# Compleat Database e-letter in Science Signaling

*Savas Paragamian*

*2017-05-24*

## Compleat Database

The data

From the COMPLEAT database and online tool we downloaded all the protein complexes of drosophila. The complexes are both from literature and predicted with a variety of tools. The methology is explained in the paper as well as in the online portal (Vinayagam et al. 2013).

```
# drosophila

compleat_drosophila <- read.delim(file = "Data/drosophila_complexes_compleat0.txt",header = F,sep = "\t
compleat_drosophila <- compleat_drosophila[!(is.na(compleat_drosophila$V2)),] # remove empty rows
# homo sapiens
compleat_homo <- read.delim(file = "Data/compleat1_homo.txt",header = F,sep = "\t")
compleat_homo <- compleat_homo[!(is.na(compleat_homo$V2)),] # remove empty rows
# yeast
compleat_yeast <- read.delim(file = "Data/compleat2_yeast.txt",header = F,sep = "\t")
compleat_yeast <- compleat_yeast[!(is.na(compleat_yeast$V2)),] # remove empty rows
```

## Distribution of the size of complexes

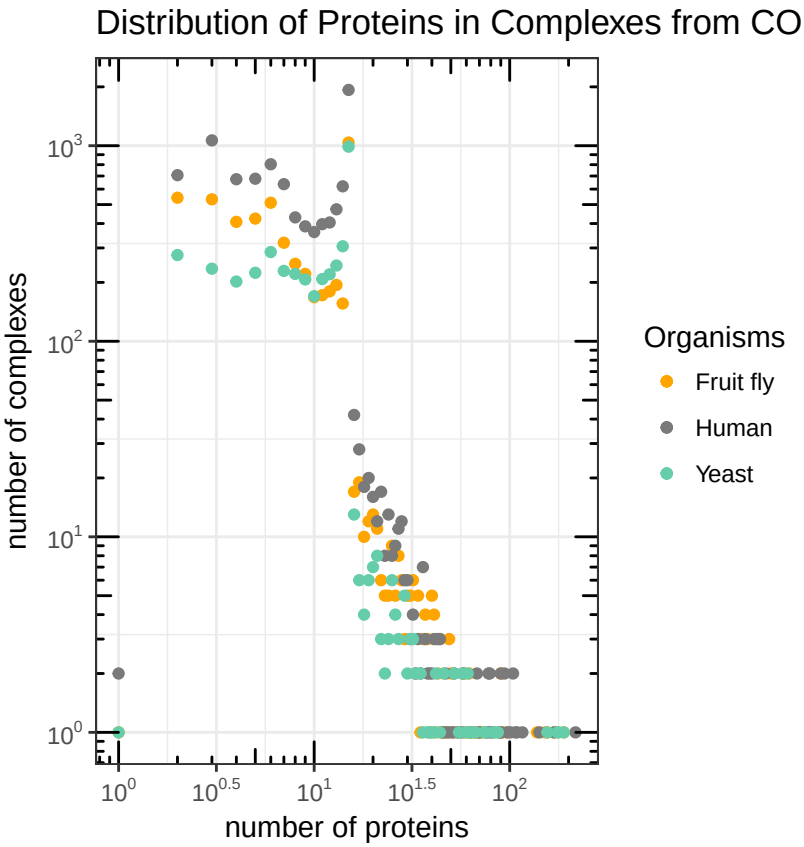
Plot the distributions

```
# drosophila
compleat_drosophila_complex_distribution <- compleat_drosophila %>% group_by(V2) %>% summarise(n_comple
colnames(compleat_drosophila_complex_distribution)[1] <- "complex_with_k_proteins"

# homo sapiens
compleat_homo_complex_distribution <- compleat_homo %>% group_by(V2) %>% summarise(n_complex_with_k_pro
colnames(compleat_homo_complex_distribution)[1] <- "complex_with_k_proteins"

# yeast
compleat_yeast_complex_distribution <- compleat_yeast %>% group_by(V2) %>% summarise(n_complex_with_k_p
colnames(compleat_yeast_complex_distribution)[1] <- "complex_with_k_proteins"
```

Whole distributions.

```
ggplot()+
  geom_point(data = compleat_drosophila_complex_distribution, aes(x = complex_with_k_proteins, y = n_co
  geom_point(data = compleat_homo_complex_distribution, aes(x = complex_with_k_proteins, y = n_complex_
  geom_point(data = compleat_yeast_complex_distribution, aes(x = complex_with_k_proteins, y = n_complex_
  ggtitle("Distribution of Proteins in Complexes from COMPLEAT")+
  scale_y_log10(breaks = trans_breaks("log10", function(x) 10^x),
                labels = trans_format("log10", math_format(10^.x))) +
  scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                labels = trans_format("log10", math_format(10^.x)))+
  annotation_logticks(sides="trbl")+
```

```
  coord_fixed(ratio = 1)+
  scale_colour_manual(values = c("Fruit fly"="orange1", "Human"= "gray48", "Yeast"= "mediumaquamarine")
  labs(x="number of proteins", y="number of complexes")+
  theme_bw()
```




The human dataset doesn't have information on the methods that the complexes were discovered.

Literature VS Predicted

```
## Methods
# drosophila
compleat_drosophila_methods <- compleat_drosophila %>% group_by(V7) %>% summarise(n_complexes=n())

# yeast
compleat_yeast_methods <- compleat_yeast %>% group_by(V7) %>% summarise(n_complexes=n())

## Predicted methods
# drosophila
compleat_drosophila_predicted <- compleat_drosophila %>% filter(V3=="Predicted") %>% group_by(V7) %>% su

# yeast
compleat_yeast_predicted <- compleat_yeast %>% filter(V3=="Predicted") %>% group_by(V7) %>% summarise(n_
```

In yeast and fruit fly the methods of prediction are NetworkBlast and CFinder. Calculate the individual distributions between prediction methods and literature.

```
## Distribution of individual methods
# drosophila
```

```
compleat_drosophila_Literature <- compleat_drosophila %>% filter(V3=="Literature") %>% group_by(V2) %>%
colnames(compleat_drosophila_Literature)[1] <- "complex_with_k_proteins"

compleat_drosophila_NetworkBlast <- compleat_drosophila %>% filter(V7=="NetworkBlast") %>% group_by(V2)
colnames(compleat_drosophila_NetworkBlast)[1] <- "complex_with_k_proteins"

compleat_drosophila_CFinder <- compleat_drosophila %>% filter(V7=="CFinder") %>% group_by(V2) %>% summa
colnames(compleat_drosophila_CFinder)[1] <- "complex_with_k_proteins"

#yeast
compleat_yeast_Literature <- compleat_yeast %>% filter(V3=="Literature") %>% group_by(V2) %>% summarise
colnames(compleat_yeast_Literature)[1] <- "complex_with_k_proteins"

compleat_yeast_NetworkBlast <- compleat_yeast %>% filter(V7=="NetworkBlast") %>% group_by(V2) %>% summa
colnames(compleat_yeast_NetworkBlast)[1] <- "complex_with_k_proteins"

compleat_yeast_CFinder <- compleat_yeast %>% filter(V7=="CFinder") %>% group_by(V2) %>% summarise(n_com
colnames(compleat_yeast_CFinder)[1] <- "complex_with_k_proteins"
```

Summary statistics.

```
summary(compleat_drosophila_NetworkBlast)
```

```
##  complex_with_k_proteins n_complex_with_k_proteins
##  Min.   : 3              Min.   :   2.0
##  1st Qu.: 6              1st Qu.: 132.0
##  Median : 9              Median : 148.0
##  Mean   : 9              Mean   : 222.5
##  3rd Qu.:12              3rd Qu.: 222.0
##  Max.   :15              Max.   :1018.0
```

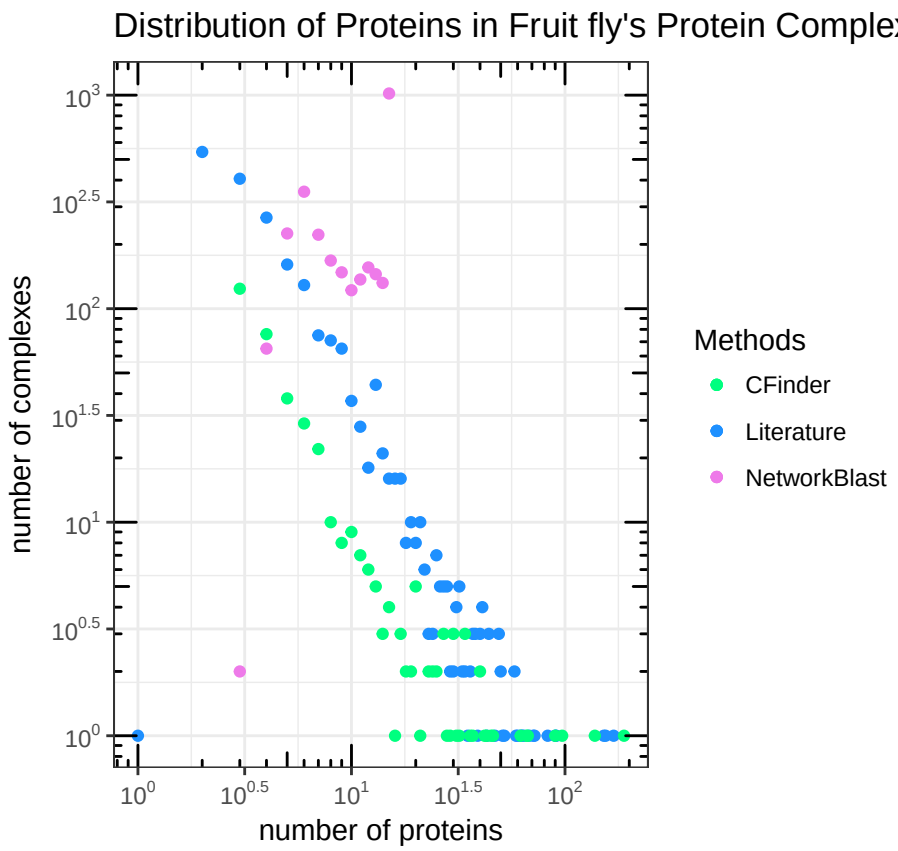
```
summary(compleat_yeast_NetworkBlast)
```

```
##  complex_with_k_proteins n_complex_with_k_proteins
##  Min.   : 3              Min.   :  3.0
##  1st Qu.: 6              1st Qu.:127.0
##  Median : 9              Median :167.0
##  Mean   : 9              Mean   :220.2
##  3rd Qu.:12              3rd Qu.:205.0
##  Max.   :15              Max.   :974.0
```

Plot drosophila.

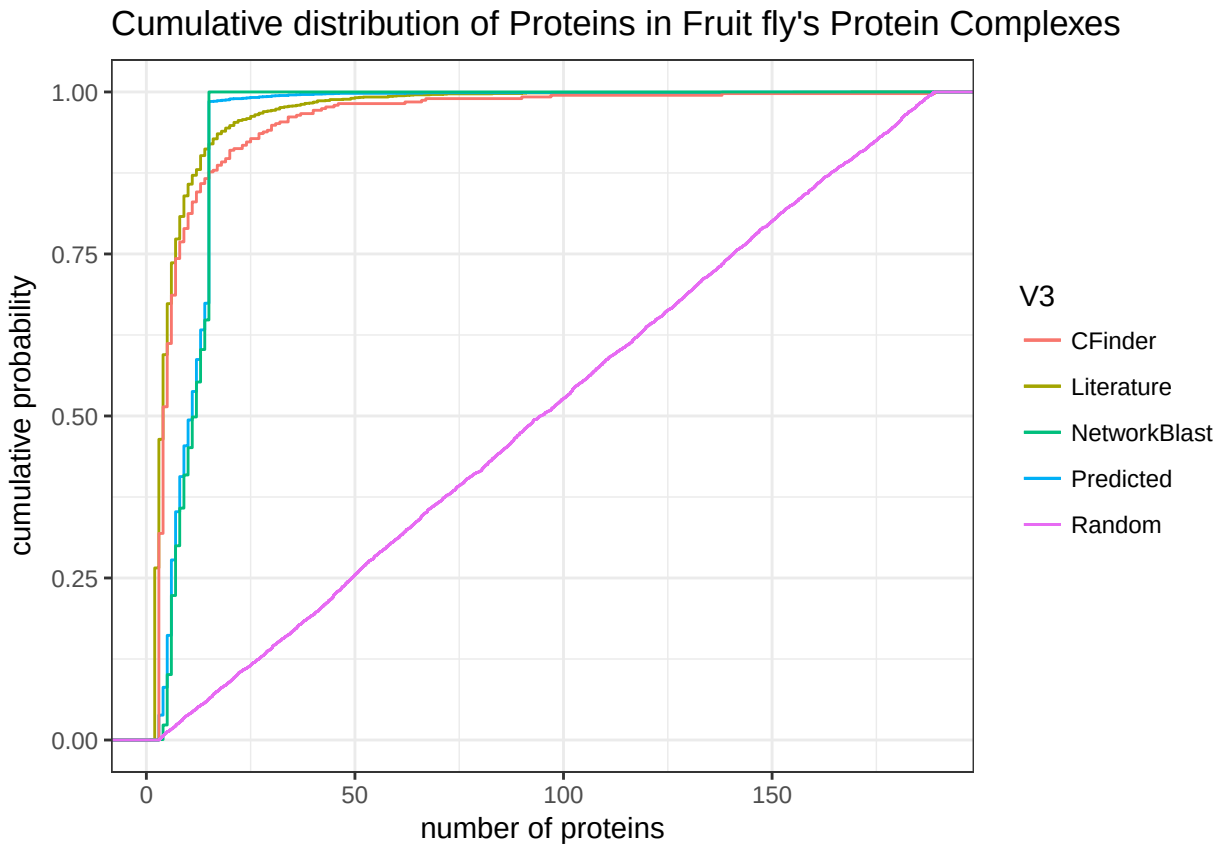
```
ggplot()+
  geom_point(data = compleat_drosophila_Literature, aes(x = complex_with_k_proteins, y = n_complex_with
  geom_point(data = compleat_drosophila_CFinder, aes(x = complex_with_k_proteins, y = n_complex_with_k_
  geom_point(data = compleat_drosophila_NetworkBlast, aes(x = complex_with_k_proteins, y = n_complex_wi
  ggtitle("Distribution of Proteins in Fruit fly's Protein Complexes")+
  scale_y_log10(breaks = trans_breaks("log10", function(x) 10^x),
                labels = trans_format("log10", math_format(10^.x))) +
  scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                labels = trans_format("log10", math_format(10^.x)))+
  annotation_logticks(sides="trbl")+
  coord_fixed(ratio = 1)+
  scale_colour_manual(values = c("Literature"="dodgerblue", "CFinder"= "springgreen", "NetworkBlast"= "
  labs(x="number of proteins", y="number of complexes")+
```

```
  theme_bw()
```




Cumulative distribution

```
compleat_drosophila_predicted2 <- compleat_drosophila %>% filter(V7=="NetworkBlast" | V7=="CFinder")

random_dist <- as.data.frame(x=runif(n = 5000,min = 3,max = 189))
colnames(random_dist) <- "Random"

ggplot()+
  stat_ecdf(data = compleat_drosophila, aes(x = V2, colour=V3),geom = "step")+
  stat_ecdf(data = compleat_drosophila_predicted2, aes(x = V2, colour=V7),geom = "step")+
  stat_ecdf(data = random_dist, aes(x = Random, colour="Random"),geom = "step")+
  ggtitle("Cumulative distribution of Proteins in Fruit fly's Protein Complexes")+
  labs(x="number of proteins", y="cumulative probability")+
  theme_bw()
```




Is the distribution of literature a power law??

```
library(poweRlaw)
```

```
## Warning: package 'poweRlaw' was built under R version 3.3.2
```

```
compleat_drosophila_literature2 <- compleat_drosophila %>% filter(V3=="Literature")

lit_power_law <- displ$new(compleat_drosophila_literature2$V2)

parest <- estimate_pars(lit_power_law)

min_est <- estimate_xmin(lit_power_law)

lit_power_law2 <- lit_power_law$setXmin(min_est)

bs <- bootstrap(lit_power_law, no_of_sims = 100, threads = 1)
```

```
## Expected total run time for 100 sims, using 1 threads is 54.8 seconds.
```

```
is_power_law <- bootstrap_p(lit_power_law, no_of_sims=100, threads=2)
```

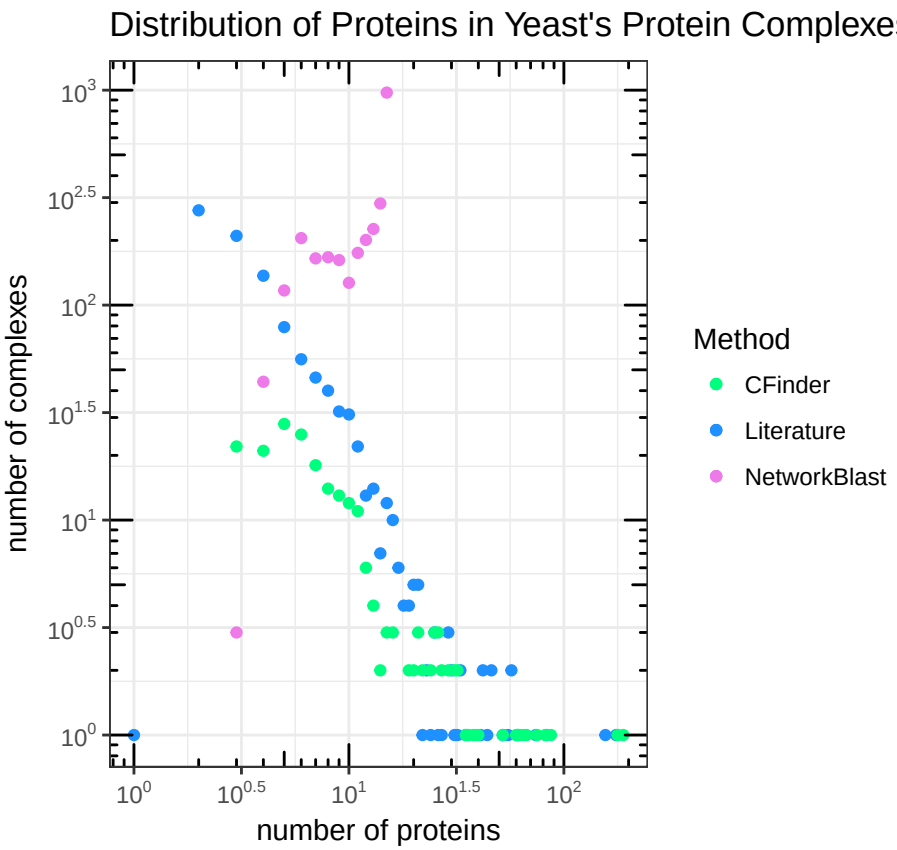
```
## Expected total run time for 100 sims, using 2 threads is 28.5 seconds.
```

```
#m_m = displ$new(moby)

#is_power_lawtest <- bootstrap_p(m_m, no_of_sims=100, threads=2)
```

It is not power law based on the Kolmogorov Smirnoff test develoveped by Clauset and implemented in the R CRAN package "poweRlaw". The p-value is 0.

Plot yeast.

```
ggplot()+
  geom_point(data = compleat_yeast_Literature, aes(x = complex_with_k_proteins, y = n_complex_with_k_pr
  geom_point(data = compleat_yeast_CFinder, aes(x = complex_with_k_proteins, y = n_complex_with_k_prote
  geom_point(data = compleat_yeast_NetworkBlast, aes(x = complex_with_k_proteins, y = n_complex_with_k_
  ggtitle("Distribution of Proteins in Yeast's Protein Complexes")+
  scale_y_log10(breaks = trans_breaks("log10", function(x) 10^x),
                labels = trans_format("log10", math_format(10^.x))) +
  scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                labels = trans_format("log10", math_format(10^.x)))+
  annotation_logticks(sides="trbl")+
  coord_fixed(ratio = 1)+
  scale_colour_manual(values = c("Literature"="dodgerblue", "CFinder"= "springgreen", "NetworkBlast"= "
  labs(x="number of proteins", y="number of complexes")+
  theme_bw()
```




This is because NetworkBlast has an inherent bias towards small sized complexes or is the result of a faulty interpetation in the article?

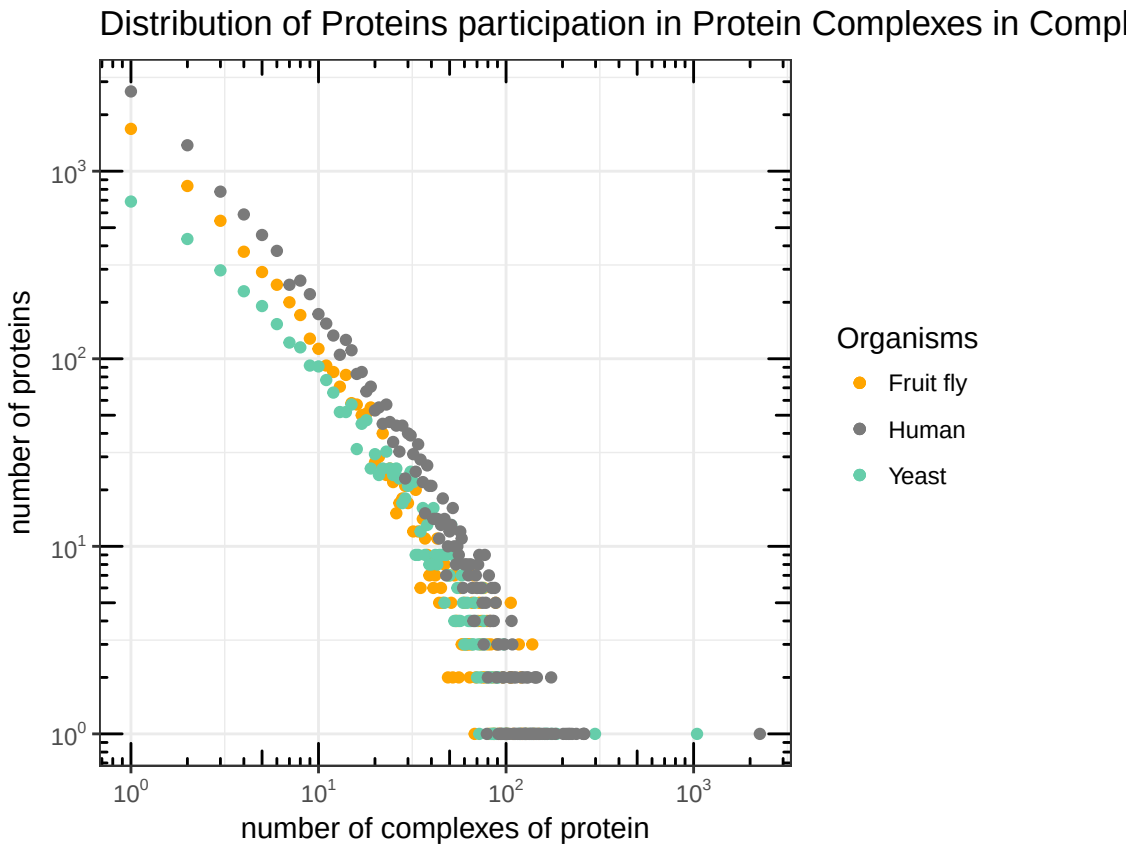
## Distribution of the membership of proteins in complexes

```
# drosophila
compleat_drosophila2 <- compleat_drosophila[,c(1,3,7,12)]
drosophila_splitted_complexes <- strsplit(x = as.character(compleat_drosophila2$V12), " ")
```

```
compleat_drosophila_long <- data.frame(Complex = rep.int(compleat_drosophila2$V1, sapply(drosophila_spli

# yeast
compleat_yeast2 <- compleat_yeast[,c(1,3,7,12)]
yeast_splitted_complexes <- strsplit(x = as.character(compleat_yeast2$V12), " ")

compleat_yeast_long <- data.frame(Complex = rep.int(compleat_yeast2$V1, sapply(yeast_splitted_complexes

# human

compleat_homo2 <- compleat_homo[,c(1,5)]
homo_splitted_complexes <- strsplit(x = as.character(compleat_homo2$V5), ";")

compleat_homo_long <- data.frame(Complex = rep.int(compleat_homo2$V1, sapply(homo_splitted_complexes, le
```

```
# drosophila
compleat_drosophila_protein_dist <- compleat_drosophila_long %>% group_by(Protein) %>% summarize(k_compl
# yeast
compleat_yeast_protein_dist <- compleat_yeast_long %>% group_by(Protein) %>% summarize(k_complexes=n())
# homo
compleat_homo_protein_dist <- compleat_homo_long %>%  group_by(Protein) %>% summarize(k_complexes=n()) %
```

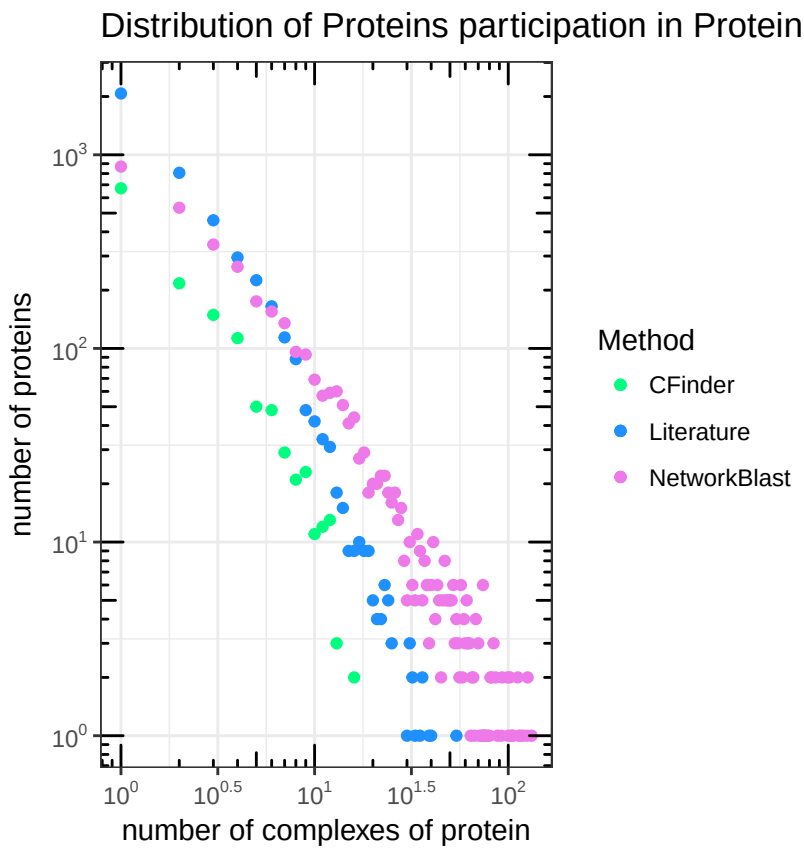
```
ggplot()+
  geom_point(data = compleat_drosophila_protein_dist, aes(x = k_complexes, y = n_protein_k_complexes,col
  geom_point(data = compleat_yeast_protein_dist, aes(x = k_complexes, y = n_protein_k_complexes,color="Y
  geom_point(data = compleat_homo_protein_dist, aes(x = k_complexes, y = n_protein_k_complexes, color="H
  scale_y_log10(breaks = trans_breaks("log10", function(x) 10^x),
                labels = trans_format("log10", math_format(10^.x))) +
  scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                labels = trans_format("log10", math_format(10^.x)))+
  annotation_logticks(sides="trbl")+
  coord_fixed(ratio = 1)+
  scale_colour_manual(values = c("Fruit fly"="orange1", "Yeast"= "mediumaquamarine", "Human"="gray48"),
  ggtitle("Distribution of Proteins participation in Protein Complexes in Compleat")+
  labs(x="number of complexes of protein", y="number of proteins")+
  theme_bw()
```



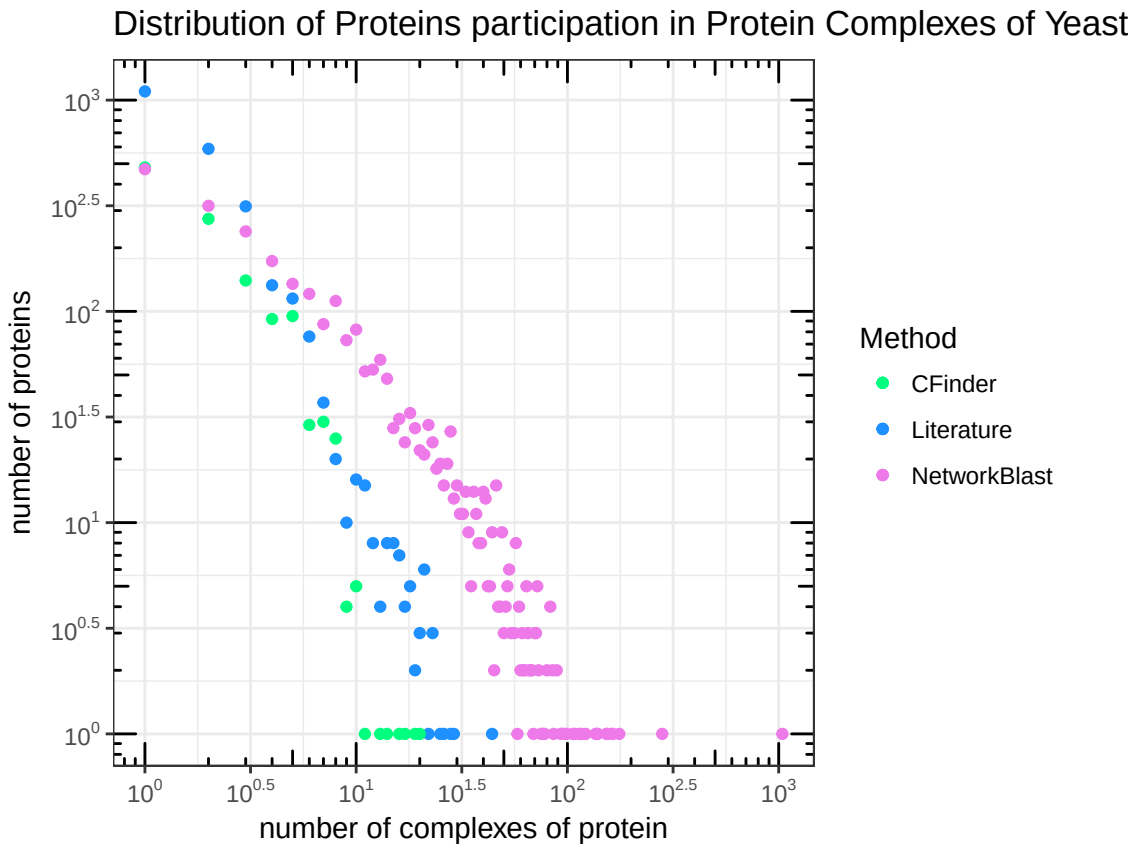

```
# drosophila
compleat_drosophila_protein_Literature_dist <- compleat_drosophila_long %>% filter(Method=="Literature"

compleat_drosophila_protein_NetworkBlast_dist <- compleat_drosophila_long %>% filter(Tool=="NetworkBlas

compleat_drosophila_protein_CFinder_dist <- compleat_drosophila_long %>% filter(Tool=="CFinder") %>% gr

# yeast
compleat_yeast_protein_Literature_dist <- compleat_yeast_long %>% filter(Method=="Literature") %>% grou

compleat_yeast_protein_NetworkBlast_dist <- compleat_yeast_long %>% filter(Tool=="NetworkBlast") %>% gr

compleat_yeast_protein_CFinder_dist <- compleat_yeast_long %>% filter(Tool=="CFinder") %>% group_by(Pro
```

```
ggplot()+
  geom_point(data = compleat_drosophila_protein_Literature_dist, aes(x = k_complexes, y = n_protein_k_c
  geom_point(data = compleat_drosophila_protein_CFinder_dist, aes(x = k_complexes, y = n_protein_k_compl
  geom_point(data = compleat_drosophila_protein_NetworkBlast_dist, aes(x = k_complexes, y = n_protein_k_
  scale_y_log10(breaks = trans_breaks("log10", function(x) 10^x),
                labels = trans_format("log10", math_format(10^.x))) +
  scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                labels = trans_format("log10", math_format(10^.x)))+
  annotation_logticks(sides="trbl")+
  coord_fixed(ratio = 1)+
  scale_colour_manual(values = c("Literature"="dodgerblue", "CFinder"= "springgreen", "NetworkBlast"= "
  ggtitle("Distribution of Proteins participation in Protein Complexes of Fruit fly")+
  labs(x="number of complexes of protein", y="number of proteins")+
```

```
  theme_bw()
```




```
ggplot()+
  geom_point(data = compleat_yeast_protein_Literature_dist, aes(x = k_complexes, y = n_protein_k_comple
  geom_point(data = compleat_yeast_protein_CFinder_dist, aes(x = k_complexes, y = n_protein_k_complexes
  geom_point(data = compleat_yeast_protein_NetworkBlast_dist, aes(x = k_complexes, y = n_protein_k_compl
  scale_y_log10(breaks = trans_breaks("log10", function(x) 10^x),
                labels = trans_format("log10", math_format(10^.x))) +
  scale_x_log10(breaks = trans_breaks("log10", function(x) 10^x),
                labels = trans_format("log10", math_format(10^.x)))+
  annotation_logticks(sides="trbl")+
  coord_fixed(ratio = 1)+
  scale_colour_manual(values = c("Literature"="dodgerblue", "CFinder"= "springgreen", "NetworkBlast"= "
  ggtitle("Distribution of Proteins participation in Protein Complexes of Yeast")+
  labs(x="number of complexes of protein", y="number of proteins")+
  theme_bw()
```




## References

Vinayagam, A, Y Hu, M Kulkarni, C Roesel, R Sopko, S E Mohr, and N Perrimon. 2013. "Protein complex-based analysis framework for high-throughput data sets." *Sci Signal* 6 (264): rs5. doi:10.1126/scisignal.2003629.