Greedy heuristic sanity check

Anne Hartebrodt

October 8, 2018

```
## Loading required package: data.table
```

```
## Loading required package: gridExtra
```

```
## Loading required package: ggplot2
```

```
## Loading required package: cowplot
```

```
##
## Attaching package: 'cowplot'
```

```
## The following object is masked from 'package:ggplot2':
##
##     ggsave
```

```
## Loading required package: viridis
```

```
## Loading required package: viridisLite
```

Data and parameters

This is a test to see if the greedy heuristic works on simplisitic toydata. A variying numer of differently sized networks have been hidden in the data, we call these "foreground nodes". P-Values for the foreground nodes have been sampled from a normal distribution with mean close to 0 and a small standard deviation. P-values for the background nodes have been sampled from a normal distribution with mean=0.6 and small standard deviation, such that the overlap between the forground and background score distributions is not given.

The greedy heuristic has been performed using the known number of "forground" nodes as maximum network size. For data with more than one "forground" network the maximum network size has been set to a slightly higher number of nodes in order to allow the greedy heuristic to jump to the next network and find all good = forground nodes.

Distribution of foreground and backgound p-values for all the datasets

```
filelocation<-"/home/anne/Masterarbeit/pipes/pipeline_easy_sub/result/biogrid/"
goldstandard<-"/home/anne/Masterarbeit/pipes/pipeline_easy_sub/sample_data/biogrid/"
fname<-"/home/anne/Masterarbeit/thesis/figures/sanity_check/sanity_check_sub.pdf"

dd<-dir(goldstandard)
dat<-data.frame()
for(pdir in dd){
  files<-grep("general", dir(file.path(goldstandard, pdir)), value = T)
for(f in files){
df<-fread(file.path(goldstandard, pdir, f))
df$file<-f
df$dir<-pdir
dat<-rbind(dat,df)
}
```

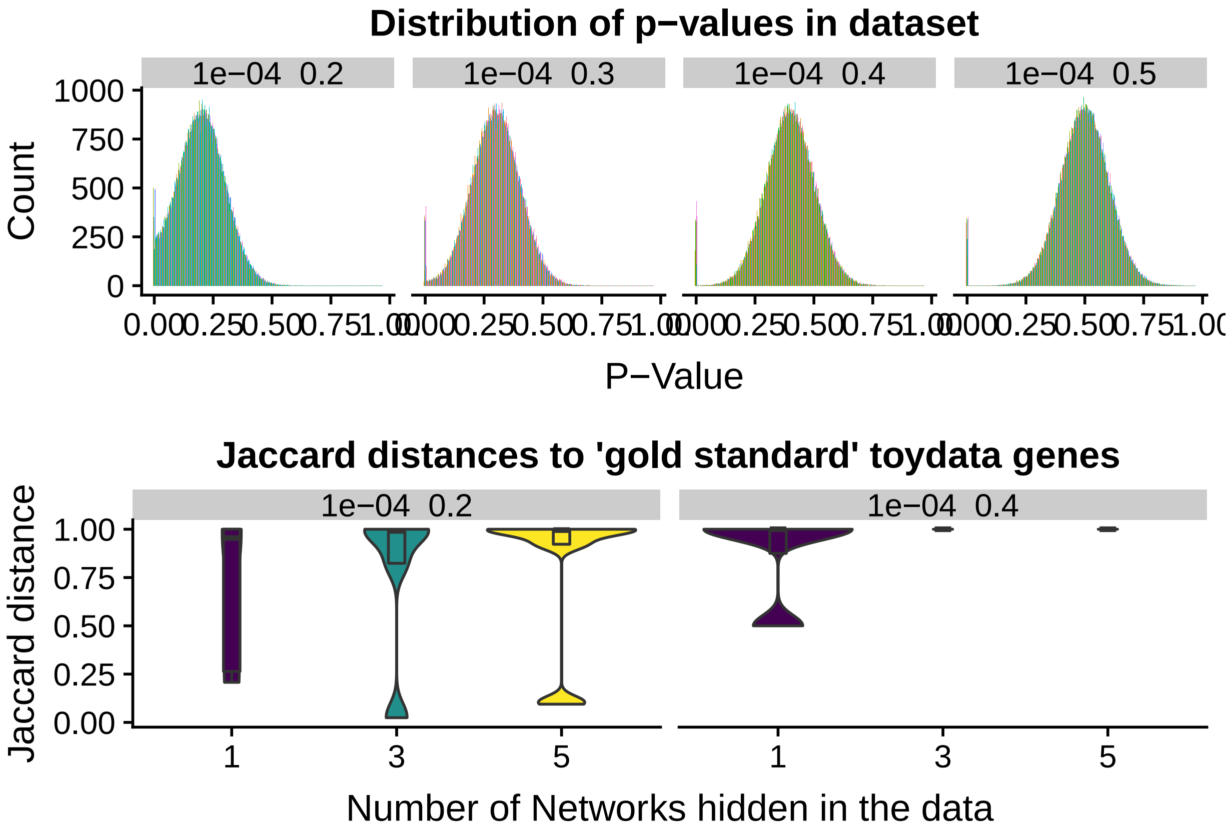
```
}
g0<-ggplot(dat, aes(x=V2, fill=file))+geom_histogram(position = "dodge", bins = 100)+ggtitle("Distribut
```

Overlap of gold standard gene set with detected (solution) networks

In order to assess wether the found solution was close enough to the foreground “gold standard” solution previously hidden in the network, the jaccard index was calculated. A Jaccard index of 1 means, that only the genes in the foreground gene set where contained in the solution.

Below, we see the distribution of the forground and background p-values, with variing degree of overlap and the corresponding performance of the greedy heuristic measured using the Jaccard index.

```
network_jaccard$nr_networks<-as.factor(network_jaccard$nr_networks)
g1 <-ggplot(network_jaccard, aes(x=nr_networks,y = jaccard_dists, fill=nr_networks))  +
  geom_violin()+
  geom_boxplot(width=.1, outlier.colour=NA)+
  guides(fill=F)+
  ggtitle("Jaccard distances to 'gold standard' toydata genes") +
  ylab("Jaccard distance") +
  theme()+xlab("Number of Networks hidden in the data")+
  facet_wrap("run", nrow = 1)+
  scale_fill_viridis_d()
pp<-plot_grid(g0, g1, ncol = 1)
plot(pp)
```



```
ggsave(plot=pp, filename=fname)
```

```
## Saving 6.5 x 4.5 in image
```