
Stat 215A - Week 9a

Zoe Vernon 10/16/2018

Lab 3 Introduction

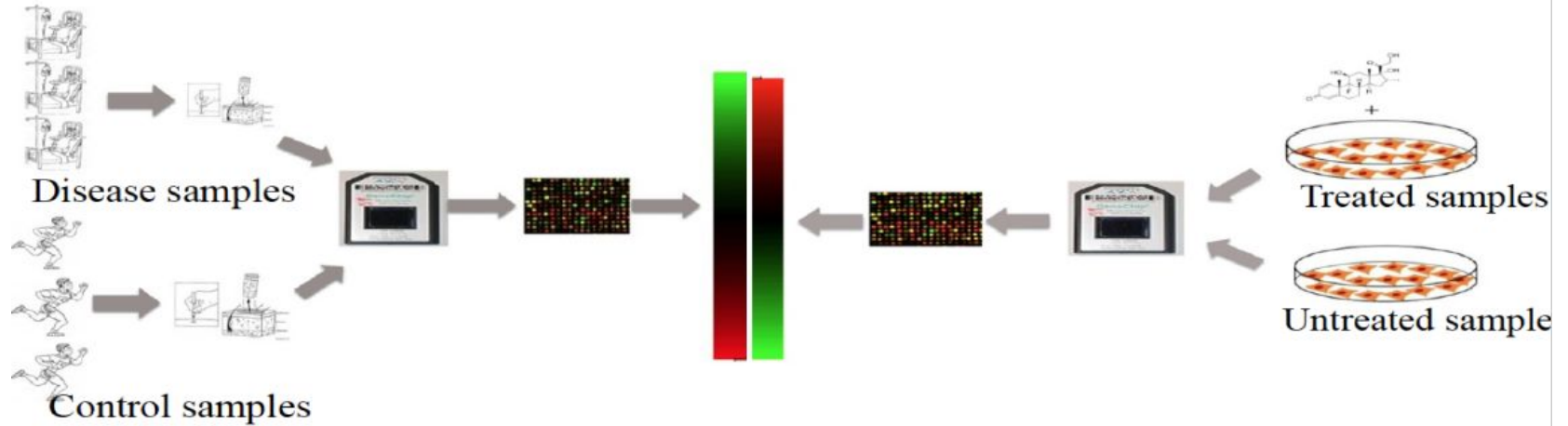
Lab 3 (extension of lab 2)

In this lab you will use the binary encoded data from lab 2.

You are asked to study the stability of k-means by randomizing subsampling the data numerous times.

In this lab you will use parallelization and source C++ code to speed up the computations.

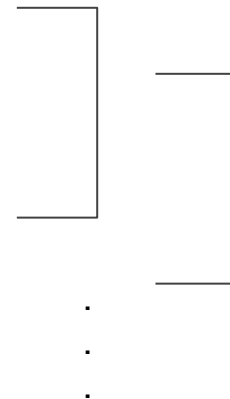
I use this stuff in my own research!



Our solution

Count number of decreasing subsequences of length k in windows of length l
(each window computes in $O(k l \log(l))$ time)

gene	rank in drug signature	rank in disease signature
555	1	1000
2	2	850
300	3	989
\vdots	\vdots	\vdots
690	998	100
690	999	10
700	1000	2



Running simulations

When I ran our method on to count these decreasing subsequences in a real dataset of 66,000 compounds it took ~22 days (with parallelization)

Running it on the same data but using Rcpp means it now only takes ~48 minutes