

Summer Institute for Training in Biostatistics 2005

Final week questions on computational biology

Sündüz Keleş

July 18, 2005

Identifying transcription factor binding sites (motifs) using microarray gene expression data

1. Data acquisition and construction of a motif count matrix.

- (a) **Obtaining candidate motifs.** Visit the The Promoter Database of *Saccharomyces cerevisiae* (SCPD) at <http://rulai.cshl.edu/cgi-bin/SCPD/getfactorlist>. Here, the link `Get consensus list` provides a list of known DNA binding motifs and their sequences. Copy and paste these into a file `SCPD_motifs.txt`.

Next, you will need to generate a file that only has the actual motif sequences but not the names. This file will be an input to a pattern search program. For example, you can do the following in R

```
motifs = read.table("SCPD_motifs.txt", col.names = c("motifName", "motifSeq"))
attach(motifs)
cat(as.character(motifSeq), file = "SCPD_motif_seqs.txt", sep = "\n")
```

Now, `SCPD_motif_seqs.txt` contains sequences for known motifs.

- (b) **DNA pattern matching using RSA-Tools.** Next, go to the Regulatory Sequence Analysis Tools (RSAT) web site at <http://rsat.ulb.ac.be/rsat/>. Click on the `Pattern matching --> dna-pattern (strings)` link. Using this web-based program, you will generate a vector of covariates for each cell cycle regulated gene. For a particular gene, each element of this vector will correspond to the number of copies of a known motif in the gene's upstream promoter region.

Fill in the form to do pattern matching as follows: As your `Query patterns`, enter all the known motifs from SCPD (contents of the above `SCPD_motif_seqs.txt` file). As your `Sequence` load the file `spellman_upstream_cellcycle.txt` file which contains the upstream promoter sequences for cell cycle regulated genes. Search both strands, prevent overlapping matches and return only the match count table. Enter your email address to the email box and hit the GO button.

- (c) **Reading in the output from RSA Tools.** After you hit the GO button, you will immediately get an url indicating where the results are posted. You can read in the data to R from this url. First examine the output at the url. You will notice that you need to read in a portion of the output on this page. You can specify this by the number of rows you want to read in and the number of rows you want to skip from the beginning.

```
#First check how many upstream sequences we are using: This is a
#total of 1584/2 = 792 sequences since each upstream sequence has
#a description line that starts with ">".
> seqData = scan("spellman_upstream_cellcycle.txt", what = "")
Read 1584 items
```

```
#Skipping first 63 lines and then reading in 792 lines with a
#header.
> countData = read.table("http://rsat.scmbb.ulb.ac.be/
rsat/tmp/result.2005_07_17.203339.txt", sep = "\t", skip = 63,
nrows = 792, header = T)
```

Now, `countData` is a data frame where the first column contains the gene IDs and the remaining columns correspond to number of occurrences of known motifs. Randomly select 5 motifs and provide summaries of their count data among cell cycle regulated genes, e.g., tables, figures etc.

(d) **Expression data.** You will also need the gene expression data in `spellman_expr_cellcycle.txt` for the rest of the analysis. An updated version of this file is available at <ftp://ftp.cs.wisc.edu/pub/users/keles/Spellman/>.

2. **Identifying potential cell cycle related motifs.** Time points `alpha0` to `alpha119` span two cycles from the yeast cell cycle. In particular, time points `alpha0` to `alpha56` correspond to one cycle and `alpha63` to `alpha119` correspond to another cycle.

For each time point between `alpha0` and `alpha119`, identify the motifs for which there is a difference between the expression levels of the genes with and without a copy of the motif. How are your results affected when you consider 1 or less copies versus more than 1 copies?

3. **Additional support for important motifs.** Recall that these experiments span two cell cycles. Using this periodicity information, can you find additional evidence, i.e., in the form of a plot, to support the important motifs you identified in the above question?

4. **Retrieving information regarding the identified motifs from TRANSFAC.** Visit the transcription factor database TRANSFAC at <http://www.gene-regulation.com/pub/databases.html> and obtain information regarding the motifs you identified as important (Follow the links Search -> Factor; you might have to become a member to browse TRANSFAC. Becoming a member is free and it only requires an email address.)