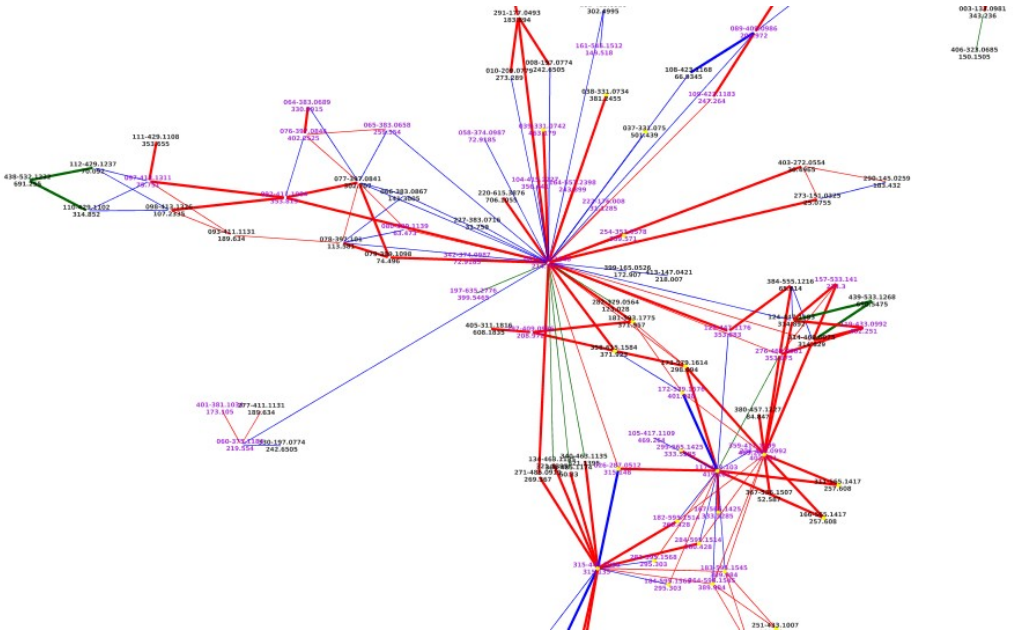


Figure 5 – Partial view of the overlaid reaction and weighted correlation network (absolute correlation value above 0.7). Nodes (in purple) represents mass peaks with mean intensity significantly different between standard watering and drought. Edges in red represents the correlations present only on standard watering condition, green only on drought condition and blue in both conditions. The node width indicates the difference between correlations is significant (thick) or not (thin).

Once one detects an interesting pathway, *ProbMetab* methods allow to export the visualization of KEGG pathway layout, making possible the inspection of the metabolic context where the putative compounds are inserted **Figure 6**, and with that a link with traditional pathway knowledge and representation, in a an environment where the user can edit the pathways and store a standard format that can later be used to modeling [43–45].

```
# see a specific pathway in a different window
classTableb <- ansConn$classTable
for(i in 1:nrow(classTableb)) if(classTableb[i,6]=="") classTableb[i,6] <- classTableb[i-
1,6]
classTableb[,6] <- as.numeric(classTableb[,6])
classTableS <- classTableb[which(classTableb[,6] %in% nodes(gr.cor)),]
kgr <- get.kgml.positions.kegg("rn00944")
cnames <- sapply(sub("^cpd:(C\\d{5}).*$", "\\1", colnames(kgr$adj)), get.name)
kgr1 <- as(kgr$adj, "graphNEL")
form <- data.frame(nodes(kgr1))
codes <- sub("^cpd:(C\\d{5}).*$", "\\1", colnames(kgr$adj))
form <- cbind(form, codes%in% classTableS[,2])
cnames2 <- c("node.name", "lcolor.#FF0000")
colnames(form) <- cnames2
export2cytoscape(kgr1, node.label=cnames, cwName="test2", node.form=form, pos=kgr)

# retrieve kegg version with ProbMetab
get.kegg.pathways(as.vector(form[form[,2],1]), 20)
```

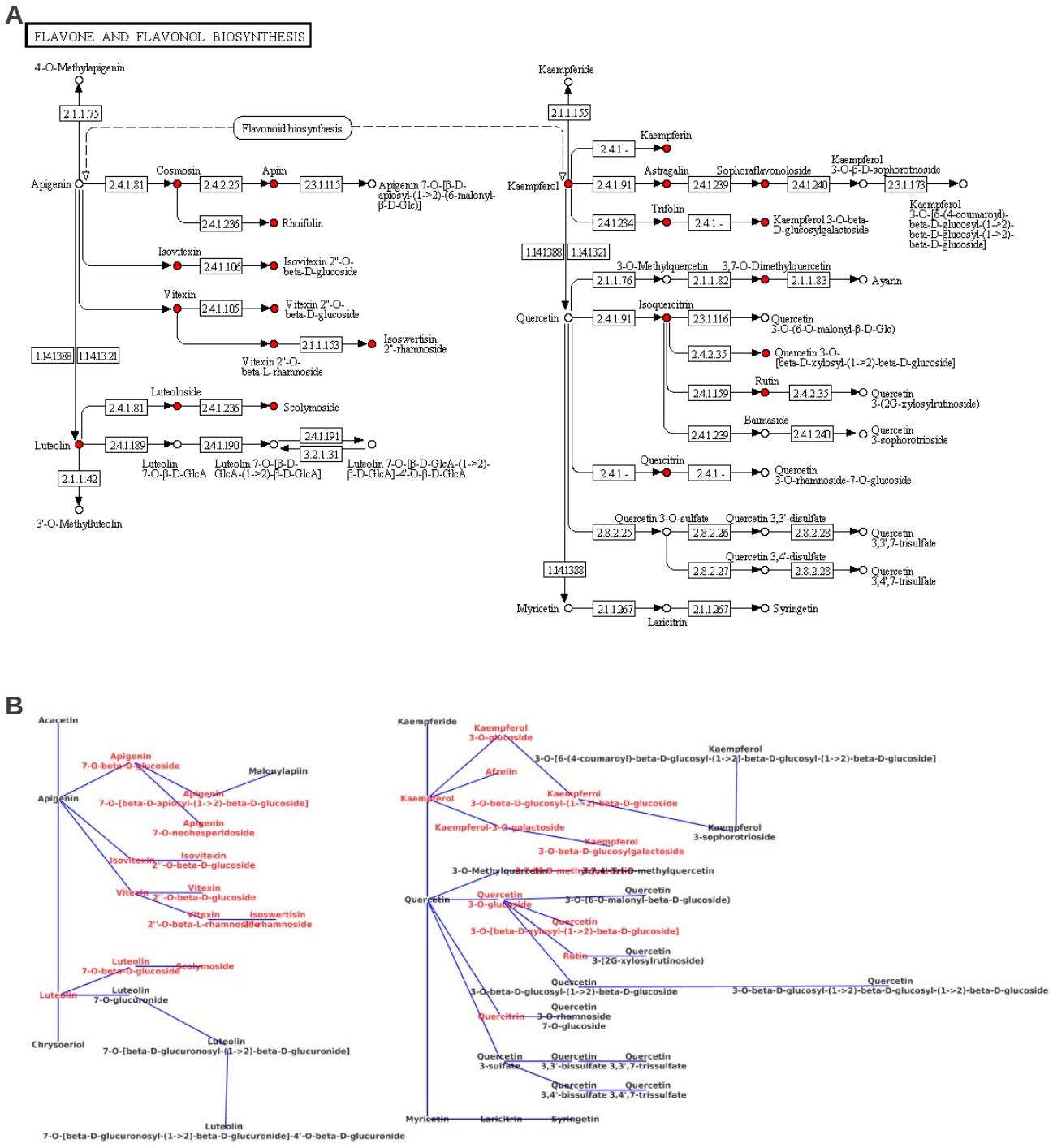


Figure 6 – Representations of a KEGG pathway map retrieved with *ProbMetab*. A – Conventional KEGG pathway map recovered through KEGG's API. B - Schematic representation of Flavone and flavonol biosynthesis pathway, automatically exported from KEGG to Cytoscape by *ProbMetab*, showing the 23 compounds that appear as putative identities in the overlaid reaction/weighted network.

Among the 31 annotated peaks, 18 present mean intensity significantly different between the

watering and drought conditions, p-value for t-test adapted to multiple testing < 0.05 [46]. Again, among the 31 peaks, 27 had an intensity decrease in drought conditions, Intensity ratio (intensity in drought/intensity in watering) < 1. Previous studies have shown that small metabolite intensity variations, from one to two times (Fold Change), are robust to experimental variation [47]. In the present experiment we observed the variation of maximum 5 times more intense on watering (putative mass 610.14) and maximum 1.3 times more intense on drought condition (putative mass 550.09).

Table – Partial view of probability tables exported by *ProbMetab*, showing in yellow the putative identities associated to Flavonoid biosynthesis.

Proposed Mass	Most Probable Compound Code	Most probable Compound	Probability	Entropy	Fold Change	Ion annotation	t-test pvalue	Compound/Pathways
286.0439463083	C01514	Luteolin	0.162	0.955	0.349	287.0512#315.148#[48][M]+#[M+H-C6H10O5]+ 448.093	0.000017476	C01514-00941;00944;01100;01110
	C05903	Kaempferol	0.147					C05903-00941;00944;01100;01110;01061
	C17786	6-Demethoxycapillarisin	0.076					
	C10510	Orobol	0.074					
	C10036	Datiscetin	0.071					
	C10097	Isoscutellarein	0.07					
	C08576	Aureusidin	0.068					
	C12134	2'-Hydroxygenistein	0.067					
	C10041	Fisetin	0.066					
	C10184	Scutellarein	0.066					
	C17810	Citreorosein	0.065					
	C08720	Maritimetin	0.058					
	C07359	Oxazepam	0.009					
330.0677718265	C17405	Penicillin O	0.202	0.969	0.736	331.075#501.439#[70][M]+#	0.3015851606	
	C01265	3',4',5-Trihydroxy-3,7-dimethoxyflavone	0.16					C01265-00944
	C10193	Tricin	0.142					
	C03036	(+)-Bisdechlorogeodin	0.09					
	C17670	Aurantio-obtusin	0.086					
	C03040	(-)-Bisdechlorogeodin	0.084					
	C10424	Hildecarpin	0.082					
	C10033	Cirsiliol	0.082					
	C16754	Aflatoxin G2	0.074					
330.0660866171	C17405	Penicillin O	0.395	0.864	0.905	331.0734#381.2455#[71][M]+#[M+Na]+ 308.088	0.7463242681	
	C01265	3',4',5-Trihydroxy-3,7-dimethoxyflavone	0.126					C01265-00944
	C10193	Tricin	0.112					
	C10033	Cirsiliol	0.069					
	C03036	(+)-Bisdechlorogeodin	0.065					
	C03040	(-)-Bisdechlorogeodin	0.064					
	C17670	Aurantio-obtusin	0.058					
	C10424	Hildecarpin	0.056					
	C16754	Aflatoxin G2	0.054					
330.0669000273	C17405	Penicillin O	0.266	0.938	0.752	331.0742#463.179#[72][M]+#[M+K]+ 292.11	0.0424501254	
	C01265	3',4',5-Trihydroxy-3,7-dimethoxyflavone	0.153					C01265-00944
	C10193	Tricin	0.14					
	C17670	Aurantio-obtusin	0.09					
	C10424	Hildecarpin	0.079					
	C03040	(-)-Bisdechlorogeodin	0.077					
	C03036	(+)-Bisdechlorogeodin	0.067					
	C16754	Aflatoxin G2	0.065					
	C10033	Cirsiliol	0.063					
432.0957515893	C06569	7a-Hydroxy-O-carbamoyl-deacetylcephalospc	0.728	0.5	0.523	433.103#419.102#[169][M]+#[M+H-C2H4]+ 460.13	0.0049787785	
	C04608	Apigenin 7-O-beta-D-glucoside	0.052					C04608-00944
	C01714	Apigenin-6-C-glucoside	0.052					C01714-00944
	C01460	Vitexin	0.046					C01460-00941;00944
	C09126	Genistein 7-O-beta-D-glucoside	0.03					
	C16911	Afzelin	0.022					C16911-00944
	C10420	Genistein 8-C-glucoside	0.022					
	C04609	Aerobacter aerogenes capsular polysacchari	0.017					
	C10345	Emodin 8-glucoside	0.016					
	C01715	Kallikrein	0.016					
432.0933965475	C06569	7a-Hydroxy-O-carbamoyl-deacetylcephalospc	0.952	0.129	0.965	433.1007#302.5615#[170][M]+#[M+H]+ 432.094	0.9982552761	
	C01460	Vitexin	0.01					C01460-00941;00944
	C04608	Apigenin 7-O-beta-D-glucoside	0.008					C04608-00944
	C01714	Apigenin-6-C-glucoside	0.007					C01714-00944
	C09126	Genistein 7-O-beta-D-glucoside	0.005					
	C10345	Emodin 8-glucoside	0.005					
	C16911	Afzelin	0.004					
	C10420	Genistein 8-C-glucoside	0.004					
	C01715	Kallikrein	0.004					C16911-00944
	C04609	Aerobacter aerogenes capsular polysacchari	0.003					
464.0853162616	C05623	Quercetin 3-O-glucoside	0.39	0.9	0.843	465.0926#212.5315#[202][M]+#[M+H]+ 464.088	0.16032	C05623-00944;01100;01110
	C16410	Bracteatin 6-O-glucoside	0.122					
	C10050	Gossypetin 8-rhamnoside	0.119					
	C10073	Hyperin	0.11					
	C12639	Quercimeritrin	0.102					
	C10108	Myricitrin	0.09					
	C06775	11-O-Demethylpradinone I	0.066					
564.1344538537	C04858	Apigenin 7-O-[beta-D-apiosyl-(1->2)-beta-D-gl	0.222	0.999	0.965	565.1417#257.608#[299][M]+#[M+H]+ 564.137	0.790713334	C04858-00944
	C16491	Isovitexin 2''-O-arabinoside	0.204					
	C16803	Glucofrangulin B	0.196					
	C10181	Schaftoside	0.194					
	C10110	Neoschaftoside	0.184					
564.1352009103	C04858	Apigenin 7-O-[beta-D-apiosyl-(1->2)-beta-D-gl	0.212	0.999	0.514	565.1425#333.5285#[300][M]+#[M+H]+ 564.136	0.0022811227	C04858-00944
	C16491	Isovitexin 2''-O-arabinoside	0.21					
	C10110	Neoschaftoside	0.206					
	C16803	Glucofrangulin B	0.188					
	C10181	Schaftoside	0.185					
578.1502992021	C17639	Procyanidin B2	0.19	0.91	0.444	579.1576#401.938#[317][M]+#	0.0002145892	
	C17640	Procyanidin B5	0.184					
	C10221	Epicatechin-(4beta->8)-ent-epicatechin	0.184					
	C10238	Procyanidin B4	0.166					
	C12628	Vitexin 2''-O-beta-L-rhamnoside	0.116					C12628-00944
	C12627	Apigenin 7-O-neohesperidoside	0.048					C12627-00944
	C16981	Kaempferitrin	0.038					
	C16802	Glucofrangulin A	0.038					
	C10196	Violanthin	0.036					

Proposed Mass		Most probable Compound	Probability	Entropy	Fold Change	Ion annotation	t-test pvalue	Compound/Pathways
578.1541496289	C12628	Vitexin 2"-O-beta-L-rhamnoside	0.343	0.894	0.797	579.1614#298.094#[318][M]+#[M+H-C6H10O5]+ 740.207	0.0804169189	C12628-00944
	C12627	Apigenin 7-O-neohesperidoside	0.14					C12627-00944
	C10196	Violanthin	0.109					
	C16802	Glucofrangulin A	0.1					
	C16981	Kaempferitrin	0.089					
	C17639	Procyanidin B2	0.066					
	C17640	Procyanidin B5	0.056					
	C10238	Procyanidin B4	0.049					
	C10221	Epicatechin-(4beta->8)-ent-epicatechin	0.048					
592.1702418137	C12629	7-O-Methylvitexin 2"-O-beta-L-rhamnoside	0.606	0.968	0.934	593.1775#371.957#[344][M]+#[M+H]+ 592.17	0.2482714845	C12629-00944
	C04275	1,2-Bis-O-sinapoyl-beta-D-glucoside	0.394					
594.1441488456	C17140	Tribuloside	0.483	0.804	0.525	595.1514#260.428#[348][M]+#[M+H]+ 594.147	0.0002277469	
	C04024	Vitexin 2"-O-beta-D-glucoside	0.074					C04024-00944
	C04199	Isovitexin 2"-O-beta-D-glucoside	0.07					C04199-00944
	C12630	Scolymoside	0.07					C12630-00944
	C10195	Vicenin-2	0.064					
	C08064	Saponarin	0.063					
	C10513	Paniculatin	0.062					
	C17600	Multiflorin B	0.057					
	C03870	Isoorientin 2"-O-rhamnoside	0.056					
594.1471744158	C17140	Tribuloside	0.144	0.996	0.441	595.1545#389.984#[349][M]+#[M+H]+ 594.148	0.0042995599	
	C12630	Scolymoside	0.118					C12630-00944
	C04024	Vitexin 2"-O-beta-D-glucoside	0.116					C04024-00944
	C10195	Vicenin-2	0.116					
	C08064	Saponarin	0.115					
	C04199	Isovitexin 2"-O-beta-D-glucoside	0.112					C04199-00944
	C17600	Multiflorin B	0.098					
	C03870	Isoorientin 2"-O-rhamnoside	0.094					
	C10513	Paniculatin	0.088					
594.1495186907	C04199	Isovitexin 2"-O-beta-D-glucoside	0.138	0.992	1.176	595.1568#295.303#[350][M]+#[M+H]+ 594.15	0.0237175346	C04199-00944
	C12630	Scolymoside	0.126					C12630-00944
	C10513	Paniculatin	0.122					
	C17600	Multiflorin B	0.118					
	C10195	Vicenin-2	0.112					
	C08064	Saponarin	0.11					
	C04024	Vitexin 2"-O-beta-D-glucoside	0.106					C04024-00944
	C03870	Isoorientin 2"-O-rhamnoside	0.106					
	C17140	Tribuloside	0.062					
596.1276060961	C12637	Quercetin 3-O-[beta-D-xylosyl-(1->2)-beta-D-g	1	0	0.946	597.1349#233.0755#[352][M]+#	0.9050104025	C12637-00944
610.138698153	C10227	Gallocatechin-(4alpha->8)-epigallocatechin	0.588	0.742	0.655	611.146#211.4575#[371][M]+#	0.0341713173	
	C05625	Rutin	0.12					C05625-00944;01100;01110
	C10102	Lucenin-2	0.078					
	C16490	Kaempferol 3-O-beta-D-glucosylgalactoside	0.076					C16490-00944
	C12634	Kaempferol 3-O-beta-D-glucosyl-(1->2)-beta-D	0.07					C12634-00944
	C17563	Multinoside A	0.068					
610.1476196543	C05625	Rutin	0.283	0.916	0.199	611.1549#288.3965#[372][M]+#[M+H+HCOOH]+ 564.139	0.023264914	C05625-00944;01100;01110
	C16490	Kaempferol 3-O-beta-D-glucosylgalactoside	0.188					C16490-00944
	C12634	Kaempferol 3-O-beta-D-glucosyl-(1->2)-beta-D	0.18					C12634-00944
	C17563	Multinoside A	0.169					
	C10102	Lucenin-2	0.165					
	C10227	Gallocatechin-(4alpha->8)-epigallocatechin	0.015					
550.0947069651	C12638	Quercetin 3-O-(6-O-malonyl-beta-D-glucoside	1	0	1.344	589.0586#35.237#[M+K]+ 550.095	0.0001756326	C12638-00944
432.0958380163	C06569	7a-Hydroxy-O-carbamoyl-deacetylcephalospo	0.716	0.514	0.707	433.0992#402.251#[171][M]+#[M+H]+ 432.096	0.0330061661	
	C04608	Apigenin 7-O-beta-D-glucoside	0.06					C04608-00944
	C01714	Apigenin-6-C-glucoside	0.057					C01714-00944
	C01460	Vitexin	0.044					C01460-00941;00944
	C09126	Genistein 7-O-beta-D-glucoside	0.031					
	C04609	Aerobacter aerogenes capsular polysaccharid	0.024					
	C16911	Afzelin	0.019					C16911-00944
	C10345	Emodin 8-glucoside	0.016					
	C10420	Genistein 8-C-glucoside	0.016					
	C01715	Kallikrein	0.016					
432.0942072809	C06569	7a-Hydroxy-O-carbamoyl-deacetylcephalospo	0.897	0.236	0.965	433.1007#302.5615#[170][M]+#[M+H]+ 432.094	0.9982552761	
	C01714	Apigenin-6-C-glucoside	0.022					C01714-00944
	C01460	Vitexin	0.019					C01460-00941;00944
	C04608	Apigenin 7-O-beta-D-glucoside	0.018					C04608-00944
	C09126	Genistein 7-O-beta-D-glucoside	0.01					C09126-00943
	C10345	Emodin 8-glucoside	0.008					
	C01715	Kallikrein	0.008					
	C10420	Genistein 8-C-glucoside	0.007					
	C16911	Afzelin	0.006					C16911-00944
	C04609	Aerobacter aerogenes capsular polysaccharid	0.005					
330.0667697971	C17405	Penicillin O	0.274	0.928	0.708	353.0578#259.571#[86][M]+#[M+Na]+ 330.067	0.0092964071	
	C01265	3',4',5-Trihydroxy-3,7-dimethoxyflavone	0.162					C01265-00944
	C10193	Tricin	0.15					
	C10033	Cirsiliol	0.075					
	C03040	(-)-Bisdechlorogeodin	0.072					
	C10424	Hildecarpin	0.07					
	C16754	Aflatoxin G2	0.068					
	C17670	Aurantio-obtusin	0.064					
	C03036	(+)-Bisdechlorogeodin	0.064					
594.1477667169	C04024	Vitexin 2"-O-beta-D-glucoside	0.122	0.999	0.441	595.1545#389.984#[349][M]+#[M+H]+ 594.148	0.0042995599	C04024-00944
	C12630	Scolymoside	0.122					C12630-00944
	C04199	Isovitexin 2"-O-beta-D-glucoside	0.118					C04199-00944
	C10513	Paniculatin	0.114					
	C17140	Tribuloside	0.112					
	C03870	Isoorientin 2"-O-rhamnoside	0.108					
	C08064	Saponarin	0.104					
	C17600	Multiflorin B	0.104					
	C10195	Vicenin-2	0.095					
594.1497401947	C12630	Scolymoside	0.14	0.984	1.176	595.1568#295.303#[350][M]+#[M+H]+ 594.15	0.0237175346	C12630-00944
	C04199	Isovitexin 2"-O-beta-D-glucoside	0.127					C04199-00944
	C04024	Vitexin 2"-O-beta-D-glucoside	0.12					C04024-00944
	C10195	Vicenin-2	0.12					
	C17600	Multiflorin B	0.118					
	C08064	Saponarin	0.115					
	C10513	Paniculatin	0.114					
	C03870	Isoorientin 2"-O-rhamnoside	0.107					
	C17140	Tribuloside	0.04					

Proposed Mass		Most probable Compound	Probability	Entropy	Fold Change	Ion annotation	t-test pvalue	Compound/Pathways
594.1471916604	C17140	Tribuloside	0.143	0.996	0.525	595.1514#260.428#[348][M]+#[M+H]+ 594.147	0.0002277469	
	C12630	Scolymoside	0.122					C12630-00944
	C04199	Isovitexin 2"-O-beta-D-glucoside	0.116					C04199-00944
	C08064	Saponarin	0.114					
	C04024	Vitexin 2"-O-beta-D-glucoside	0.11					C04024-00944
	C17600	Multiflorin B	0.106					
	C10513	Paniculatin	0.102					
	C10195	Vicenin-2	0.094					
	C03870	Isoorientin 2"-O-rhamnoside	0.094					
610.1451587718	C05625	Rutin	0.265	0.956	0.887	633.1316#211.4535#[394][M]+#[M+Na]+ 610.145	0.4343258201	C05625-00944;01100;01110
	C12634	Kaempferol 3-O-beta-D-glucosyl-(1->2)-beta-[	0.19					C12634-00944
	C16490	Kaempferol 3-O-beta-D-glucosylgalactoside	0.18					C16490-00944
	C10102	Lucenin-2	0.162					
	C17563	Multinoside A	0.15					
	C10227	Gallocatechin-(4alpha->8)-epigallocatechin	0.054					
564.136319278	C10181	Schaftoside	0.212	0.999	0.514	565.1425#333.5285#[300][M]+#[M+H]+ 564.136	0.0022811227	
	C16491	Isovitexin 2"-O-arabinoside	0.204					
	C16803	Glucofrangulin B	0.202					
	C04858	Apigenin 7-O-[beta-D-apiosyl-(1->2)-beta-D-gl	0.196					C04858-00944
	C10110	Neoschaftoside	0.187					
564.1366987421	C16491	Isovitexin 2"-O-arabinoside	0.215	0.999	0.965	565.1417#257.608#[299][M]+#[M+H]+ 564.137	0.790713334	
	C04858	Apigenin 7-O-[beta-D-apiosyl-(1->2)-beta-D-gl	0.209					C04858-00944
	C10110	Neoschaftoside	0.202					
	C10181	Schaftoside	0.193					
	C16803	Glucofrangulin B	0.18					
448.0933731465	C01821	Isoorientin	0.23	0.954	0.473	449.0995#315.135##[M+H]+ 448.093	7.24857998668504E-005	
	C03951	Luteolin 7-O-beta-D-glucoside	0.146					C03951-00944
	C12626	Kaempferol 3-O-beta-D-galactoside	0.104					C12626-00944
	C12249	Astragalin	0.1					C12249-00944;01110
	C16409	Aureusidin 6-O-glucoside	0.086					
	C01750	Quercitrin	0.076					C01750-00944
	C08598	Carthamone	0.073					
	C10042	Fisetin 8-C-glucoside	0.071					
	C10114	Orientin	0.063					
	C17056	Plantaginin	0.05					
464.0877596974	C05623	Quercetin 3-O-glucoside	0.408	0.842	0.843	465.0926#212.5315#[202][M]+#[M+H]+ 464.088	0.1603182746	C05623-00944;01100;01110
	C12639	Quercimeritrin	0.128					
	C10108	Myricitrin	0.121					
	C10050	Gossypetin 8-rhamnoside	0.118					
	C16410	Bracteatin 6-O-glucoside	0.115					C16410-00941
	C10073	Hyperin	0.108					
	C06775	11-O-Demethylpradinone I	0.002					
592.1698116346	C12629	7-O-Methylvitexin 2"-O-beta-L-rhamnoside	0.608	0.966	1.212	615.1584#371.923#[374][M]+#[M+Na]+ 592.17	0.078735621	C12629-00944
	C04275	1,2-Bis-O-sinapoyl-beta-D-glucoside	0.392					

Conclusions

New experiments are being carried out to obtain MS/MS spectras from putative Flavonoid peaks in order to confirm the annotated identities. Although many studies have shown at the transcriptional level, that the Flavonoid production is increased in response to drought [42], Yang et al. (2007) [43] have shown for *Glycyrrhiza inflata*, that, even though the production increase, the total content Flavonoid decrease. A possible explanation for this decrease, also observed in the present experiment, could be the change of Flavonoid observable form due its scavenging role in the drought condition.

The differential correlation analysis [49] can provide evidences of carbon flux changes under drought stress, once a correlation between two mass peaks present on normal condition and absent under stress can point to a new reaction happening. The understanding of carbon partition under stress conditions is essential to provide increments of cultivated plants. Hofmann & Jahufer, (2011) [50], have shown a negative correlation between Flavonoid production and drought mass accumulation in white clover genotypes, and suggest that Flavonoids can be used as biomarkers in breeding programs to control the tradeoff between production and stress tolerance.

The analysis automated in *ProbMetab* were able to recover changes in a known plant stress response metabolic pathway, showing its potential to unravel interesting mechanistic changes in the metabolism. Only a small fraction of the reaction network was analyzed, and the powerful bridge

between R and Cytoscape, with the condensed information provided by *ProbMetab*, allows to further explore the metabolic changes under stress conditions.

References

1. Fiehn O, Sumner LW, Rhee SY, Ward J, Dickerson J, Lange BM, Lane G, Roessner U, Last R, Nikolau B: **Minimum reporting standards for plant biology context information in metabolomic studies**. *Metabolomics* 2007, **3**:195–201.

2. Vincent IM, Creek DJ, Burgess K, Woods DJ, Burchmore RJS, Barrett MP: **Untargeted metabolomics reveals a lack of synergy between nifurtimox and eflornithine against Trypanosoma brucei.** *PLoS Negl. Trop. Dis.* 2012, **6**:e1618.

3. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak M-Y, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P: **A cross-platform toolkit for mass spectrometry and proteomics**. *Nat. Biotechnol.* 2012, **30**:918–920.

4. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G: **XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification.** *Anal. Chem.* 2006, **78**:779–87.

5. Kuhl C, Tautenhahn R, Bottcher C, Larson TR, Neumann S: **CAMERA: An integrated strategy for compound spectra extraction and annotation of LC/MS data sets.** *Anal. Chem.* 2011.

6. Scheltema RA, Jankevics A, Jansen RC, Swertz MA, Breitling R: **PeakML/mzMatch: a file format, Java library, R library, and tool-chain for mass spectrometry data analysis.** *Anal. Chem.* 2011, **83**:2786–93.

7. Patti GJ, Tautenhahn R, Siuzdak G: **Meta-analysis of untargeted metabolomic data from multiple profiling experiments.** *Nat. Protoc.* 2012, **7**:508–16.

8. Alonso A, Julià A, Beltran A, Vinaixa M, Díaz M, Ibañez L, Correig X, Marsal S: **AStream: An R package for annotating LC/MS metabolomic data.** *Bioinformatics* 2011, **27**:1339–1340.

9. Jankevics A, Merlo ME, de Vries M, Vonk RJ, Takano E, Breitling R: **Separating the wheat from the chaff: a prioritisation pipeline for the analysis of metabolomics datasets.** *Metabolomics* 2012, **8**:29–36.

10. Creek DJ, Jankevics A, Burgess KE V, Breitling R, Barrett MP: **IDEOM: An Excel interface for analysis of LC-MS based metabolomics data.** *Bioinformatics* 2012, **28**:1048–1049.

11. Pence HE, Williams A: **ChemSpider: an online chemical information resource**. *J. Chem. Educ.* 2010, **87**:1123–1124.

12. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH: **PubChem: a public information system for analyzing bioactivities of small molecules.** *Nucleic Acids Res.* 2009, **37**:W623–33.

13. Tautenhahn R, Cho K, Uritboonthai W, Zhu Z, Patti GJ, Siuzdak G: **An accelerated workflow for untargeted metabolomics using the METLIN database**. *Nat. Biotechnol.* 2012, **30**:826–828.

14. Kind T, Fiehn O: **Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm.** *BMC Bioinformatics* 2006, **7**:234.

15. Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Pujar A, Shearer AG, Travers M, Weerasinghe D, Zhang P, Karp PD: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases**. *Nucleic Acids Res.* 2011, **40**:D742–753.

16. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res.* 2000, **28**:27–30.

17. Kim TY, Sohn SB, Kim Y Bin, Kim WJ, Lee SY: **Recent advances in reconstruction and applications of genome-scale metabolic models.** *Curr. Opin. Biotechnol.* 2012, **23**:617–23.

18. Hucka M, Finney a., Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin a. P, Bornstein BJ, Bray D, Cornish-Bowden a., Cuellar a. a., Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr J-H, Hunter PJ, Juty NS, Kasberger JL, Kremling a., Kummer U, Le Novere N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J: **The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models**. *Bioinformatics* 2003, **19**:524–531.

19. Cottret L, Wildridge D, Vinson F, Barrett MP, Charles H, Sagot M-F, Jourdan F: **MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks.** *Nucleic Acids Res.* 2010, **38**:W132–7.

20. Barupal DK, Haldiya PK, Wohlgemuth G, Kind T, Kothari SL, Pinkerton KE, Fiehn O: **MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity**. *BMC Bioinformatics* 2012, **13**:99.

21. Weber RJM, Southam AD, Sommer U, Viant MR: **Characterization of isotopic abundance measurements in high resolution FT-ICR and Orbitrap mass spectra for improved confidence of metabolite identification.** *Anal. Chem.* 2011, **83**:3737–43.

22. Pluskal T, Uehara T, Yanagida M: **Highly accurate chemical formula prediction tool utilizing high-resolution mass spectra, MS/MS fragmentation, heuristic rules, and isotope pattern matching.** *Anal. Chem.* 2012.

23. Maechling CR, Clemett SJ, Zare RN: **13C/12C ratio measurements of aromatic molecules using photoionization with TOF mass spectrometry**. *Chem. Phys. Lett.* 1995, **241**:301–310.

24. Breitling R, Ritchie S, Goodenowe D, Stewart ML, Barrett MP: **Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data**. *Metabolomics* 2006, **2**:155–164.

25. Jourdan F, Breitling R, Barrett MP, Gilbert D: **MetaNetter: inference and visualization of high-resolution metabolomic networks.** *Bioinformatics* 2008, **24**:143–5.

26. Creek DJ, Jankevics A, Breitling R, Watson DG, Barrett MP, Burgess KE V: **Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography-Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction.** *Anal. Chem.* 2011.

27. Miller AJ: **Selection of Subsets of Regression Variables**. *J. R. Stat. Soc. Ser. A* 1984, **147**:pp. 389–425.

28. Kaliszan R: **QSRR: quantitative structure-(chromatographic) retention relationships.** *Chem.*

Rev. 2007, **107**:3212–46.

29. Allwood J, Weber R, Zhou J, He S, Viant M, Dunn W: **CASMI—The Small Molecule Identification Process from a Birmingham Perspective**. *Metabolites* 2013, **3**:397–411.

30. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res.* 2003, **13**:2498–504.

31. Shannon PT, Grimes M, Kutlu B, Bot JJ, Galas DJ: **RCytoscape: tools for exploratory network analysis**. *BMC Bioinformatics* 2013, **14**:217.

32. Oliver S: **Systematic functional analysis of the yeast genome**. *Trends Biotechnol.* 1998, **16**:373–378.

33. Beltran A, Suarez M, Rodríguez MA, Vinaixa M, Samino S, Arola L, Correig X, Yanes O: **Assessment of compatibility between extraction methods for NMR- and LC/MS-based metabolomics.** *Anal. Chem.* 2012, **84**:5838–44.

34. Steinbeck C, Conesa P, Haug K, Mahendraker T, Williams M, Maguire E, Rocca-Serra P, Sansone S-A, Salek RM, Griffin JL: **MetaboLights: towards a new COSMOS of metabolomics data management**. *Metabolomics* 2012.

35. Steuer R: **Review: on the analysis and interpretation of correlations in metabolomic data.** *Brief. Bioinform.* 2006, **7**:151–8.

36. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ: **Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data.** *BMC Syst. Biol.* 2011, **5**:21.

37. Moore P: **Temporal and Spatial Regulation of Sucrose Accumulation in the Sugarcane Stem**. *Aust. J. Plant Physiol.* 1995, **22**:661.

38. RAE A, GROF C, CASU R, BONNETT G: **Sucrose accumulation in the sugarcane stem: pathways and control points for transport and compartmentation**. *F. Crop. Res.* 2005, **92**:159–168.

39. Lisec J, Meyer RC, Steinfath M, Redestig H, Becher M, Witucka-Wall H, Fiehn O, Törjék O, Selbig J, Altmann T, Willmitzer L: **Identification of metabolic and biomass QTL in Arabidopsis thaliana in a parallel analysis of RIL and IL populations.** *Plant J.* 2008, **53**:960–72.

40. Ji Y, Hebbring S, Zhu H, Jenkins GD, Biernacka J, Snyder K, Drews M, Fiehn O, Zeng Z, Schaid D, Mrazek DA, Kaddurah-Daouk R, Weinshilboum RM: **Glycine and a glycine dehydrogenase (GLDC) SNP as citalopram/escitalopram response biomarkers in depression: pharmacometabolomics-informed pharmacogenomics.** *Clin. Pharmacol. Ther.* 2011, **89**:97–104.

41. De Vos RCH, Moco S, Lommen A, Keurentjes JJB, Bino RJ, Hall RD: **Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry.** *Nat. Protoc.* 2007, **2**:778–91.

42. Pandey N, Ranjan A, Pant P, Tripathi RK, Ateek F, Pandey HP, Patre U V, Sawant S V: **CAMTA 1 regulates drought responses in Arabidopsis thaliana.** *BMC Genomics* 2013, **14**:216.

43. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, Pico AR: **WikiPathways: building research communities on biological pathways.** *Nucleic Acids Res.* 2011:gkr1074–.

44. Bourqui R, Cottret L, Lacroix V, Auber D, Mary P, Sagot M-F, Jourdan F: **Metabolic network visualization eliminating node redundance and preserving metabolic pathways.** *BMC Syst. Biol.* 2007, **1**:29.

45. Jourdan F, Cottret L, Huc L, Wildridge D, Scheltema R, Hillenweck A, Barrett MP, Zalko D, Watson DG, Debrauwer L: **Use of reconstituted metabolic networks to assist in metabolomic data visualization and mining.** *Metabolomics* 2010, **6**:312–321.

46. Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S: *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer New York; 2005, **746718470**.

47. Vinaixa M, Samino S, Saez I, Duran J, Guinovart JJ, Yanes O: **A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data**. *Metabolites* 2012, **2**:775–795.

48. Yang Y, He F, Yu L, Chen X, Lei J, Ji J: **Influence of drought on oxidative stress and flavonoid production in cell suspension culture of Glycyrrhiza inflata Batal.** *Z. Naturforsch. C.* , **62**:410–6.

49. Fukushima A: **DiffCorr: an R package to analyze and visualize differential correlations in biological networks.** *Gene* 2013, **518**:209–14.

50. Hofmann RW, Jahufer MZZ: **Tradeoff between biomass and flavonoid accumulation in white clover reflects contrasting plant strategies.** *PLoS One* 2011, **6**:e18949.