EUFLOW:

Expected Utility Workflow Evaluation Package

Kevin K. McDade

September 9, 2016

I.Introduction

The data in bioinformatics is often in some "raw" form which is not the intended analysis goal. Processing of this data often involves a multistep process which to as a workflow, pipeline, or protocol. One major obstacle to data reproducibility is workflows are rarely identical and the differences between workflow steps may not be the optimal way to process the data. We define a workflow here as a series of steps that a user takes to arrive at the analysis goal. The magnitude of the path choices can be considered to be a great benefit to the analyst, however, it is often the case that the user does not know which path to take.

In regards to bioinformatics workflow options can include ready-to-go data (a finished data set), custom workflows (user decides what steps to take), or use of a tuning parameter (i.e. requiring a certain level of data quality). Any change in a workflow step or change of a parameter setting constitutes a new workflow option. To what extent do these choices affect the final dataset to be analyzed? If the datasets differ substantially, will they differ in quality? If so, how can we tell which is best? Finally, will soundness of the scientific conclusions be harmed by worse workflow options? Surprisingly little is known.

The contents of this package are:

```
search()
```

```
##  [1] ".GlobalEnv"                "package:EUFLOW"
##  [3] "package:IdMappingAnalysis" "package:rChoiceDialogs"
##  [5] "package:rJava"             "package:R.oo"
##  [7] "package:R.methodsS3"       "package:stats"
##  [9] "package:graphics"          "package:grDevices"
## [11] "package:utils"             "package:datasets"
## [13] "package:methods"           "Autoloads"
## [15] "package:base"
```

```
ls(pos=2)
```

```
##  [1] "EvaluationExperimentSet"    "expectedUtility"
##  [3] "fit2clusters.workflow"      "make.workflow.map"
##  [5] "merge_tag_options"          "Model.quality.list"
##  [7] "RANDOMSET"                  "RNASEQ_PLUS_RANDOM"
##  [9] "RNASEQDATA"                 "RPPADATA.original"
## [11] "Workflow.Criterion"         "Workflow.Evaluation.table"
## [13] "Workflow.posteriorestimate" "WorkflowEvaluationData"
```

We now test the package using How-touse-this-package.Rmd contents.

Users are required to input two separate data files, regardless of the analysis type. The EUFLOW package depends upon a model quality heuristic which we will demonstate here as correlation between gene expression and protein expression. The first data file, for example, we will use ovarian TCGA RNASeq data on two

separate workflows provided on the same samples. Only samples for which protein expression data was available are used in the RNASEQDATA file.

A small slice of this data file (RNASEQDATA) demonstrates the data structure of numbered row names followed by a column of identifiers including a tag _v1 and _v2. The same samples were processed with two different workflow options RnaSeqv1 and RnaSeqv2. The output below shows the gene expression values for the first 9 identifiers (in this case gene names) with a tag to distinguish workflow option. For brevity only 4 of the 198 sample identifiers are represented.

```
RNASEQDATA<-read.csv(file="../data/RNASEQDATA.csv",header=TRUE)
RNASEQDATA[1:9,1:5]
```

```
##           X TCGA.04.1348 TCGA.04.1357 TCGA.04.1362 TCGA.04.1514
## 1  ACACA_v1       2.7465       1.7817       3.1331       7.7577
## 2   AKT1_v1      59.0196      56.3896      35.2372      53.8846
## 3   AKT2_v1      38.4897      20.8841      26.0161      19.7833
## 4   AKT3_v1       1.1253       1.3015       0.3474       1.5601
## 5  ANXA1_v1      77.4788     120.5428     303.1342       8.0768
## 6     AR_v1       0.8918       1.8014       1.1988       4.6952
## 7    BAX_v1      20.9399      26.1528      10.2788      12.9121
## 8   BCL2_v1       0.6704       3.9945       0.8274       8.8969
## 9 BCL2L1_v1      82.5929      85.8818      70.9154      53.7558
```

```
RNASEQDATA[68:77,1:5]
```

```
##             X TCGA.04.1348 TCGA.04.1357 TCGA.04.1362 TCGA.04.1514
## 68   ACACA_v2     9.464835     8.707251     9.558908    10.713177
## 69    AKT1_v2    12.544332    12.433390    11.711924    12.155235
## 70    AKT2_v2    12.525374    11.598174    11.863642    11.323640
## 71    AKT3_v2     6.889721     7.060301     4.974580     7.203402
## 72   ANXA1_v2    11.953121    12.550434    13.832331     8.456068
## 73      AR_v2     6.812444     7.832033     7.125823     8.970759
## 74     BAX_v2     8.916460     9.251187     7.777588     7.953272
## 75    BCL2_v2     6.596938     9.205781     6.749019    10.135617
## 76  BCL2L1_v2    12.295913    12.308044    11.964698    11.380043
## 77 BCL2L11_v2    10.157993    10.397630    10.602247    10.940016
```

In the second data file we will use TCGA protein expression data on the sampe samples. As above we will output only a subset. Alos note that as this is the reference dataset in this example only one identifier is represented rather than multiple workflow options.

```
RPPADATA<-read.csv(file="../data/RPPADATA.original.csv",header=TRUE)

RPPADATA[1:9,1:5]
```

```
##        X TCGA.04.1348 TCGA.04.1357 TCGA.04.1362 TCGA.04.1514
## 1  ACACA     0.137025    -1.878242    -0.043323    -0.337896
## 2   AKT1     0.164359     0.893065    -1.453180     0.620281
## 3   AKT2     0.164359     0.893065    -1.453180     0.620281
## 4   AKT3     0.164359     0.893065    -1.453180     0.620281
## 5  ANXA1    -0.169000     0.096700     1.540100    -2.791800
## 6     AR    -0.359340     0.277150    -0.466700     0.397730
```

```
## 7    BAX     0.011804     0.726132    -0.494385    -1.037562
## 8   BCL2    -0.704400     1.398200    -0.870200     1.851100
## 9 BCL2L1     0.358680     1.733370     1.612330    -1.059790
```

We now have all that is required to run the EUFLOW package. For the purpose of clarity we define the RNASEQDATA set as the EvaluationExperimentSet, which includes 2 workflow options (RnaSeqV1 and RnaSeqv2). We also define the RPPADATA as the ReferenceSet. If it is the choice of the user multiple reference sets can be utilized.

```
EvaluationExperimentSet<-RNASEQDATA
ReferenceSet<-RPPADATA
```

Now that we have the data for our example, the WorkflowEvaluationData function will modify the separate dataframes into one data structure to prepare to calcuate the model quality and perform the evaluation. The first item in the list is the Reference data and the second item in the list is the evaluation data.

```
Workflow.Data<-WorkflowEvaluationData(EvaluationExperimentSet,ReferenceSet)
Workflow.Data[[1]][1:9,1:5]
```

```
##        Symbol TCGA.04.1348 TCGA.04.1357 TCGA.04.1362 TCGA.04.1514
## ACACA   ACACA     0.137025    -1.878242    -0.043323    -0.337896
## AKT1     AKT1     0.164359     0.893065    -1.453180     0.620281
## AKT2     AKT2     0.164359     0.893065    -1.453180     0.620281
## AKT3     AKT3     0.164359     0.893065    -1.453180     0.620281
## ANXA1   ANXA1    -0.169000     0.096700     1.540100    -2.791800
## AR         AR    -0.359340     0.277150    -0.466700     0.397730
## BAX       BAX     0.011804     0.726132    -0.494385    -1.037562
## BCL2     BCL2    -0.704400     1.398200    -0.870200     1.851100
## BCL2L1 BCL2L1     0.358680     1.733370     1.612330    -1.059790
```

Further data processing to determine the number of workflow options and structure tags will be used to name the output by creating a Merged.options object.

```
Merged.options<-merge_tag_options(Workflow.Data)
Merged.options[1:9,1:5]
```

```
##               Symbol TCGA.04.1348 TCGA.04.1357 TCGA.04.1362 TCGA.04.1514
## ACACA_DRIVER   ACACA     0.137025    -1.878242    -0.043323    -0.337896
## AKT1_DRIVER     AKT1     0.164359     0.893065    -1.453180     0.620281
## AKT2_DRIVER     AKT2     0.164359     0.893065    -1.453180     0.620281
## AKT3_DRIVER     AKT3     0.164359     0.893065    -1.453180     0.620281
## ANXA1_DRIVER   ANXA1    -0.169000     0.096700     1.540100    -2.791800
## AR_DRIVER         AR    -0.359340     0.277150    -0.466700     0.397730
## BAX_DRIVER       BAX     0.011804     0.726132    -0.494385    -1.037562
## BCL2_DRIVER     BCL2    -0.704400     1.398200    -0.870200     1.851100
## BCL2L1_DRIVER BCL2L1     0.358680     1.733370     1.612330    -1.059790
```

The Model.quality.object is created to create a map between the Reference ids and the evaluation ids using the Model.quality.list function.

```
Model.quality.object<-Model.quality.list(Merged.options)
```

Next, Model Quality is an object which contains the model quality values for each of the pairs. How the Model quality is determined is specified by the user.

```
Model.Quality<-Workflow.Criterion(Model.quality.object,method="pearson")
head(as.data.frame(Model.Quality))
```

```
##        drivers workflow_options_merged     pearson
## 1 ACACA_DRIVER          ACACA_WFO_RS_1  0.54909334
## 2 ACACA_DRIVER          ACACA_WFO_RS_2  0.55458298
## 3  AKT1_DRIVER           AKT1_WFO_RS_1  0.62910048
## 4  AKT1_DRIVER           AKT1_WFO_RS_2  0.59568345
## 5  AKT2_DRIVER           AKT2_WFO_RS_1 -0.07874835
## 6  AKT2_DRIVER           AKT2_WFO_RS_2 -0.06027781
```

```
Model.Quality<-Workflow.Criterion(Model.quality.object,method="spearman")
head(as.data.frame(Model.Quality))
```

```
##        drivers workflow_options_merged    spearman
## 1 ACACA_DRIVER          ACACA_WFO_RS_1  0.53894840
## 2 ACACA_DRIVER          ACACA_WFO_RS_2  0.56452003
## 3  AKT1_DRIVER           AKT1_WFO_RS_1  0.52934338
## 4  AKT1_DRIVER           AKT1_WFO_RS_2  0.54777270
## 5  AKT2_DRIVER           AKT2_WFO_RS_1 -0.04651546
## 6  AKT2_DRIVER           AKT2_WFO_RS_2 -0.04524932
```

```
Posterior.dataframe<-Workflow.posteriorestimate(Model.quality.object,Model.Quality)
```

```
## Performing bootstrap R= 200 on correlations...
##
processed: 1 %

processed: 1 %

processed: 2 %

processed: 3 %

processed: 4 %

processed: 4 %

processed: 5 %

processed: 6 %

processed: 7 %

processed: 7 %
```

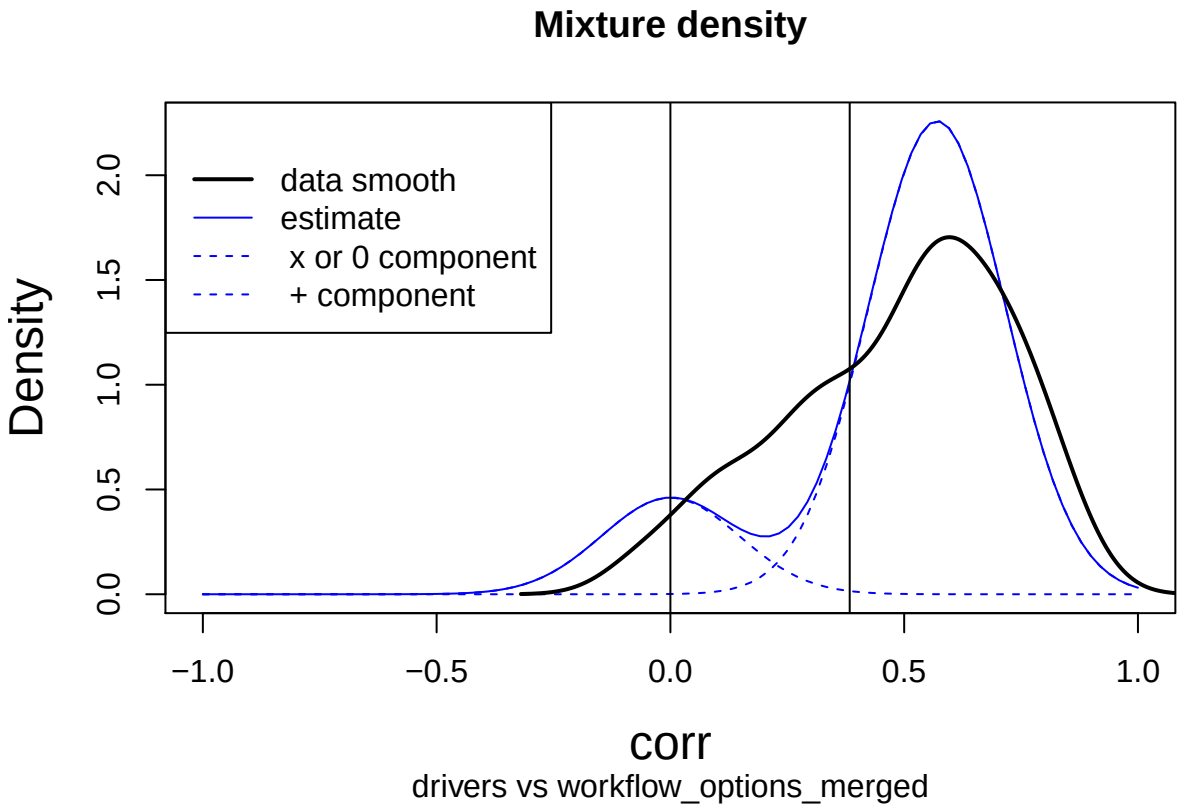
```
processed: 8 %

processed: 9 %

processed: 10 %

processed: 10 %

processed: 11 %

processed: 12 %

processed: 13 %

processed: 13 %

processed: 14 %

processed: 15 %

processed: 16 %

processed: 16 %

processed: 17 %

processed: 18 %

processed: 19 %

processed: 19 %

processed: 20 %

processed: 21 %

processed: 22 %

processed: 22 %

processed: 23 %

processed: 24 %

processed: 25 %

processed: 25 %

processed: 26 %

processed: 27 %

processed: 28 %
```

```
processed: 28 %
processed: 29 %
processed: 30 %
processed: 31 %
processed: 31 %
processed: 32 %
processed: 33 %
processed: 34 %
processed: 34 %
processed: 35 %
processed: 36 %
processed: 37 %
processed: 37 %
processed: 38 %
processed: 39 %
processed: 40 %
processed: 40 %
processed: 41 %
processed: 42 %
processed: 43 %
processed: 43 %
processed: 44 %
processed: 45 %
processed: 46 %
processed: 46 %
processed: 47 %
processed: 48 %
```

```
processed: 49 %
processed: 49 %
processed: 50 %
processed: 51 %
processed: 51 %
processed: 52 %
processed: 53 %
processed: 54 %
processed: 54 %
processed: 55 %
processed: 56 %
processed: 57 %
processed: 57 %
processed: 58 %
processed: 59 %
processed: 60 %
processed: 60 %
processed: 61 %
processed: 62 %
processed: 63 %
processed: 63 %
processed: 64 %
processed: 65 %
processed: 66 %
processed: 66 %
processed: 67 %
processed: 68 %
```

```
processed: 69 %
processed: 69 %
processed: 70 %
processed: 71 %
processed: 72 %
processed: 72 %
processed: 73 %
processed: 74 %
processed: 75 %
processed: 75 %
processed: 76 %
processed: 77 %
processed: 78 %
processed: 78 %
processed: 79 %
processed: 80 %
processed: 81 %
processed: 81 %
processed: 82 %
processed: 83 %
processed: 84 %
processed: 84 %
processed: 85 %
processed: 86 %
processed: 87 %
processed: 87 %
processed: 88 %
```

```
processed: 89 %

processed: 90 %

processed: 90 %

processed: 91 %

processed: 92 %

processed: 93 %

processed: 93 %

processed: 94 %

processed: 95 %

processed: 96 %

processed: 96 %

processed: 97 %

processed: 98 %

processed: 99 %

processed: 99 %

processed: 100 %
## 36 .  Converged.
```




```
Workflow.Evaluation.table(Posterior.dataframe)
```

```
##           label Utp Lfp deltaPlus nPairs    PrPlus    PrTrue    PrFalse
## Use All Use All   1   1         1    133 0.9552771 0.9552771 0.04472292
## 1             1   1   1         1     67 0.9400748 0.9400748 0.05992517
## 2             2   1   1         1     66 0.9707970 0.9707970 0.02920299
##             Utrue     Lfalse Eutility1  Eutility
## Use All 0.9552771 0.04472292 0.9105542 121.10370
## 1       0.9400748 0.05992517 0.8801497  58.97003
## 2       0.9707970 0.02920299 0.9415940  62.14521
```

```
Workflow.Evaluation.table(Posterior.dataframe,deltaPlus = 2)
```

```
##           label Utp Lfp deltaPlus nPairs    PrPlus    PrTrue   PrFalse
## Use All Use All   1   1         2    133 0.9552771 0.4776385 0.5223615
## 1             1   1   1         2     67 0.9400748 0.4700374 0.5299626
## 2             2   1   1         2     66 0.9707970 0.4853985 0.5146015
##             Utrue    Lfalse   Eutility1  Eutility
## Use All 0.4776385 0.5223615 -0.04472292 -5.948148
## 1       0.4700374 0.5299626 -0.05992517 -4.014987
## 2       0.4853985 0.5146015 -0.02920299 -1.927397
```

```
Workflow.Evaluation.table(Posterior.dataframe,Utp=3)
```

```
##           label Utp Lfp deltaPlus nPairs    PrPlus    PrTrue    PrFalse
## Use All Use All   3   1         1    133 0.9552771 0.9552771 0.04472292
## 1             1   3   1         1     67 0.9400748 0.9400748 0.05992517
## 2             2   3   1         1     66 0.9707970 0.9707970 0.02920299
##             Utrue     Lfalse Eutility1 Eutility
## Use All 2.865831 0.04472292  2.821108 375.2074
## 1       2.820224 0.05992517  2.760299 184.9401
## 2       2.912391 0.02920299  2.883188 190.2904
```

```
Workflow.Evaluation.table(Posterior.dataframe,Utp=1)
```

```
##           label Utp Lfp deltaPlus nPairs    PrPlus    PrTrue    PrFalse
## Use All Use All   1   1         1    133 0.9552771 0.9552771 0.04472292
## 1             1   1   1         1     67 0.9400748 0.9400748 0.05992517
## 2             2   1   1         1     66 0.9707970 0.9707970 0.02920299
##             Utrue     Lfalse Eutility1  Eutility
## Use All 0.9552771 0.04472292 0.9105542 121.10370
## 1       0.9400748 0.05992517 0.8801497  58.97003
## 2       0.9707970 0.02920299 0.9415940  62.14521
```