Two criterions to benchmark batch effect correction algorithms with an application to the integration of Baseline RNASeq datasets from Expression Atlas

Guillaume Heger

11/06/2019

I present here two different methods to be used in the context of the integration of datasets from different sources, subject to batch effect. I extracted one of them from an existing paper and I introduce a new one in this paper. Although both of them are based on Principal Components Analysis, the first one has what I would call a "sample-based" approach while the second one focuses more on the information given about genes.

The final aim of this work is the creation of several Expression Atlases by merging the RNASeq datasets available on the existing Expression Atlas. The building of such atlases shall be done using heterogenous datasets which can definitely not be supposed to follow the same distribution. Therefore many batch effect correction algorithms can already be considered as irrelevant for this problem (batch-mean centering to name but one, as well as every method which doesn't take biological factors into account).

The data : Mouse Expression Atlas

In this section, I show a typical workflow of data preparation before integration.

```
#relative path to the data of all the experiments
batch_data<-"Mouse Expression Atlas/batch data/"
#list of experiments
experiments<-list()
for(filename in dir(batch_data)){
  experiments[[filename %>% str_split("-") %>% unlist %>% extract(2:3) %>% paste(collapse="")]] <- get(
}
```

```
#tissues investigated in each experiment
tissues<-experiments %>% map(~.$organism_part)
#matrix of intersections between batches
intersections<-NULL
for(i in tissues %>% seq_along){
  for(j in tissues %>% seq_along){
    intersections%<>%c(length(intersect(tissues[[i]],tissues[[j]])))
  }
}
intersections %<>% matrix(length(tissues)) %<>% set_colnames(names(tissues)) %<>% set_rownames(names(tis
library(igraph)
intersections %>% graph_from_adjacency_matrix %>% plot
```

As three experiments are isolated, one shall not try to integrate them, as it is not possible to correct batch effect with these experiments (since it is not possible then to dissociate batch effect from biological variations).

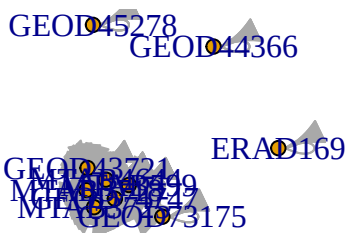

Figure 1: Graph of intersections between the experiments

```
experiments[c('GEOD45278','GEOD44366','ERAD169')]<-NULL
```

After these steps, we can use the package **eigenangles** to integrate the datasets. Let's do it first without any batch effect correction :

```
library(eigenangles)
#integrate.experiments(list=experiments, model=~organism_part, method="none")->all
```