Finding Transcription Factors Binding Sites

Introduction

Eukaryotic gene regulation can be very complex. Transcription factor binding to promoter DNA sequences is a stochastic process, and imperfect matches can be sufficient for binding. Chromatin remodeling, methylation, histone modification, chromosome interaction, distal enhancers, and the cooperative binding of transcription co-factors all play an important role. We avoid most of this complexity in this demonstration workflow in order to examine transcription factor binding sites in a small set of seven broadly co-expressed Saccharomyces cerevisiae genes of related function. These genes exhibit highly correlated mRNA expression across 200 experimental conditions, and are annotated to Nitrogen Catabolite Repression (NCR), the means by which yeast cells switch between using rich and poor nitrogen sources.

We will see, however, that even this small collection of co-regulated genes of similar function exhibits considerable regulatory complexity, with (among other things) activators and repressors competing to bind to the same DNA promoter sequence. Our case study sheds some light on this complexity, and demonstrates how several new Bioconductor packages and methods allow us to

- Search and retrieve DNA-binding motifs from the MotifDb package
- Extract the DNA sequence of the promoter regions of genes of interest
- Locate motifs in the promoter sequence

Installation

```
> source("http://bioconductor.org/biocLite.R")
> biocLite(c("MotifDb", "GenomicFeatures",
           "TxDb.Scerevisiae.UCSC.sacCer3.sgdGene",
           "org.Sc.sgd.db", "BSgenome.Scerevisiae.UCSC.sacCer3",
           "motifStack", "seqLogo"))
```

Load packages into your session

```
> library(MotifDb)
> library(seqLogo)
> library(motifStack)
> library(Biostrings)
> library(GenomicFeatures)
> library(org.Sc.sgd.db)
> library(BSgenome.Scerevisiae.UCSC.sacCer3)
> library(TxDb.Scerevisiae.UCSC.sacCer3.sgdGene)
```

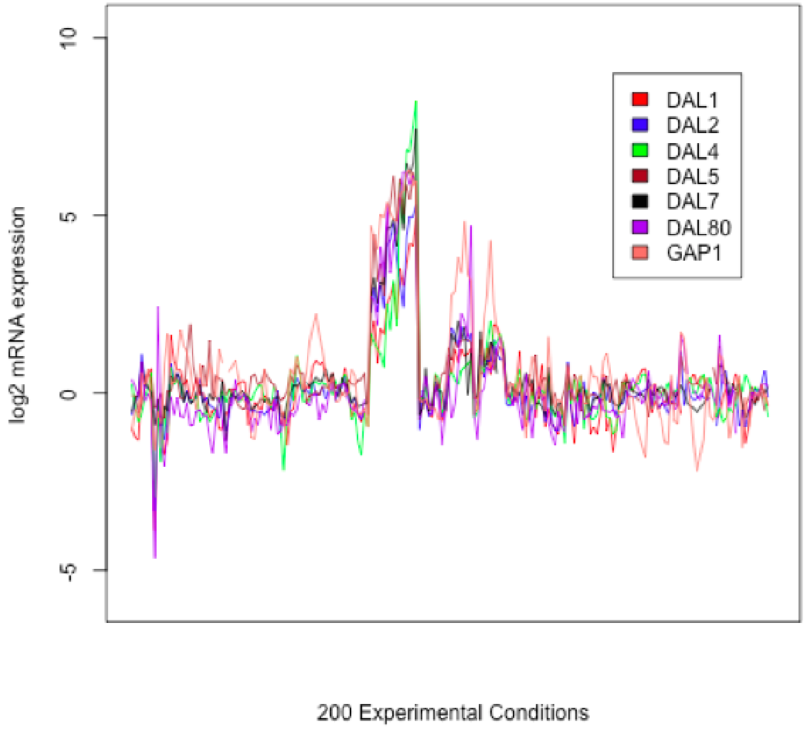
Biological background

The Figure below displays expression levels of seven genes across 200 conditions, from a compendium of yeast expression data which accompanies Allocco et al, 2004, `Quantifying the relationship between co-expression, coregulation and gene function`

Allocco et al establish that

In S. cerevisiae, two genes have a 50% chance of having a common transcription factor binder if the correlation between their expression profiles is equal to 0.84.

These seven highly-correlated (> 0.85) NCR genes form a connected subnetwork within the complete co–expresson network derived from the compendium data (work not shown).



Sequence Search

Sequence-based transcription factor binding site search methods answer two questions:

- For a given TF, what DNA sequence pattern/s does it preferentially bind to?
- Are these patterns present in the promoter region of some gene X?

A genes promoter region is traditionally (if loosely) defined with respect to its transcription start site (TSS): 1000-3000 base pairs upstream, and 100-300 basepairs downstream. For the purposes of this workflow, we will focus only on these cis-regulatory regions, ignoring enhancer regions, which are also protein/DNA binding sites, but typically at a much greater distance from the TSS. An alternative and more inclusive proximal regulatory region may be appropriate for metazoans: 5000 base pairs up- and down stream of the TSS.

For simplicitys sake we will use a uniform upstream distance of 1000 bp, and 0 bp downstream in the analyses below.

Search for position frequency matrix (PFM) DAL80 motif

1.
```
> query(MotifDb, "DAL80")
MotifDb object of length 3
| Created from downloaded public sources: 2013-Aug-30
| 3 position frequency matrices from 3 sources:
|        JASPAR_2014:    1
|        JASPAR_CORE:    1
|             ScerTF:    1
| 1 organism/s
|        Scerevisiae:    3
Scerevisiae-JASPAR_CORE-DAL80-MA0289.1
Scerevisiae-JASPAR_2014-DAL80-MA0289.1
Scerevisiae-ScerTF-DAL80-harbison
```

```
> pfm.dal80.jaspar <- query(MotifDb,"DAL80")[[1]]
```

```
> pfm.dal80.jaspar
           1 2 3 4 5          6    7
A 0.10891089 0 1 0 1 0.90909091 0.03
C 0.66336634 0 0 0 0 0.01010101 0.19
G 0.05940594 1 0 0 0 0.01010101 0.75
T 0.16831683 0 0 1 0 0.07070707 0.03
```

- Plot logo for the motif

```
> seqLogo(pfm.dal80.jaspar)
```

Question 1:

. How many position frequency matrices are found in the motif databases? What are they?

. Save the motif logo of JASPAR motif

. Write the consensus of the JASPAR motif

. Plot ScerTF motif logo and write the consensus of the motif

Extract promoter region sequence of dal1 gene

```
> dal1 <- "YIR027C"
> chromosomal.loc <-
  transcriptsBy(TxDb.Scerevisiae.UCSC.sacCer3.sgdGene, by="gene") [dal1]
> promoter.dal1 <-
  getPromoterSeq(chromosomal.loc, Scerevisiae, upstream=1000, downstream=0)
```

Search for candidate binding sites of the motif

- Convert PFM (Position Frequency Matrix) to PCM (Position Count Matrix)

```
> pcm.dal80.jaspar <- round(100 * pfm.dal80.jaspar)
> pcm.dal80.jaspar
   1   2   3   4   5  6  7
A 11   0 100   0 100 91  3
C 66   0   0   0   0  1 19
G  6 100   0   0   0  1 75
T 17   0   0 100   0  7  3
```

- Search DAL80 using PCM in the promoter regions of DAL1

```
matchPWM(pcm.dal80.jaspar, unlist(promoter.dal1)[[1]], "90%")
```

Question 2:

. How many possible binding sites you find?

Compare motifs

* The seqlogo package has been the standard tool for viewing sequence logos, but can only portray one logo at a time.

```
> dal80.jaspar2014 <- query(MotifDb,"DAL80")[[1]]
> dal80.jasparCORE <- query(MotifDb,"DAL80")[[2]]
> dal80.scertf <-query(MotifDb,"DAL80")[[3]]
> seqLogo(dal80.jaspar2014)
> seqLogo(dal80.jasparCORE)
> seqLogo(dal80.scertf)
```

* The new (October 2012) package motifStack can plot multiple motifs together.
Here, you need to have ghostscript installed in your computer.

For Window

```
Sys.setenv(R_GSCMD="C:/Program Files/gs/gs9.18/bin/gswin64c.exe")
```

First, create instances of the pfm class:

```
pfm.dal80.jaspar2014 <- new("pfm", mat=query(MotifDb, "dal80")[[1]],
                         name="DAL80-JASPAR2014")
pfm.dal80.jasparCORE <- new("pfm", mat=query(MotifDb, "dal80")[[2]],
                         name="DAL80-JASPARCORE")
pfm.dal80.scertf <- new("pfm", mat=query(MotifDb, "dal80")[[3]],
                         name="DAL80-ScerTF")
plotMotifLogoStack(DNAmotifAlignment(c(pfm.dal80.jaspar2014,pfm.dal80.jasparCORE, pfm.dal80.
    scertf)))
```

Question 3

. Save the stack plot of the motifs

. Is there any difference between dal80 motif in different databases?

`Reference`
Finding Candidate Binding Sites for Known Transcription Factors via Sequence Matching