ChemModLab Full Dataset Test and New Functions

Jeremy Ash

October 18, 2016

Preparing Data: All descriptor sets

```
yfilein <- read.csv("AID_364.csv")
# xfilein <- read.csv("BurdenNumbers.csv")
# data <- cbind(yfilein, xfilein[,-1])
# head(data[1:6])

data <- yfilein
desc_lengths <- c()
for(desc in c("BurdenNumbers.csv","Pharmacophores.csv","AtomPairs.csv",
              "FragmentPairs.csv","Carharts.csv")){
  d <- read.csv(desc)[-1]
  data <- cbind(data, d)
  desc_lengths <- c(desc_lengths, ncol(d))
}
desc_idx <- list()
desc_idx[[1]] <- 1:desc_lengths[1]
for(i in 2:length(desc_lengths)){
  l1 <- desc_idx[[i-1]][length(desc_idx[[i-1]])]
  l2 <- desc_lengths[i]
  desc_idx[[i]] <- (l1+1):(l1+l2)
}
for(i in 1:length(desc_idx)){
  desc_idx[[i]] <- desc_idx[[i]] + 2
}

head(data[1:6])
```

```
##        CID Outcome WBN_GC_L_0.25 WBN_GC_H_0.25 WBN_GC_L_0.50 WBN_GC_H_0.50
## 1  5388992       1      -2.40010       1.98339      -2.52864       2.50835
## 2  5388983       1      -2.40010       1.98240      -2.52868       2.50398
## 3   663143       1      -2.41650       1.32890      -2.53910       2.05778
## 4    10607       1      -2.38337       2.17677      -2.52643       2.33232
## 5  5388972       1      -2.29039       1.97468      -2.41743       2.46177
## 6 11970251       1      -2.29039       2.22488      -2.41748       2.56161
```

```
ncol(data)
```

```
## [1] 6116
```

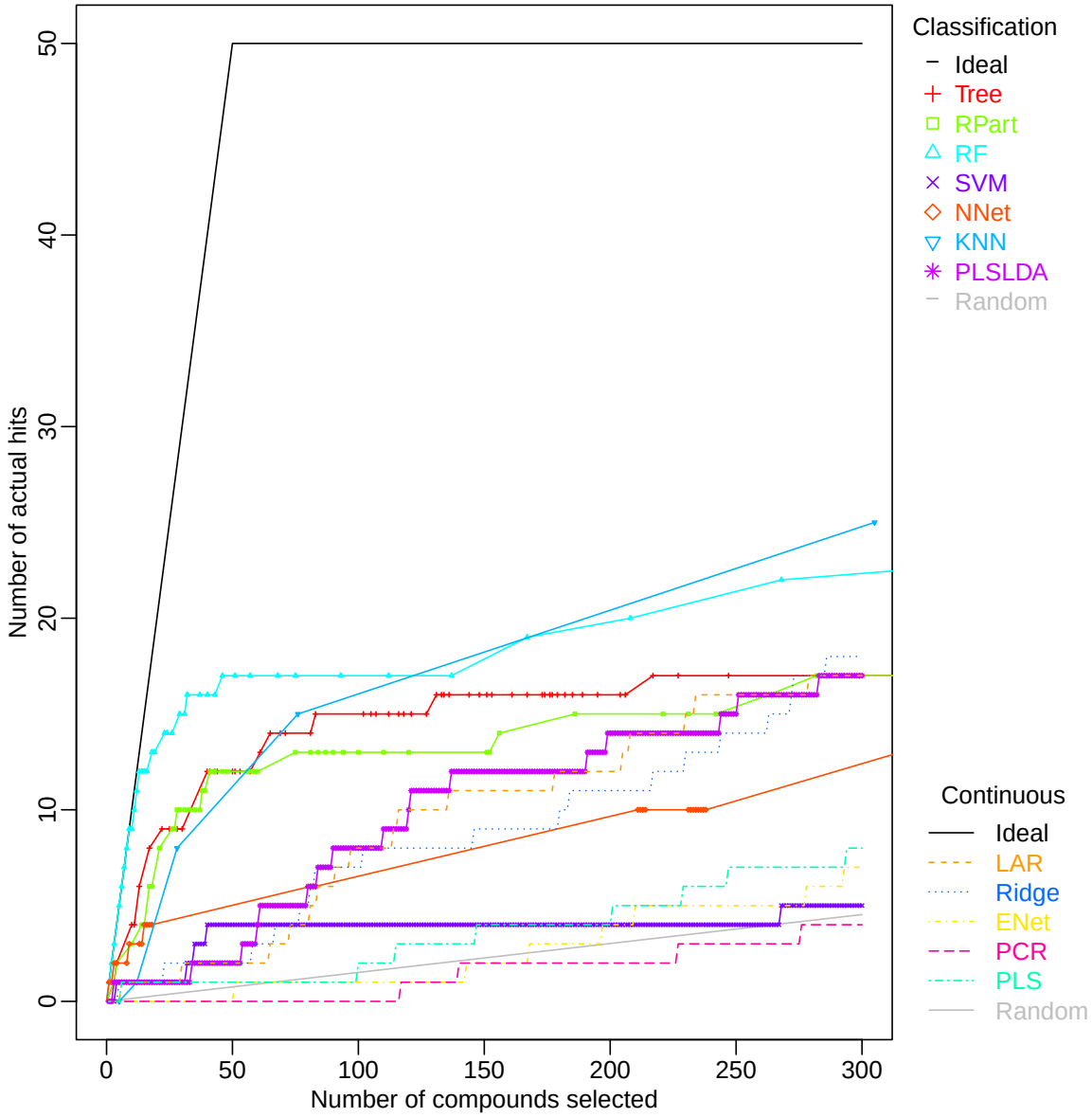
```
source("../background_test2.R")

# bb <- ModelTrain(data, idcol=1,
#                  models = c("NNet","PCR","ENet","PLS","Ridge","LARs",
```

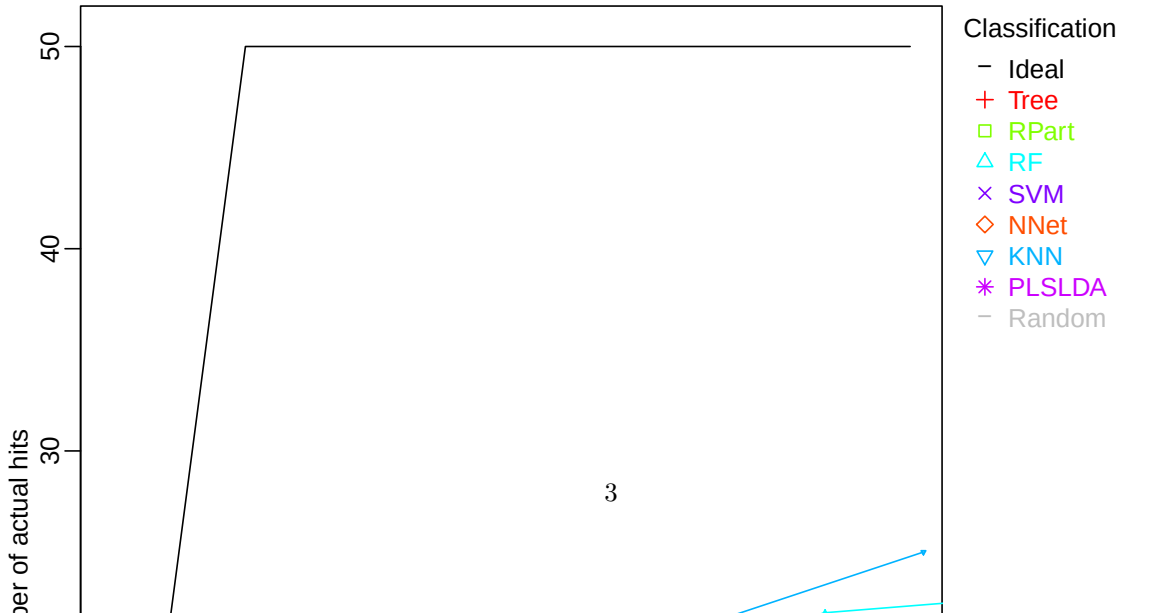
```
#                               "PLSLDA","RPart","Tree","SVM","KNN","Forest"),
#                   xcols = desc_idx, nsplits = 3, des.names =
#                     c("BurdenNumbers","Pharmacophores","AtomPairs","FragmentPairs","Carharts"),
#                   nfolds=10, seed.in=c(11111,22222,33333))

load("fullrun.RData")
results <- bb

format(object.size(results), units="Mb")
```

```
## [1] "47.6 Mb"
```

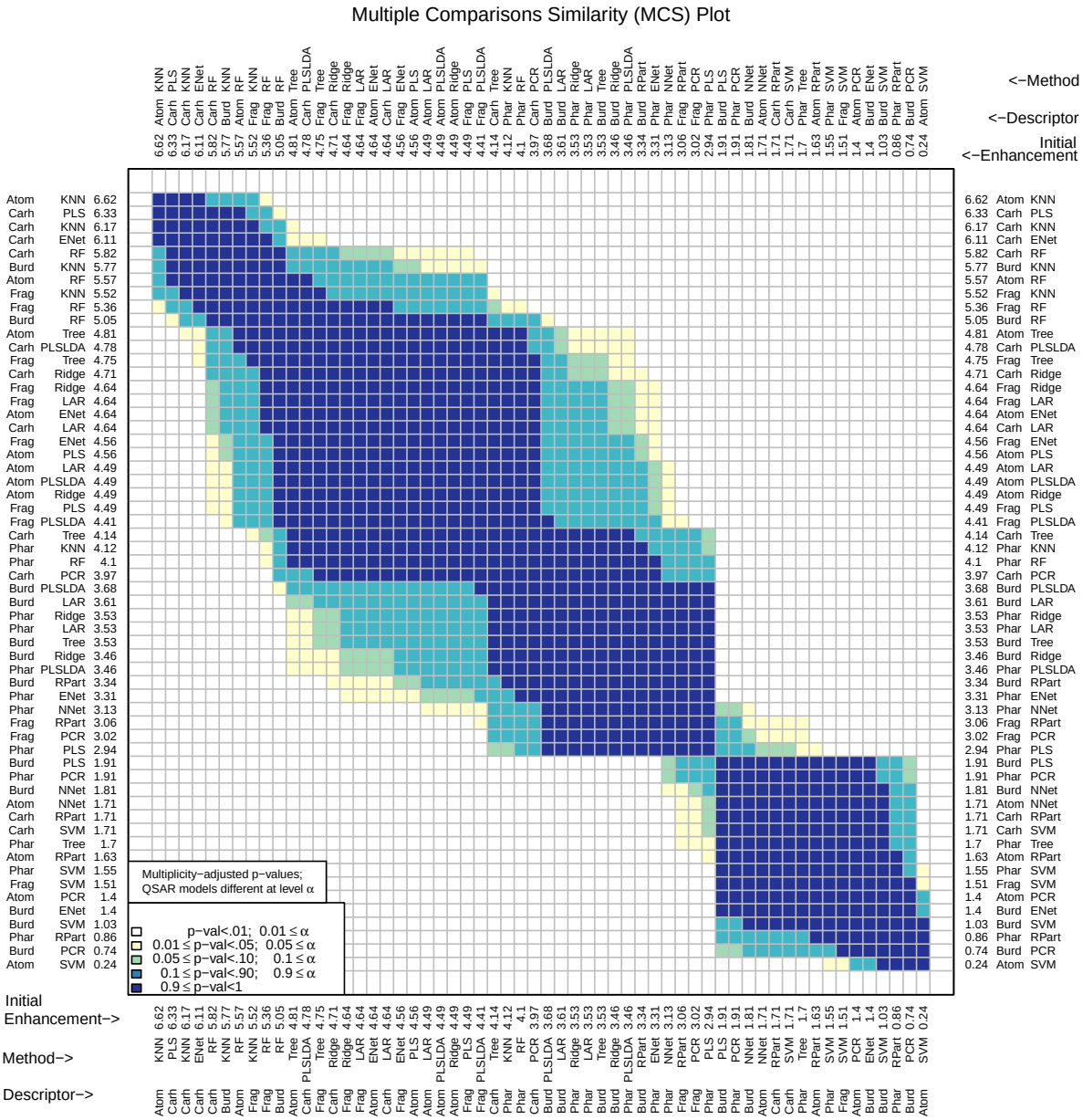
Old Code with 2008 packages

Installing the most current packages at the time the original ChemModLab output was generated.


Number of actual hits
0
10
20
30
40
50
0
50
100
150
200
250
300
Number of compounds selected
Classification
Ideal
Tree
RPart
RF
SVM
NNet
KNN
PLSLDA
Random
Continuous
Ideal
LAR
Ridge
ENet
PCR
PLS
Random


Number of actual hits
30
40
50
Classification
Ideal
Tree
RPart
RF
SVM
NNet
KNN
PLSLDA
Random

```
##    Analysis of Variance on Initial Enhancement @ 300
##  Using factors: Split and Descriptor/Method combination
## Source     DF          SS          MS           F   p-value
## Model      59    448.3911      7.5998     57.3784    <.0001
## Error     114     15.0995      0.1325
## Total     173    463.4906
##       R-Square    Coef Var    Root MSE        Mean
##         0.9674     10.0431      0.3639      3.6238
## Source       DF          SS          MS          F   p-value
## Split         2       2.341       1.170      8.836     3e-04
## Desc/Meth    57     446.050       7.825     59.082    <.0001
```



Data: Burden Number Descriptors and Active/Inactive Response

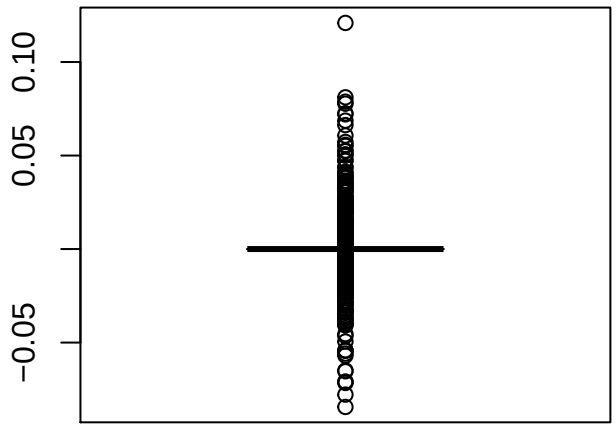
```
pred1 <- read.csv("../../ChemModLab_old/Split1/pred (2014_07_23 17_49_31 UTC).csv"
                  , skip = 1, row.names = 1)
pred2 <- read.csv("../../ChemModLab_old/Split2/pred (2014_07_23 17_49_31 UTC).csv"
                  , skip = 1, row.names = 1)
pred3 <- read.csv("../../ChemModLab_old/Split3/pred (2014_07_23 17_49_31 UTC).csv"
                  , skip = 1, row.names = 1)

pred_old <- list()
pred_old[[1]] <- list(pred1[,2:13],pred1[,grep("\\.1",colnames(pred1))],
                      pred1[,grep("\\.2",colnames(pred1))],pred1[,grep("\\.3",colnames(pred1))],
                      pred1[,grep("\\.4",colnames(pred1))])
pred_old[[2]] <- list(pred2[,2:13],pred2[,grep("\\.1",colnames(pred2))],
                      pred2[,grep("\\.2",colnames(pred2))],pred2[,grep("\\.3",colnames(pred2))],
                      pred2[,grep("\\.4",colnames(pred2))])
pred_old[[3]] <- list(pred3[,2:13],pred3[,grep("\\.1",colnames(pred3))],
                      pred3[,grep("\\.2",colnames(pred3))],pred3[,grep("\\.3",colnames(pred3))],
                      pred3[,grep("\\.4",colnames(pred3))])

for(i in 1:3){
  for(j in 1:5){
    colnames(pred_old[[i]][[j]]) <- sub("\\.1", "",colnames(pred_old[[i]][[j]]))
    colnames(pred_old[[i]][[j]]) <- sub("\\.2", "",colnames(pred_old[[i]][[j]]))
    colnames(pred_old[[i]][[j]]) <- sub("\\.3", "",colnames(pred_old[[i]][[j]]))
    colnames(pred_old[[i]][[j]]) <- sub("\\.4", "",colnames(pred_old[[i]][[j]]))
    rownames(pred_old[[i]][[j]]) <- as.character(rownames(pred_old[[i]][[j]]))
  }
}

desc <- c("BurdenNumbers","Pharmacophores","AtomPairs","FragmentPairs","Carharts")

for(i in 1:3){
  for(j in 1:5){
    cat(paste0("\nSplit ",i," Descriptor Set: ", desc[j],"\n"))
    print(all.equal(bb$all.preds[[i]][[j]][,-1], pred_old[[i]][[j]]))
#     print(head(bb$all.preds[[i]][[j]][,-1]))
#     print(head(pred_old[[i]][[j]]))
  }
}
```

```
##
## Split 1 Descriptor Set: BurdenNumbers
## [1] "Component \"LAR\": Mean relative difference: 9.138816e-08"
## [2] "Component \"Ridge\": Mean relative difference: 9.361895e-08"
## [3] "Component \"ENet\": Mean relative difference: 1.125892e-07"
## [4] "Component \"PCR\": Mean relative difference: 1.603592e-07"
## [5] "Component \"PLS\": Mean relative difference: 1.074425e-07"
##
## Split 1 Descriptor Set: Pharmacophores
## [1] "Component \"LAR\": Mean relative difference: 9.384864e-08"
## [2] "Component \"Ridge\": Mean relative difference: 9.531583e-08"
## [3] "Component \"ENet\": Mean relative difference: 9.458268e-08"
```

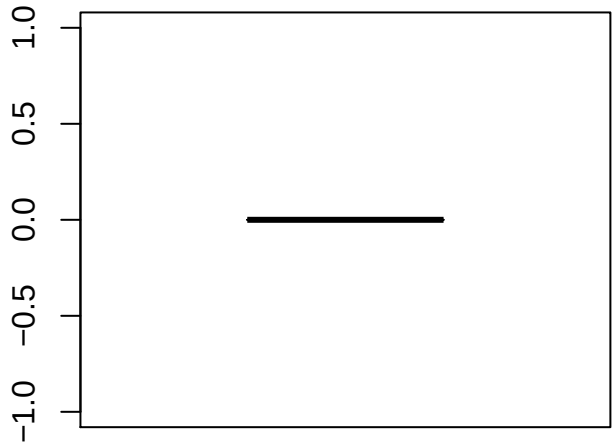
```
## [4] "Component \"PCR\": Mean relative difference: 1.259553e-07"
## [5] "Component \"PLS\": Mean relative difference: 9.268394e-08"
##
## Split 1 Descriptor Set: AtomPairs
## [1] "Component \"LAR\": Mean relative difference: 9.400163e-08"
## [2] "Component \"Ridge\": Mean relative difference: 0.05015466"
## [3] "Component \"ENet\": Mean relative difference: 8.980585e-08"
## [4] "Component \"PCR\": Mean relative difference: 1.2234e-07"
## [5] "Component \"PLS\": Mean relative difference: 9.589835e-08"
##
## Split 1 Descriptor Set: FragmentPairs
## [1] TRUE
##
## Split 1 Descriptor Set: Carharts
## [1] TRUE
##
## Split 2 Descriptor Set: BurdenNumbers
## [1] "Component \"LAR\": Mean relative difference: 9.298265e-08"
## [2] "Component \"Ridge\": Mean relative difference: 9.286989e-08"
## [3] "Component \"ENet\": Mean relative difference: 1.100751e-07"
## [4] "Component \"PCR\": Mean relative difference: 1.598104e-07"
## [5] "Component \"PLS\": Mean relative difference: 1.051322e-07"
##
## Split 2 Descriptor Set: Pharmacophores
## [1] "Component \"LAR\": Mean relative difference: 9.465777e-08"
## [2] "Component \"Ridge\": Mean relative difference: 9.498522e-08"
## [3] "Component \"ENet\": Mean relative difference: 9.397921e-08"
## [4] "Component \"PCR\": Mean relative difference: 1.268782e-07"
## [5] "Component \"PLS\": Mean relative difference: 9.980835e-08"
##
## Split 2 Descriptor Set: AtomPairs
## [1] "Component \"LAR\": Mean relative difference: 8.632387e-08"
## [2] "Component \"Ridge\": Mean relative difference: 8.588506e-08"
## [3] "Component \"ENet\": Mean relative difference: 8.845162e-08"
## [4] "Component \"PCR\": Mean relative difference: 1.164711e-07"
## [5] "Component \"PLS\": Mean relative difference: 9.057669e-08"
##
## Split 2 Descriptor Set: FragmentPairs
## [1] TRUE
##
## Split 2 Descriptor Set: Carharts
## [1] TRUE
##
## Split 3 Descriptor Set: BurdenNumbers
## [1] "Component \"LAR\": Mean relative difference: 9.085901e-08"
## [2] "Component \"Ridge\": Mean relative difference: 9.417634e-08"
## [3] "Component \"ENet\": Mean relative difference: 1.104423e-07"
## [4] "Component \"PLS\": Mean relative difference: 1.069732e-07"
##
## Split 3 Descriptor Set: Pharmacophores
## [1] "Component \"LAR\": Mean relative difference: 9.186979e-08"
## [2] "Component \"Ridge\": Mean relative difference: 9.303742e-08"
## [3] "Component \"ENet\": Mean relative difference: 9.322317e-08"
## [4] "Component \"PLS\": Mean relative difference: 9.636422e-08"
```

```
##
## Split 3 Descriptor Set: AtomPairs
## [1] "Component \"LAR\": Mean relative difference: 8.620801e-08"
## [2] "Component \"Ridge\": Mean relative difference: 9.052535e-08"
## [3] "Component \"ENet\": Mean relative difference: 8.859883e-08"
## [4] "Component \"PCR\": Mean relative difference: 1.222492e-07"
## [5] "Component \"PLS\": Mean relative difference: 8.54949e-08"
##
## Split 3 Descriptor Set: FragmentPairs
## [1] TRUE
##
## Split 3 Descriptor Set: Carharts
## [1] TRUE
```

```
# Ridge regression results in different predictions for Split 1 Atom Pairs?
boxplot(bb$all.preds[[1]][[3]][,"Ridge"] - pred_old[[1]][[3]][,"Ridge"])
```




```
# no longer any differences in RF?
boxplot(bb$all.preds[[i]][[j]][,"RF"] - pred_old[[i]][[j]][,"RF"])
```



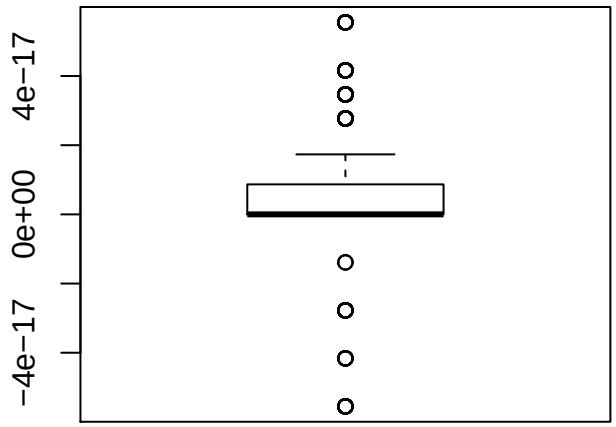

```
prob1 <- read.csv("../../ChemModLab_old/Split1/prob (2014_07_23 17_49_31 UTC).csv",
                  skip = 1, row.names = 1)
prob2 <- read.csv("../../ChemModLab_old/Split2/prob (2014_07_23 17_49_31 UTC).csv",
                  skip = 1, row.names = 1)
prob3 <- read.csv("../../ChemModLab_old/Split3/prob (2014_07_23 17_49_31 UTC).csv",
                  skip = 1, row.names = 1)

prob_old <- list()
prob_old[[1]] <- list(prob1[,2:8],prob1[,grep("\\.1",colnames(prob1))],
                      prob1[,grep("\\.2",colnames(prob1))],prob1[,grep("\\.3",colnames(prob1))],
                      prob1[,grep("\\.4",colnames(prob1))])
prob_old[[2]] <- list(prob2[,2:8],prob2[,grep("\\.1",colnames(prob2))],
                      prob2[,grep("\\.2",colnames(prob2))],prob2[,grep("\\.3",colnames(prob2))],
                      prob2[,grep("\\.4",colnames(prob2))])
prob_old[[3]] <- list(prob3[,2:8],prob3[,grep("\\.1",colnames(prob3))],
                      prob3[,grep("\\.2",colnames(prob3))],prob3[,grep("\\.3",colnames(prob3))],
                      prob3[,grep("\\.4",colnames(prob3))])

for(i in 1:3){
  for(j in 1:5){
    colnames(prob_old[[i]][[j]]) <- sub("\\.1", "",colnames(prob_old[[i]][[j]]))
    colnames(prob_old[[i]][[j]]) <- sub("\\.2", "",colnames(prob_old[[i]][[j]]))
    colnames(prob_old[[i]][[j]]) <- sub("\\.3", "",colnames(prob_old[[i]][[j]]))
    colnames(prob_old[[i]][[j]]) <- sub("\\.4", "",colnames(prob_old[[i]][[j]]))
    rownames(prob_old[[i]][[j]]) <- as.character(rownames(prob_old[[i]][[j]]))
  }
}

desc <- c("BurdenNumbers","Pharmacophores","AtomPairs","FragmentPairs","Carharts")
```

```
for(i in 1:3){
  for(j in 1:5){
    cat(paste0("\nSplit ",i," Descriptor Set: ", desc[j],"\n"))
    print(all.equal(bb$all.probs[[i]][[j]][,-1], prob_old[[i]][[j]]))
#     print(head(bb$all.probs[[i]][[j]][,-1]))
#     print(head(prob_old[[i]][[j]]))
  }
}
```

```
##
## Split 1 Descriptor Set: BurdenNumbers
## [1] "Component \"SVM\": Mean relative difference: 0.001881219"
##
## Split 1 Descriptor Set: Pharmacophores
## [1] "Component \"SVM\": Mean relative difference: 0.0005316252"
##
## Split 1 Descriptor Set: AtomPairs
## [1] "Component \"SVM\": Mean relative difference: 0.0002962165"
##
## Split 1 Descriptor Set: FragmentPairs
## [1] "Component \"SVM\": Mean relative difference: 0.0005473334"
##
## Split 1 Descriptor Set: Carharts
## [1] "Component \"SVM\": Mean relative difference: 0.0005134851"
##
## Split 2 Descriptor Set: BurdenNumbers
## [1] "Component \"SVM\": Mean relative difference: 0.001647672"
##
## Split 2 Descriptor Set: Pharmacophores
## [1] "Component \"SVM\": Mean relative difference: 0.0008089722"
##
## Split 2 Descriptor Set: AtomPairs
## [1] "Component \"SVM\": Mean relative difference: 0.0003991254"
##
## Split 2 Descriptor Set: FragmentPairs
## [1] "Component \"SVM\": Mean relative difference: 0.001046245"
##
## Split 2 Descriptor Set: Carharts
## [1] "Component \"SVM\": Mean relative difference: 0.0005919026"
##
## Split 3 Descriptor Set: BurdenNumbers
## [1] "Component \"SVM\": Mean relative difference: 0.001893699"
##
## Split 3 Descriptor Set: Pharmacophores
## [1] "Component \"SVM\": Mean relative difference: 0.0004224082"
##
## Split 3 Descriptor Set: AtomPairs
## [1] "Component \"SVM\": Mean relative difference: 0.0004903599"
##
## Split 3 Descriptor Set: FragmentPairs
## [1] "Component \"SVM\": Mean relative difference: 0.0005455603"
##
## Split 3 Descriptor Set: Carharts
```

```
## [1] "Component \"SVM\": Mean relative difference: 0.0004986973"
```

```
# no substantial differences in RF anymore
boxplot(bb$all.probs[[i]][[j]][,"RF"] - prob_old[[i]][[j]][,"RF"])
```



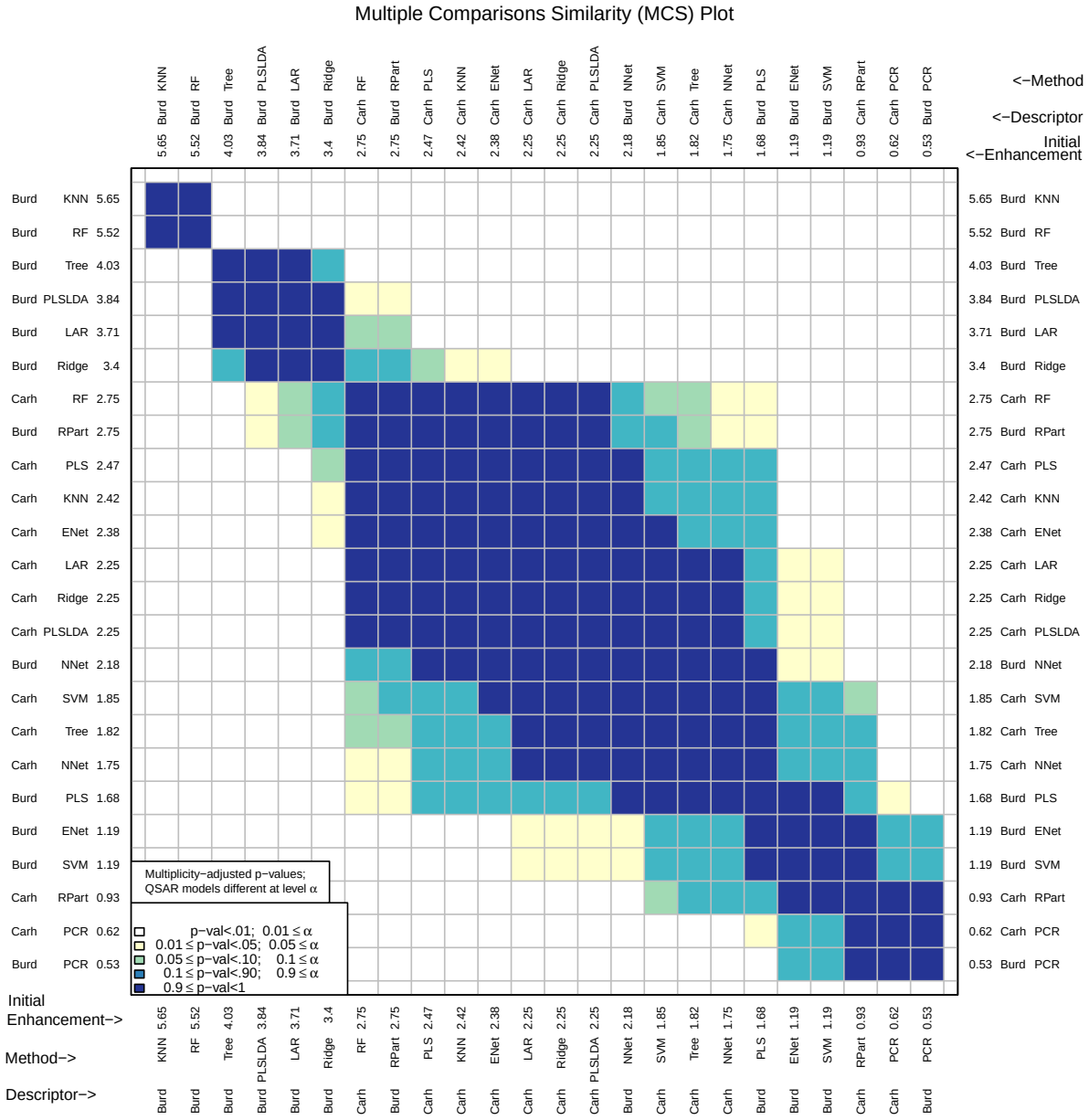
New Functions of ChemModLab

I have changed the ChemModLab code so that it takes any descriptor set, allows you to specify the names of the descriptor sets, and will flexibly incorporate new methods. Previously the set of methods, descriptor sets, and number of splits were assumed in the code. I have tested that the analyses are still working properly and that the labels on the plots appropriatey reflect the changes I have made.

```
setwd("C:/Users/Vestige/Dropbox/ChemModLab/example_run/")

yfilein <- read.csv("AID_364.csv")
xfilein1 <- read.csv("BurdenNumbers.csv")
xfilein2 <- read.csv("Carharts.csv")[, 1:26]
data <- cbind(yfilein, xfilein1[,-1], xfilein2[,-1])

source("../background_test2_name_change.R")

# bb <- ModelTrain(data, idcol=1,
#                  models = c("NNet","PCR","ENet","PLS","Ridge",
#                             "Lasso","PLSLDA","RPart","Tree","SVM","KNN","Forest"),
#                  xcols = list(seq(3,25+3),seq(25+4,ncol(data))),
#                  des.names = c("Burden Numbers", "Carharts"),
#                  nsplits = 5, nfolds=10, seed.in=c(12,34,56,78,910))
```

```
load("5split_run.RData")
```

```
##    Analysis of Variance on Initial Enhancement @ 300
##  Using factors: Split and Descriptor/Method combination
## Source     DF          SS          MS           F   p-value
## Model      27   208.9218     7.7378    46.5028    <.0001
## Error      92    15.3084     0.1664
## Total     119   224.2302
##       R-Square   Coef Var   Root MSE       Mean
##         0.9317    16.4805     0.4079     2.4751
## Source       DF         SS         MS         F   p-value
## Split         4      2.269      0.567     3.409    0.0276
## Desc/Meth    23    206.653      8.985    53.997    <.0001
```



The treatments and blocks are being assigned properly when the number of splits are increased. The descriptor set names are being set properly. The seeds are being set properly. I have also tested that when I change the name of a method the treatments are assigned properly and the label in the plots is correct.

```
source("../background_test_name_change.R")
out <- CombineSplits(result)
head(out)
```

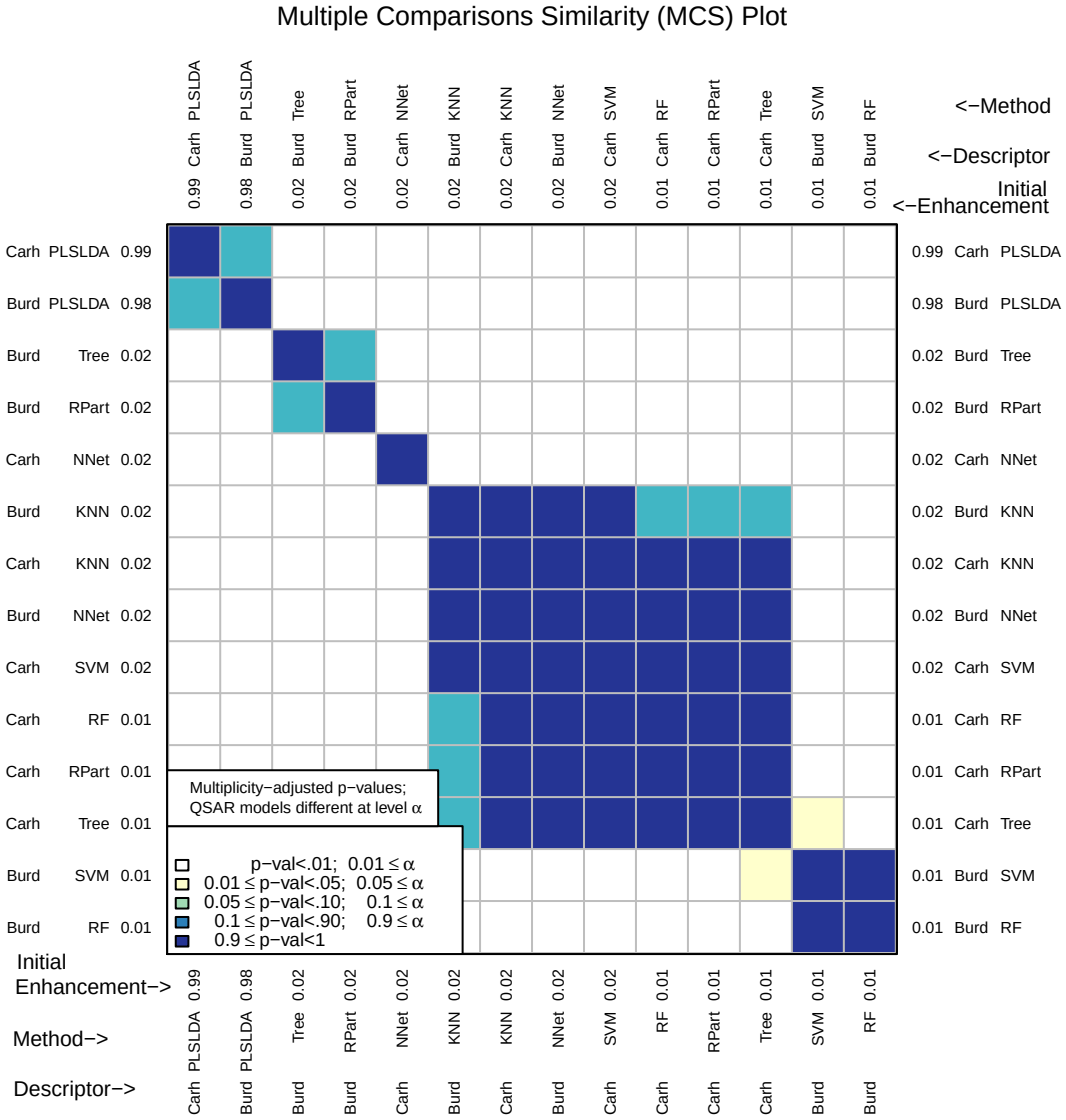
```
##   Split Descriptor Method     E300 Trmt
## 1     1       Burd   Tree 4.312298  101
## 2     1       Burd  RPart 2.851139  102
## 3     1       Burd     RF 5.534100  103
## 4     1       Burd    SVM 1.103667  104
## 5     1       Burd   NNet 2.536451  105
## 6     1       Burd    KNN 6.081156  106
```

```
tail(out)
```

```
##     Split Descriptor Method      E300 Trmt
## 178     5       Burd    PLS 1.5451333  112
## 186     5       Carh    LAR 2.2073333  208
## 187     5       Carh  Ridge 2.2073333  209
## 188     5       Carh   ENet 2.2073333  210
## 189     5       Carh    PCR 0.4414667  211
## 190     5       Carh    PLS 2.4280667  212
```

```
source("../background_test_new_summary.R")
CombineSplits(result,metric="error rate")
```

```
##    Analysis of Variance on Initial Enhancement @ 300
##  Using factors: Split and Descriptor/Method combination
## Source    DF          SS          MS           F   p-value
## Model     17   8.058e+00   4.740e-01   2.264e+06    <.0001
## Error     52   1.089e-05   2.094e-07
## Total     69   8.058e+00
##       R-Square   Coef Var   Root MSE       Mean
##       0.999999   0.296906   0.000458   0.154105
## Source       DF         SS         MS          F   p-value
## Split         4   4.25e-07   1.06e-07   5.07e-01    0.6824
## Desc/Meth    13   8.06e+00   6.20e-01   2.96e+06    <.0001
```


Multiple Comparisons Similarity (MCS) Plot
<−Method
<−Descriptor
Initial
<−Enhancement
Carh PLSLDA 0.99
Burd PLSLDA 0.98
Burd Tree 0.02
Burd RPart 0.02
Carh NNet 0.02
Burd KNN 0.02
Carh KNN 0.02
Burd NNet 0.02
Carh SVM 0.02
Carh RF 0.01
Carh RPart 0.01
Carh Tree 0.01
Burd SVM 0.01
Burd RF 0.01
Multiplicity−adjusted p−values;
QSAR models different at level α
p−val<.01; 0.01 ≤ α
0.01 ≤ p−val<.05; 0.05 ≤ α
0.05 ≤ p−val<.10; 0.1 ≤ α
0.1 ≤ p−val<.90; 0.9 ≤ α
0.9 ≤ p−val<1
Initial
Enhancement−>
Method−>
Descriptor−>