Análisis de Reporte de “Trinotate”

Marcel Gustavo Alamán Zárate

Terminal

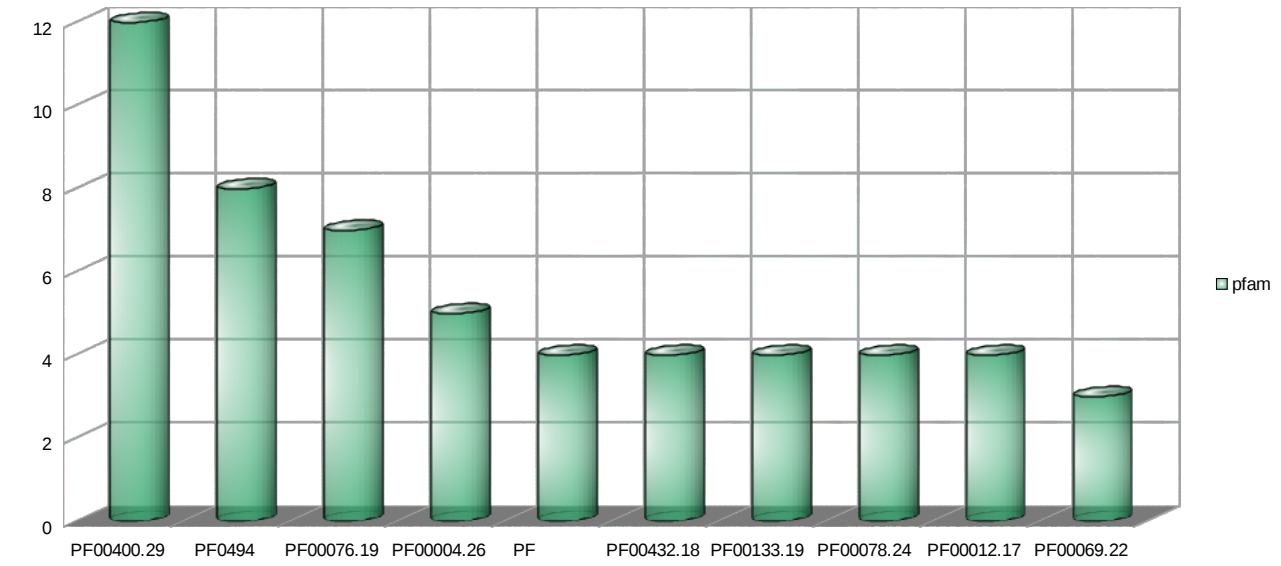
A partir del archivo Trinotate_annotation_report.xls, se empezó procesando en la Terminal los datos con la intención de obtener únicamente las columnas:

```
grep -o "PF[^[:space:]\^_]*" *xls | sort | uniq -c | sed 's/^  *//g' > gus.txt
```

El archivo *gus.txt* contiene el número de ocurrencias del nombre de cada PF y el nombre completo del PF, estos datos a su vez fueron procesados en R:

```
q<-unique(head(sort(gus$V1,TRUE),n=10))
for (i in 1:6){
 print(gus$V2[which(gus$V1==q[i])])
}
```

De ahí se obtienen la gráfica con los 10 dominios más abunantes:



Gráfica A: Gráfica de barras de 10 dominios más abundates.

Dentro de la relevancia que tiene el resultado es primordialmente que refuerza el estudio, por ejemplo el dominio más abundante PF00400.29 corresponde a un motivo (WD40) únicamente presente en eucariotas, como sabemos, los datos transcriptómicos son de una levadura que pertenece a los eucariotas y es razonable observar una ocurrencia alta del motivo. En términos generales los dominios más abundantes indicarían que clase de proteinas son traducidas en las condiciones a las que sometimos al microorganismo, sin embargo no es tan sencillo delimitar que motivos corresponden exactamente a una condición específica del microorganismo; puesto que es razonable pensar que existe una transcripción basal alta de algunas proteinas, que son necesarias para el microorganismo independientemente de la situación a la que esté sometido (proteinas de andamiaje, estructurales, enzimas, etc.). Ahora bien, la solución que se me ocurre para poder asociar qué motivos pertenecen a qué situación en particular, es durante el ensamblado se ensamblar aparte los transcritos de cada fase, de esta manera al comparar los dominios más abundantes dentro de cada fase sería más fácil discernir cuales son dominos de ocurrencia basal y cuales varían de acuerdo al estado de la levadura.