Evaluating Connection Score on Unordered Query Gene Signatures

24 June, 2017

# Introduction

This document attempts to establish the properties of varying certain properties of unordered query gene signatures in performing Connectivity Map. For details of the connectivity map method employed, see (Zhang and Gant 2008).

Specifically, we will look at a synthetic unordered query gene signature that matches the top m of N genes in a reference profile exactly (in terms of sign - rank is not included in unordered signatures) and establish the distribution of connection score as m tends to N.

We will establish how varying the length of m will affect the connection score and furthermore we will calculate the effect this has on estimated p values.

Finally, we will look at how connection score might be affected by offsetting our synthetic query gene signature such that it matches genes that are ranked lower in the reference profile.

# Creating a synthetic Reference Profile and Query Gene Signature

We will create our synthetic Reference Profile as simply an ordered rank of probes from N to 1 with alternative signs to represent upregulation and downregulation of the genes. The higher the rank number, the higher the change in expression. Note that we select N = 500.

reference.profile <- seq(from=params$N, to=1)  
even.probes <- seq(2, params$N, by = 2)  
  
reference.profile[even.probes] <- -reference.profile[even.probes]

To create the unordered query gene signatures, we can derive these values by taking the sign of these reference profile values:

query.gene.sig <- sign(reference.profile)

# Methods

## Calculate Maximum Connection Strength

max.connection.strength <- function(N, m) {  
 x <- sapply(seq(m), function(i) {  
 N - i + 1  
 })  
   
 sum(x)  
}

## Calculate Connection Strength

calc.connection.strength <- function(query.gene.sig, reference.profile, m, sig.offset = 0) {  
 sig.idx <- seq(from=1+sig.offset, by=1, length.out=m)  
 query.gene.sig.sub <- query.gene.sig[sig.idx]  
 reference.profile.sub <- reference.profile[sig.idx]  
   
 sum(query.gene.sig.sub \* reference.profile.sub)  
}

## Calculate Connection Score

Connection score is simply dividing the connection strength by the theoretical maximum connection strength.

## Random signatures

Let's generate 2000 random signatures.

random.signatures <- lapply(seq(params$Random.Sigs), function(i) {  
 sign(rnorm(n= params$N, mean = 0, sd=1))  
})

# Running the example

Given N = 500 and m from sizes 1 to N...

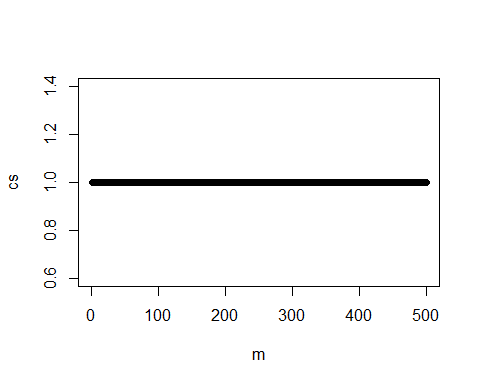
library(parallel)  
  
Ms <- seq(params$N)  
n.cores <- 4  
groups <- sort(rep(seq(4), params$N/n.cores))  
split.Ms <- split(Ms, groups)  
scores.group <- mclapply(split.Ms, function(Ms.sublist) {  
 scores <- sapply(Ms, function(m) {  
 max.cs <- max.connection.strength(params$N, m)  
 cs <- calc.connection.strength(query.gene.sig, reference.profile, m)  
   
 cs/max.cs  
 })  
   
 return(scores)  
})  
  
# flatten the list of scores into one list  
scores <- unlist(scores.group)  
  
result <- data.frame(m=Ms, cs=scores)

## Lowest Connection Scores

|  |  |
| --- | --- |
| m | cs |
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |
| 5 | 1 |
| 6 | 1 |
| 7 | 1 |
| 8 | 1 |
| 9 | 1 |
| 10 | 1 |
| 11 | 1 |
| 12 | 1 |
| 13 | 1 |
| 14 | 1 |
| 15 | 1 |
| 16 | 1 |
| 17 | 1 |
| 18 | 1 |
| 19 | 1 |
| 20 | 1 |

## Highest Connection Scores

|  |  |
| --- | --- |
| m | cs |
| 481 | 1 |
| 482 | 1 |
| 483 | 1 |
| 484 | 1 |
| 485 | 1 |
| 486 | 1 |
| 487 | 1 |
| 488 | 1 |
| 489 | 1 |
| 490 | 1 |
| 491 | 1 |
| 492 | 1 |
| 493 | 1 |
| 494 | 1 |
| 495 | 1 |
| 496 | 1 |
| 497 | 1 |
| 498 | 1 |
| 499 | 1 |
| 500 | 1 |



As we can see from the plot, Connection Scores are all 1 for all m, assuming we start with the most significant rank.

# Calculate P Values

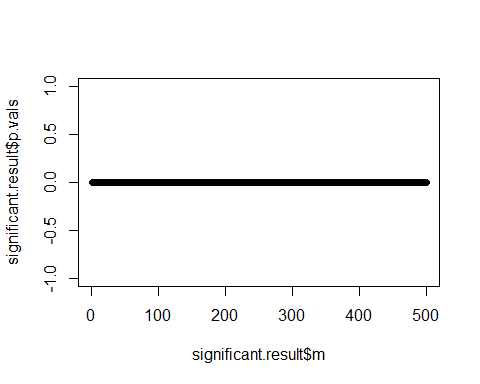
p.vals <- apply(result, 1, function(row) {  
 m <- row[1]  
 cs <- row[2]  
   
 theoretical.max.score <- max.connection.strength(params$N, m)  
   
 random.connection.scores <- lapply(random.signatures, function(rand.sig) {  
 random.strength <- calc.connection.strength(rand.sig, reference.profile, m)  
 random.score <- random.strength / theoretical.max.score  
 })  
   
 num.stronger.scores <- length(which(random.connection.scores > cs))  
 num.stronger.scores / params$Random.Sigs  
})  
  
significant.result<- cbind(result, p.vals)

### P vals for Lowest m

|  |  |  |
| --- | --- | --- |
| m | cs | p.vals |
| 1 | 1 | 0 |
| 2 | 1 | 0 |
| 3 | 1 | 0 |
| 4 | 1 | 0 |
| 5 | 1 | 0 |
| 6 | 1 | 0 |
| 7 | 1 | 0 |
| 8 | 1 | 0 |
| 9 | 1 | 0 |
| 10 | 1 | 0 |
| 11 | 1 | 0 |
| 12 | 1 | 0 |
| 13 | 1 | 0 |
| 14 | 1 | 0 |
| 15 | 1 | 0 |
| 16 | 1 | 0 |
| 17 | 1 | 0 |
| 18 | 1 | 0 |
| 19 | 1 | 0 |
| 20 | 1 | 0 |

### P vals for Highest m

|  |  |  |
| --- | --- | --- |
| m | cs | p.vals |
| 481 | 1 | 0 |
| 482 | 1 | 0 |
| 483 | 1 | 0 |
| 484 | 1 | 0 |
| 485 | 1 | 0 |
| 486 | 1 | 0 |
| 487 | 1 | 0 |
| 488 | 1 | 0 |
| 489 | 1 | 0 |
| 490 | 1 | 0 |
| 491 | 1 | 0 |
| 492 | 1 | 0 |
| 493 | 1 | 0 |
| 494 | 1 | 0 |
| 495 | 1 | 0 |
| 496 | 1 | 0 |
| 497 | 1 | 0 |
| 498 | 1 | 0 |
| 499 | 1 | 0 |
| 500 | 1 | 0 |



As our synthetic query gene signature is a theoretical maximum for the genes selected by design, it impossible for any of the random query gene signatures to have a higher connection score and thus, all p values are 0.

# Offset effect

When I refer to Offset effect, I refer to the effect of the genes in the query gene signature being offset from the highest ranked to the lower ranked. For example, if I selected the 50 most highly ranked genes from the reference profile for my query gene signature, how would the connection score be affected if I took the 50 most significant except one and so forth.

If the synthetic query gene signature is offset by an amount F, then I anticipate that the connection score will go down. This is because, by offsetting the synthetic gene signature, lower ranked, and therefore smaller rank values will be multiplied together with the reference profile. This  
implies that the connection score is biased towards query signatures with higher ranked matching genes, which is what we want.

## Example

Let us choose a nominal m:

m <- 50

If we assume m = 50 and leave N = 500, we can perform connectivity map for all offsets from starting at offset F = 0 until the offset is N - m which is 450. I believe that if we plot the offset of our matching synthetic query gene signatures against connection score that the connection score will increase significantly.

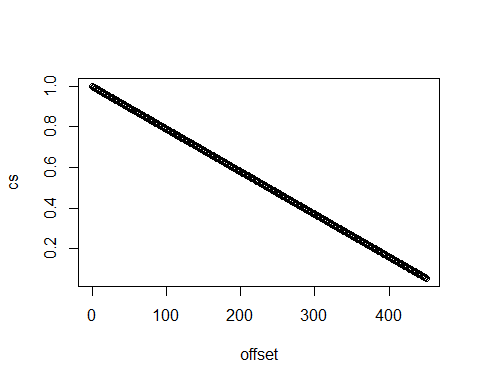
offsets <- seq(0, params$N - m)  
  
offset.connection.scores <- sapply(offsets, function(offset) {  
 max.cs <- max.connection.strength(params$N, m)  
 cs <- calc.connection.strength(query.gene.sig, reference.profile, m, sig.offset = offset)  
   
 cs/max.cs  
})  
  
offset.result <- data.frame(offset=offsets, cs=offset.connection.scores)

### Lowest Offset Scores

|  |  |
| --- | --- |
| offset | cs |
| 0 | 1.0000000 |
| 1 | 0.9978970 |
| 2 | 0.9957939 |
| 3 | 0.9936909 |
| 4 | 0.9915878 |
| 5 | 0.9894848 |
| 6 | 0.9873817 |
| 7 | 0.9852787 |
| 8 | 0.9831756 |
| 9 | 0.9810726 |
| 10 | 0.9789695 |
| 11 | 0.9768665 |
| 12 | 0.9747634 |
| 13 | 0.9726604 |
| 14 | 0.9705573 |
| 15 | 0.9684543 |
| 16 | 0.9663512 |
| 17 | 0.9642482 |
| 18 | 0.9621451 |
| 19 | 0.9600421 |

### Highest Offset Scores

|  |  |
| --- | --- |
| offset | cs |
| 431 | 0.0935857 |
| 432 | 0.0914826 |
| 433 | 0.0893796 |
| 434 | 0.0872766 |
| 435 | 0.0851735 |
| 436 | 0.0830705 |
| 437 | 0.0809674 |
| 438 | 0.0788644 |
| 439 | 0.0767613 |
| 440 | 0.0746583 |
| 441 | 0.0725552 |
| 442 | 0.0704522 |
| 443 | 0.0683491 |
| 444 | 0.0662461 |
| 445 | 0.0641430 |
| 446 | 0.0620400 |
| 447 | 0.0599369 |
| 448 | 0.0578339 |
| 449 | 0.0557308 |
| 450 | 0.0536278 |



These data show a significant effect when the offset of the synthetic query gene signature is increased. In a real world setting, this would imply that, if the unordered query signature consisted of a set of genes exactly matching the sign of the ranks of those in the reference profile, a lower connection score would be observed if those genes were ranked lower in the reference profile, i.e. larger integer values.

This proves that the Connection Score over for an unordered query gene signature against a reference profile is sensitive to how important (i.e. how highly ranked) the selected genes are.

### References

Zhang, Shu-Dong, and Timothy W Gant. 2008. “A Simple and Robust Method for Connecting Small-Molecule Drugs Using Gene-Expression Signatures.” *BMC Bioinformatics* 9 (1). BioMed Central: 258.