# Bayesian vs Frequentist Accumulation Curve CIs

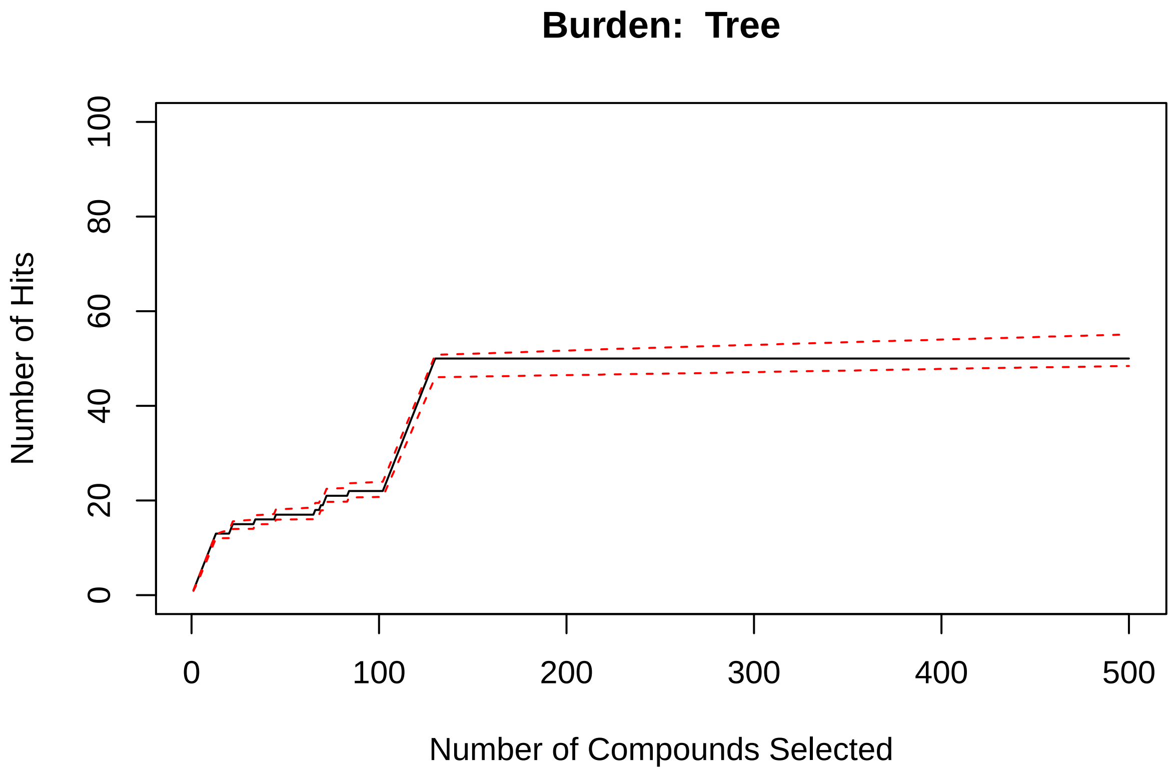
*April 21, 2018*

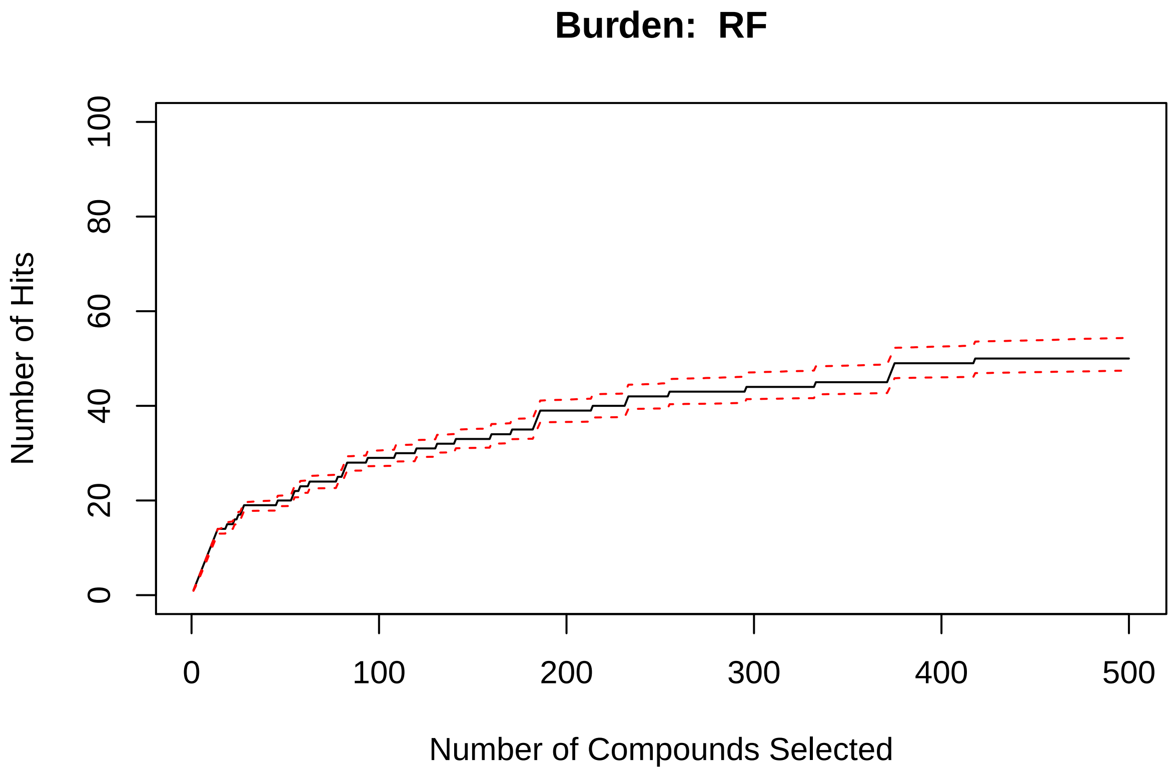
## Accumulation Curve CI

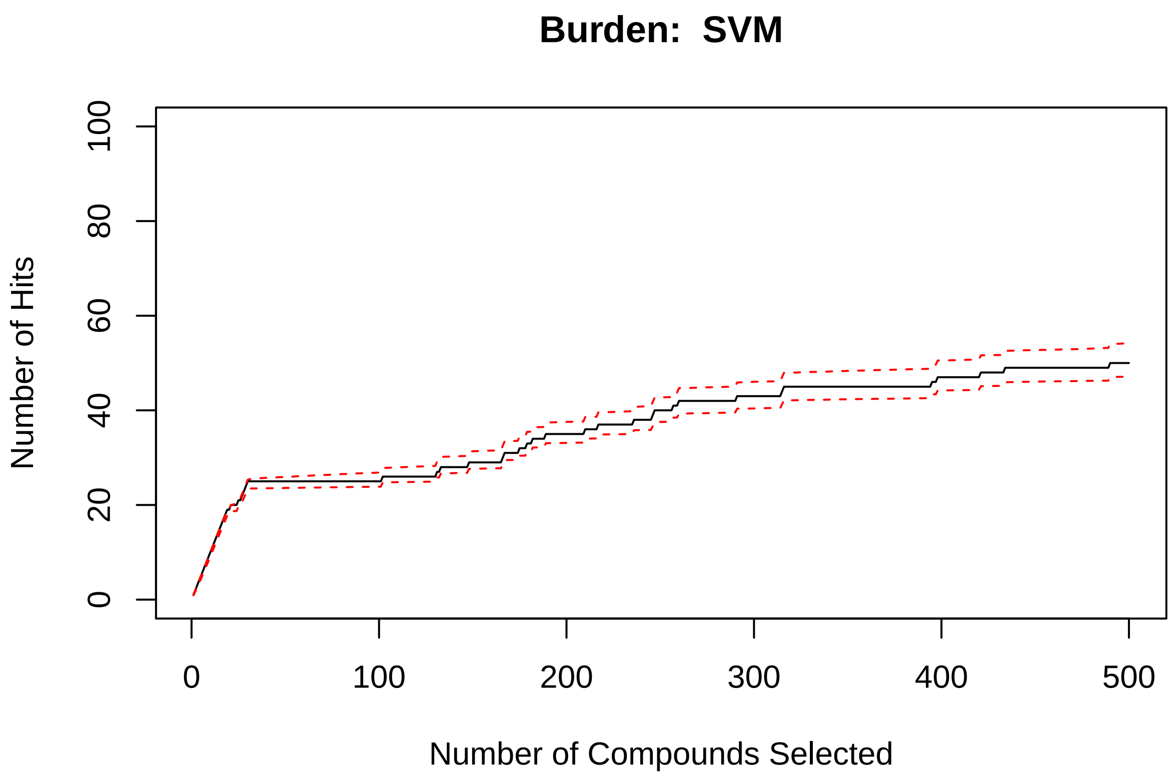
The credibility intervals for different machine-learning models and Burden/Pharmacophore descripters. First with the prior being dependent on the ordered indexing of the compounds selected: $Beta(.01, 0.01 + .0005 * u)$, where $i$ is the number of compounds selected.

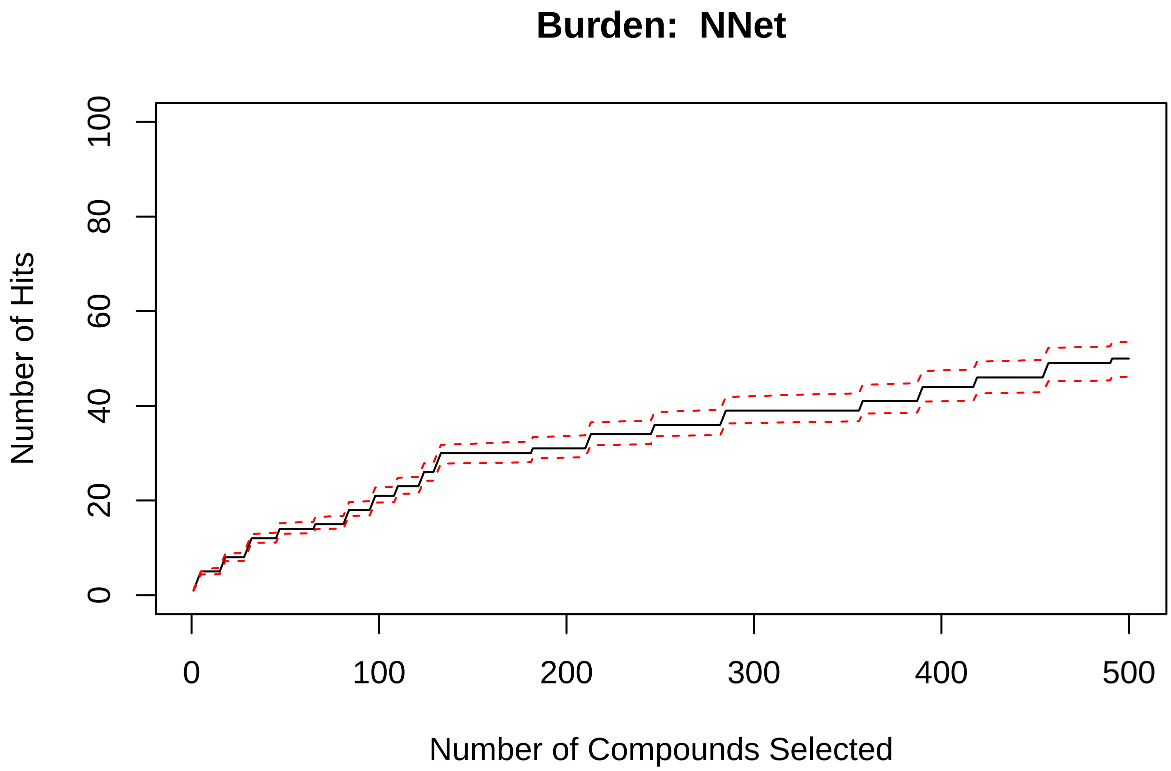
For the Burden Descriptors

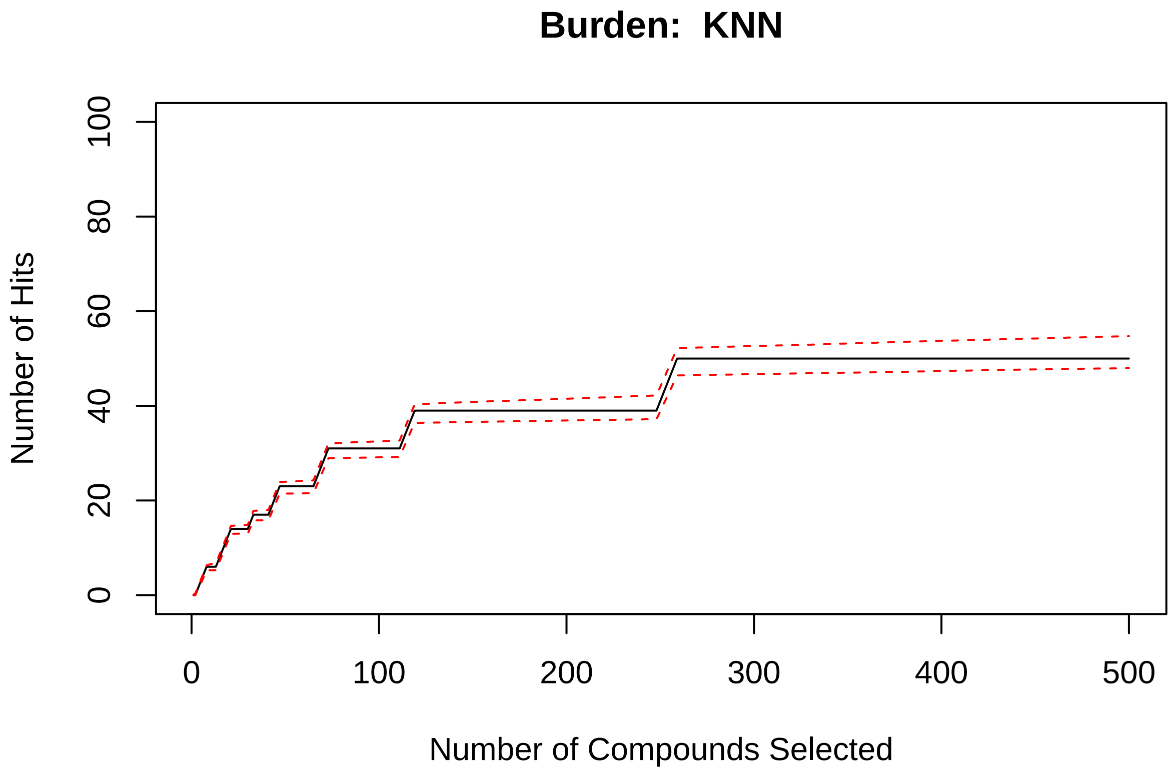
```r
modelType <- c("Tree", "RF", "SVM", "NNet", "KNN", "PLSLDA")

u <- 1:500

int.list.burd.Bayes <- list(length = ncol(probs.Burd) - 1)

for (i in 2:7) {

  hit.vec <- probs.Burd$Observed
  probs <- probs.Burd[, i]

  order.idx <- order(probs, decreasing = T)
  probs <- probs[order.idx]
  hit.vec <- hit.vec[order.idx]

  m <- length(probs)
  int.mat <- matrix(ncol = 3, nrow = m)
  colnames(int.mat) <- c("NHits", "LB", "UB")

  a <- hit.vec + .01
  b <- 1 - hit.vec + .01 + .0005*u

  sum.samp <- vector(length = 10000)
  for(j in 1:500) {
    int.mat[j, 1] <- sum(hit.vec[1:j])
    sum.samp <- sum.samp + rbeta(10000, a[j], b[j])
    int.mat[j, 2] <- quantile(sum.samp, probs = .025)
    int.mat[j, 3] <- quantile(sum.samp, probs = .975)
  }

  par(mfrow = c(1, 1))
  plot(int.mat[, 1], type = "l", ylim = c(0, 100),
       main = paste("Burden: ", colnames(probs.Burd)[i]), ylab = "Number of Hits",
       xlab = "Number of Compounds Selected")
  lines(int.mat[, 2], type = "l", lty = "dashed", col = "red")
  lines(int.mat[, 3], type = "l", lty = "dashed", col = "red")

  int.list.burd.Bayes[[i-1]] <- int.mat

}
```

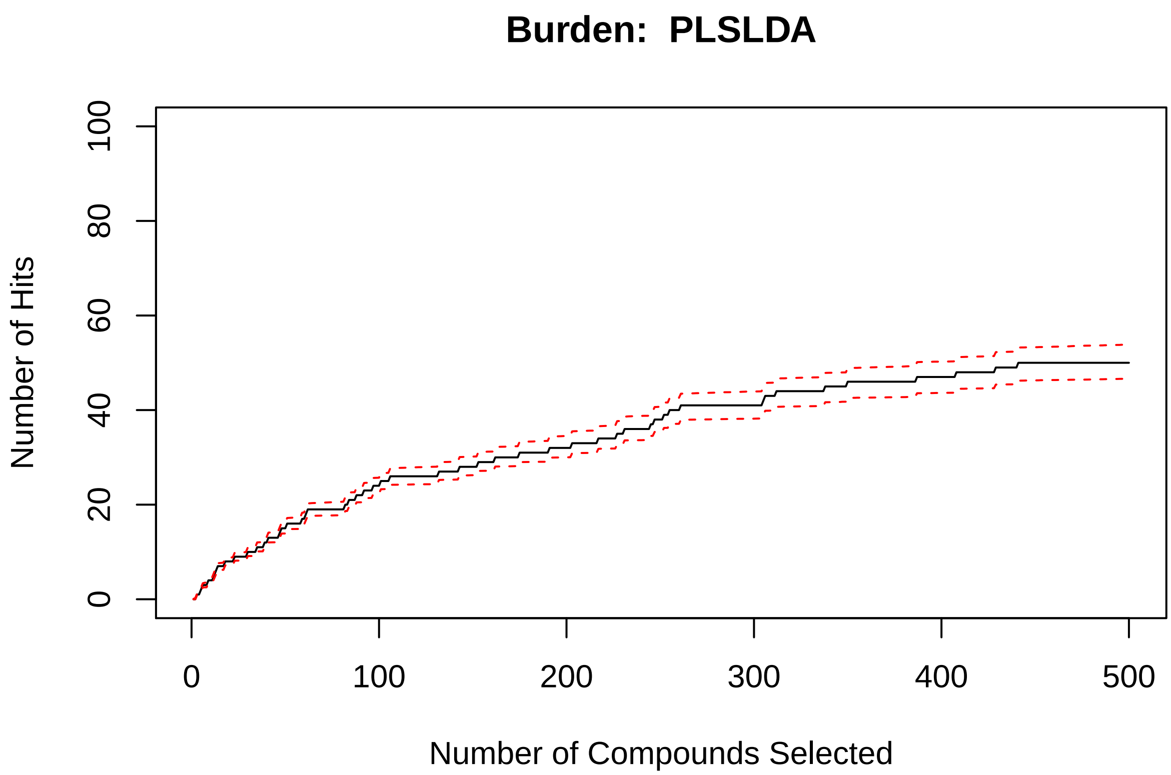

Burden: Tree
Number of Hits
0
20
40
60
80
100
0
100
200
300
400
500
Number of Compounds Selected


Burden: RF
Number of Hits
0
20
40
60
80
100
0
100
200
300
400
500
Number of Compounds Selected


Burden: SVM
Number of Hits
0
20
40
60
80
100
0
100
200
300
400
500
Number of Compounds Selected

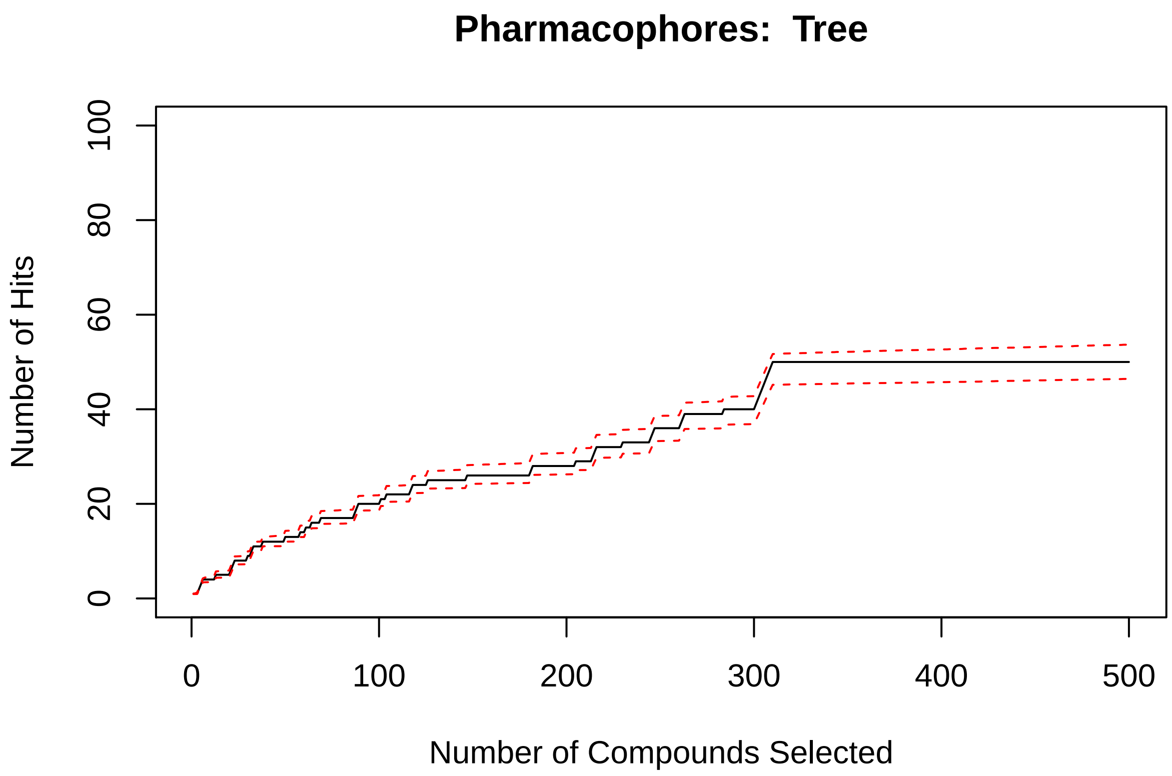

Burden:  NNet
Number of Hits
0
20
40
60
80
100
0
100
200
300
400
500
Number of Compounds Selected

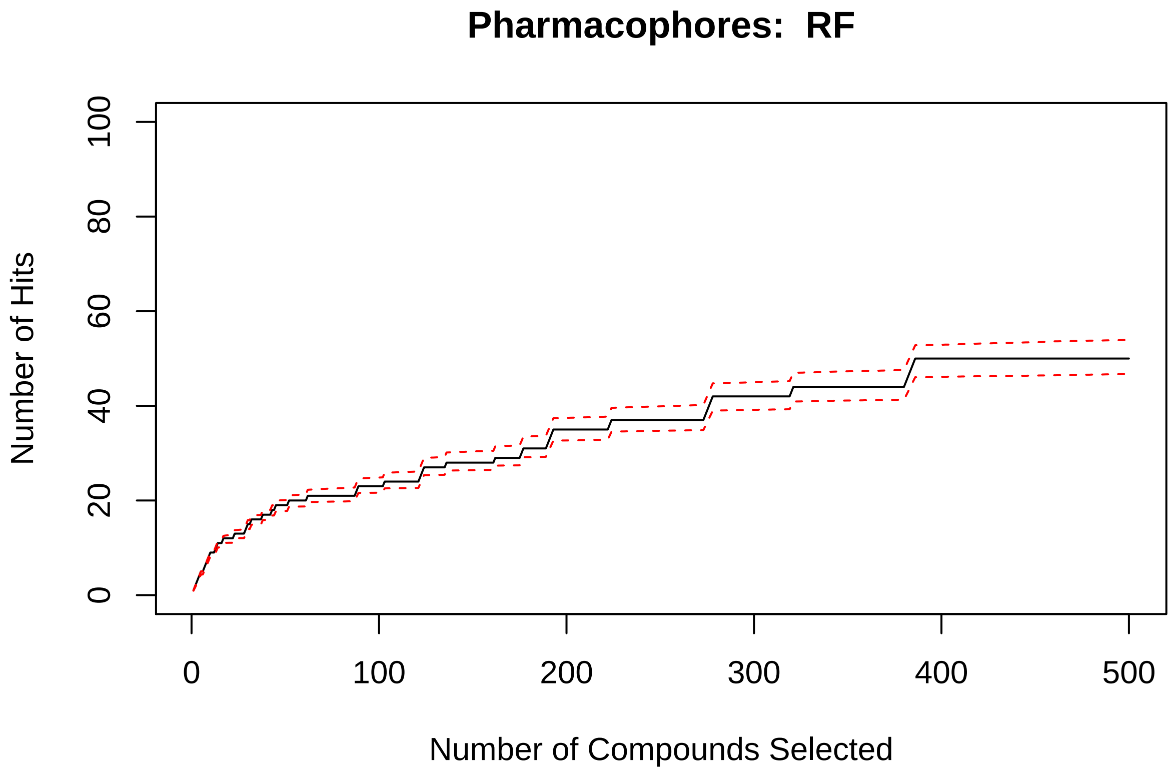

Burden:  KNN
Number of Hits
Number of Compounds Selected
0
20
40
60
80
100
0
100
200
300
400
500

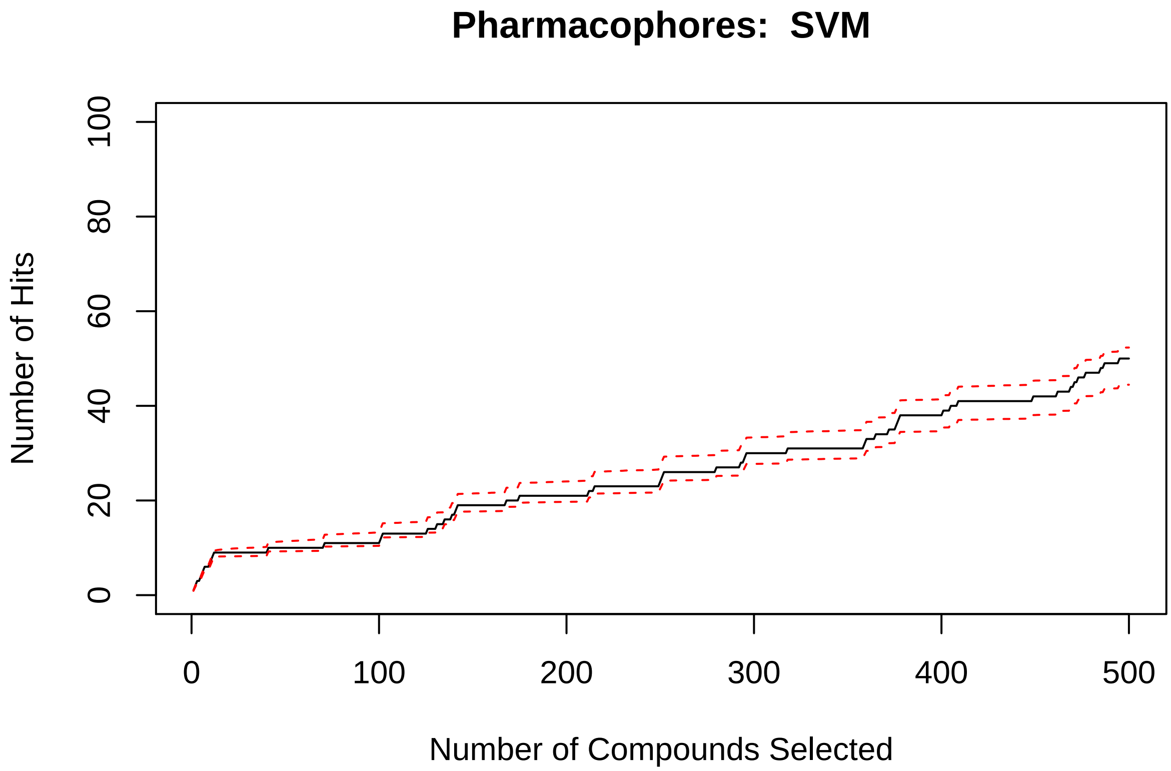


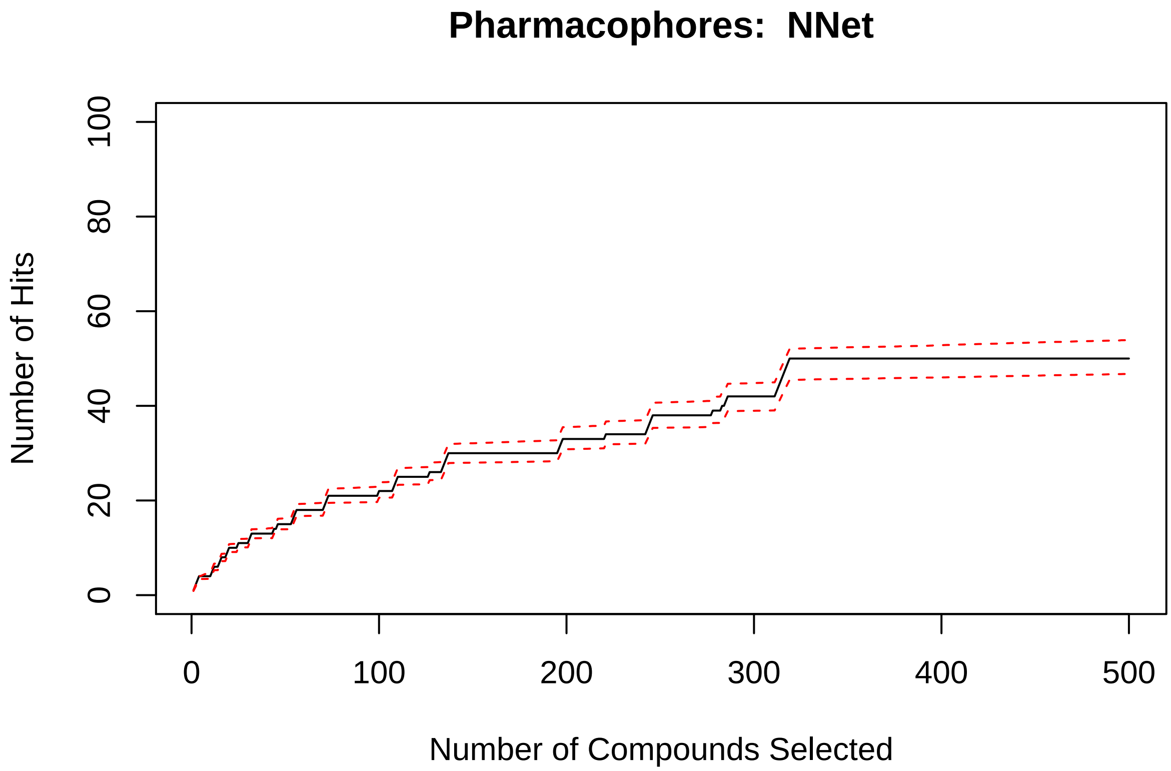
For the Pharmocophores

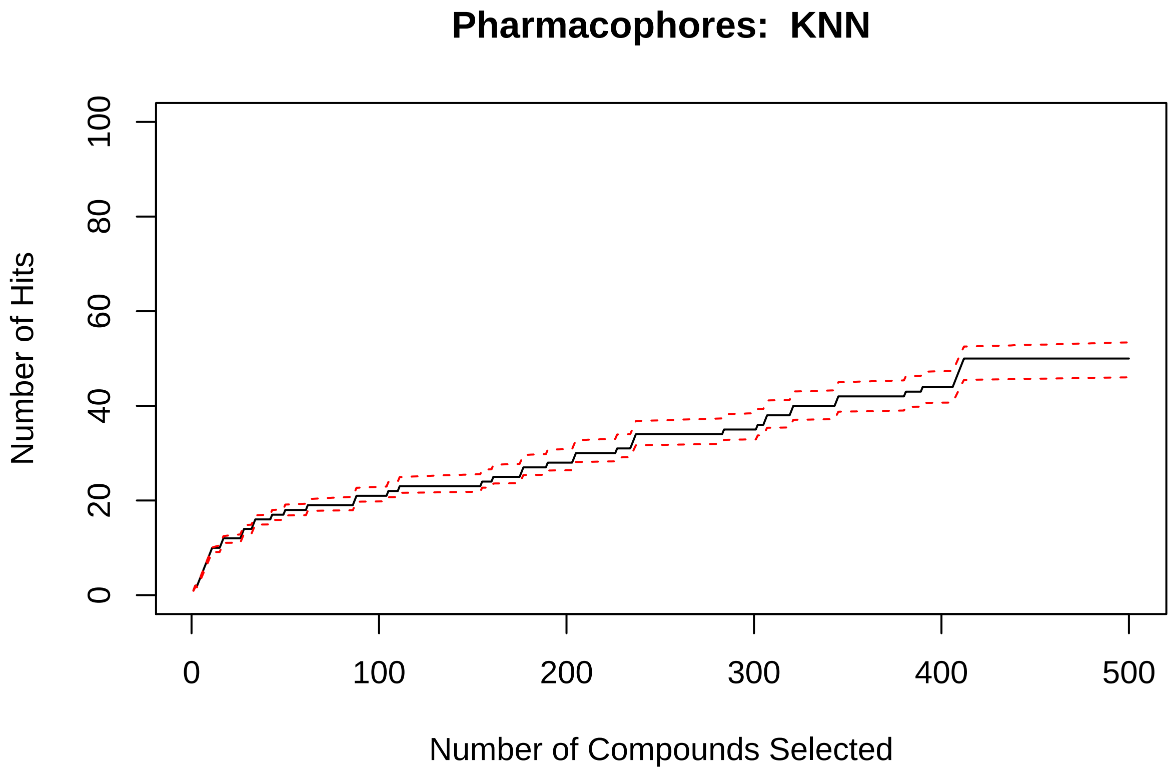
```
int.list.phar.Bayes <- list(length = ncol(probs.Phar) - 1)

for (i in 2:7) {

  hit.vec <- probs.Phar$Observed
  probs <- probs.Phar[, i]

  order.idx <- order(probs, decreasing = T)
  probs <- probs[order.idx]
  hit.vec <- hit.vec[order.idx]

  m <- length(probs)
  int.mat <- matrix(ncol = 3, nrow = m)
  colnames(int.mat) <- c("NHits", "LB", "UB")

  a <- hit.vec + .01
  b <- 1 - hit.vec + .01 + .0005*u

  sum.samp <- vector(length = 10000)
  for(j in 1:500) {
    int.mat[j, 1] <- sum(hit.vec[1:j])
    sum.samp <- sum.samp + rbeta(10000, a[j], b[j])
    int.mat[j, 2] <- quantile(sum.samp, probs = .025)
    int.mat[j, 3] <- quantile(sum.samp, probs = .975)
  }
  par(mfrow = c(1, 1))
  plot(int.mat[, 1], type = "l", ylim = c(0, 100),
       main = paste("Pharmacophores: ", colnames(probs.Phar)[i]), ylab = "Number of Hits",
       xlab = "Number of Compounds Selected")
  lines(int.mat[, 2], type = "l", lty = "dashed", col = "red")
  lines(int.mat[, 3], type = "l", lty = "dashed", col = "red")

  int.list.phar.Bayes[[i-1]] <- int.mat

}
```

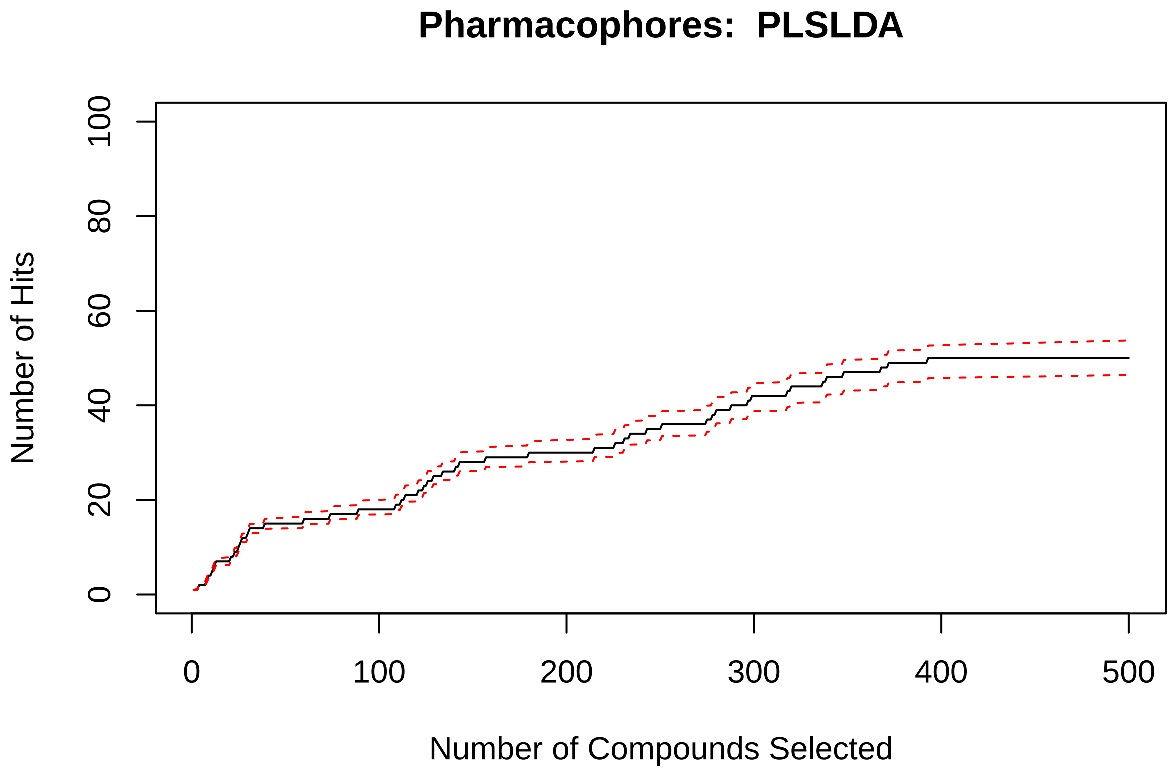



Pharmacophores: RF
Number of Hits
0
20
40
60
80
100
0
100
200
300
400
500
Number of Compounds Selected


Pharmacophores:  SVM
Number of Hits
0
20
40
60
80
100
0
100
200
300
400
500
Number of Compounds Selected


Pharmacophores:  NNet
Number of Hits
0
20
40
60
80
100
0
100
200
300
400
500
Number of Compounds Selected


Pharmacophores: KNN
Number of Hits
0
20
40
60
80
100
0
100
200
300
400
500
Number of Compounds Selected

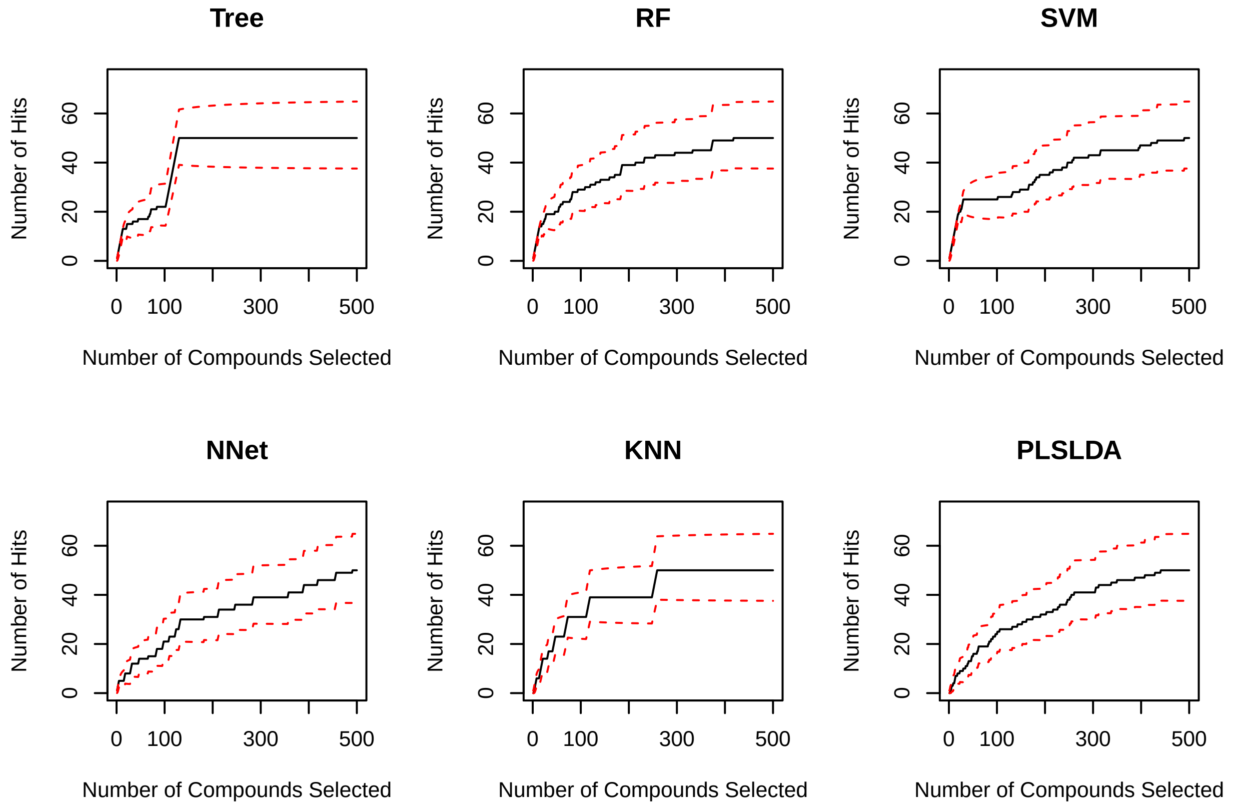


And then the frequentist runs

Burden

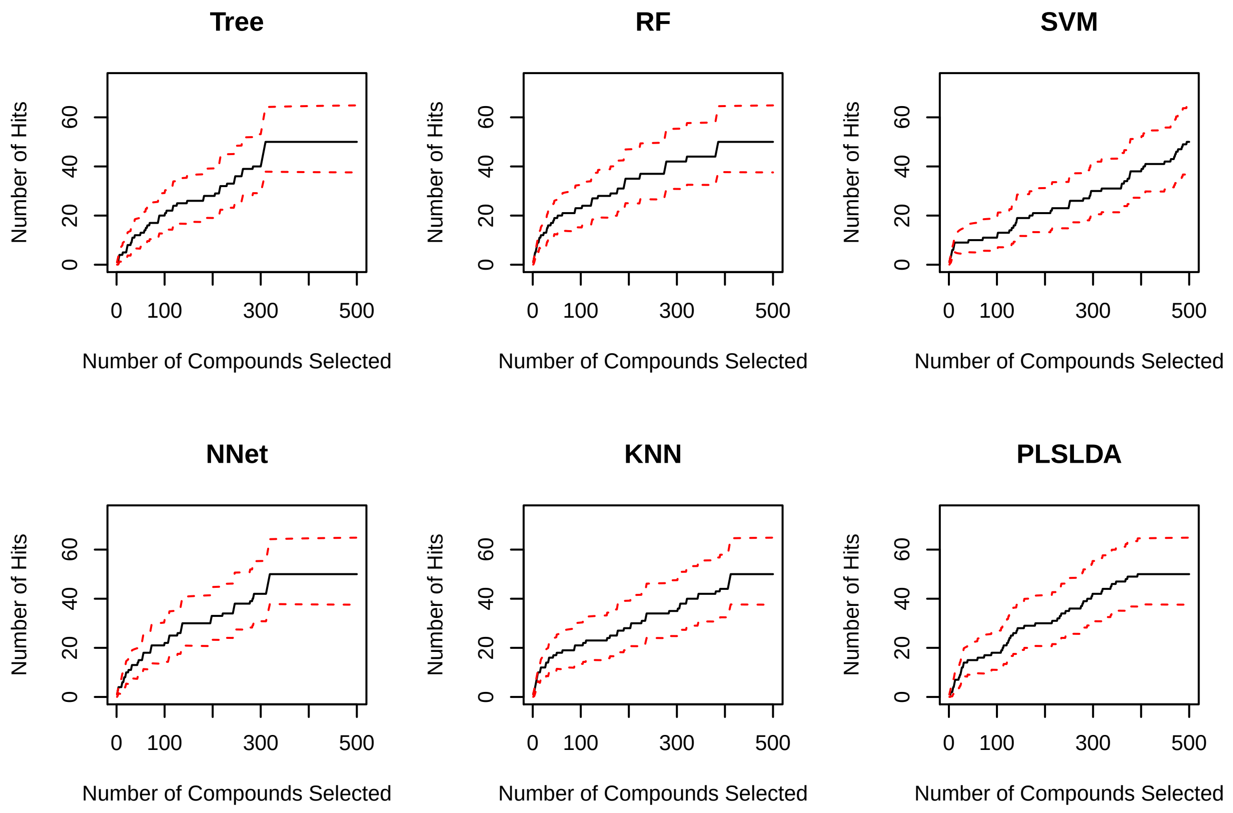
```
int.list.burd.Freq <- list(length = ncol(probs.Burd) - 1)
for(i in 2:ncol(probs.Burd)) {
  probs <- probs.Burd[, i]
  hit.vec <- probs.Burd$Observed

  order.idx <- order(probs, decreasing = T)
  probs <- probs[order.idx]
  hit.vec <- hit.vec[order.idx]

  m <- length(probs)

  # Matrix containing the number of hits and lower and upper bounds for 95%
  # confidence intervals for each number of tests
  int.mat <- matrix(ncol = 3, nrow = m)
  colnames(int.mat) <- c("NHits", "LB", "UB")

  for(j in 1:m) {
    int.mat[j, 1] <- sum(hit.vec[1:j])
    int.mat[j, 2:3] <- j * CPInt(x = sum(hit.vec[1:j]), p.vec = probs[1:j])
  }
  int.list.burd.Freq[[i-1]] <- int.mat
}
```

```
# Plotting the accumulation curves and confidence band for each modeling method
par(mfrow = c(2, 3))
for (i in seq_along(int.list.burd.Freq)) {
  plot(int.list.burd.Freq[[i]][, 1], type = "l", ylim = c(0, 75),
       main = colnames(probs.Burd)[i+1], ylab = "Number of Hits",
       xlab = "Number of Compounds Selected")
  lines(int.list.burd.Freq[[i]][, 2], type = "l", lty = "dashed", col = "red")
  lines(int.list.burd.Freq[[i]][, 3], type = "l", lty = "dashed", col = "red")
}
```



Pharmacophore

```
# Confidence intervals for the accumulation curves for Pharmamocophores descriptors Models

int.list.phar.Freq <- list(length = ncol(probs.Burd) - 1)
for(i in 2:ncol(probs.Phar)) {
  probs <- probs.Phar[, i]
  hit.vec <- probs.Phar$Observed

  order.idx <- order(probs, decreasing = T)
  probs <- probs[order.idx]
  hit.vec <- hit.vec[order.idx]

  m <- length(probs)
  int.mat <- matrix(ncol = 3, nrow = m)
  colnames(int.mat) <- c("NHits", "LB", "UB")
  for(j in 1:m) {
    int.mat[j, 1] <- sum(hit.vec[1:j])
    int.mat[j, 2:3] <- j * CPInt(x = sum(hit.vec[1:j]), p.vec = probs[1:j])
  }
  int.list.phar.Freq[[i-1]] <- int.mat
}

par(mfrow = c(2, 3))
for (i in seq_along(int.list.phar.Freq)) {
  plot(int.list.phar.Freq[[i]][, 1], type = "l", ylim = c(0, 75),
       main = colnames(probs.Burd)[i+1], ylab = "Number of Hits",
       xlab = "Number of Compounds Selected")
  lines(int.list.phar.Freq[[i]][, 2], type = "l", lty = "dashed", col = "red")
  lines(int.list.phar.Freq[[i]][, 3], type = "l", lty = "dashed", col = "red")
}
```



```
intervals <- list(Burd.Freq = int.list.burd.Freq, Phar.Freq = int.list.phar.Freq,
                  Burd.Bayes = int.list.burd.Bayes, Phar.Bayes = int.list.phar.Bayes)


sig.difs <- vector(length = 4)
names(sig.difs) <- c("BF", "PF", "BB", "PB")

av.width <- vector(length = 4)
names(av.width) <- c("BF", "PF", "BB", "PB")
```

```r
av.dist.5 <- vector(length = 4)
names(av.width) <- c("BF", "PF", "BB", "PB")

for (l in 1:4) {
  overlap.count <- 0
  total.count <- 0
  width <- 0
  dist.5 <- 0
  for (i in 1:6) {
    for (k in 1:500) {
      for (j in i:6) {
        if ((intervals[[l]][[i]][k, 3] > intervals[[l]][[j]][k, 2] &
            intervals[[l]][[i]][k, 2] < intervals[[l]][[j]][k, 2]) |
            (intervals[[l]][[j]][k, 3] > intervals[[l]][[i]][k, 2] &
            intervals[[l]][[j]][k, 2] < intervals[[l]][[i]][k, 2])) {
          overlap.count <- overlap.count + 1
        }
        total.count <- total.count + 1
      }
      width <- width + intervals[[l]][[i]][k, 3] - intervals[[l]][[i]][k, 2]
      dist.5 <- dist.5 + abs((intervals[[l]][[i]][k, 3] + intervals[[l]][[i]][k, 2])/(k*2) - .5)
    }
  }
  sig.difs[l] <- 1 - overlap.count/total.count
  av.width[l] <- width/(6*500)
  av.dist.5[l] <- dist.5/(6*500)
}
sig.difs
```

```
##        BF        PF        BB        PB
## 0.3989524 0.4296190 0.6037143 0.5868571
```

```r
av.width
```

```
##        BF        PF        BB        PB
## 21.924483 20.939654  4.860216  4.941814
```

```r
av.dist.5
```

```
## [1] 0.3027097 0.3259283 0.3111292 0.3335081
```

```r
df <- data.frame(rbind(sig.difs, av.width, av.dist.5))
df <- round(df, 2)
df <- df[, c(1,3,2,4)]
rownames(df) <- c("Fraction of Signficant Differences", "Average Interval Width", "Average Distance fro
colnames(df) <- c("Frequentist", "Bayes", "Frequentist", "Bayes")

kable(df, format = "latex", align = "c", booktabs = T, caption = "Summary measures of all point wise co
```

```r
sig.difs.freq <- (sig.difs[1] + sig.difs[2])/2
sig.difs.bayes <- (sig.difs[3] + sig.difs[4])/2
```

Table 1: Summary measures of all point wise confidence/credible intervals

| | ***Burden Numbers*** | | ***Pharmacophores*** | |
|---|---|---|---|---|
| | Frequentist | Bayes | Frequentist | Bayes |
| **Fraction of Signficant Differences** | 0.40 | 0.60 | 0.43 | 0.59 |
| **Average Interval Width** | 21.92 | 4.86 | 20.94 | 4.94 |
| **Average Distance from .5** | 0.30 | 0.31 | 0.33 | 0.33 |