Sample CoGAPS Analysis

In this example, we reproduce the DESIDE analysis performed in Ochs et al. [4] by applying CoGAPS to gene sets defined by TRANSFAC to microarray data from gastrointestinal stromal tumor (GIST) cell lines . We begin this analysis by loading the GIST expression measurements provided as an example data set with our CoGAPS R library.

```
> library('CoGAPS')
> data('GIST_TS_20084')
> ls()
> GISTDims <- dim(GIST.D)
```

This data set contains two `data.frame` objects: 1) the mean gene expression data in `GIST.D` ($\mathbf{D}$ in eq. (1)), and 2) the uncertainty on the means in `GIST.S` ($\boldsymbol{\sigma}$). The gene expression `data.frame` objects contain measurements for 1363 genes indexed by UniGene identifiers in rows and 9 timecourse samples along columns.

The CoGAPS analysis also requires gene sets. As required by the application of CoGAPS in the DESIDE algorithm, we load the set of genes regulated by transcription factors provided with our CoGAPS R library.

```
> data('TFGSList')
> ls()
> TFDims <- ncol(tf2ugFC)
```

The 230 gene sets defined from TRANSFAC are transcription factor targets stored in the `data.frame` object `tf2ugFC` in accordance with the format described in the CoGAPS User's Manual provided with the package. The user may also specify gene sets defined in a `list` object described in the User's Manual.

Because this example reproduces the analysis of [4], we will decompose the GIST data into 5 patterns as determined by empirical analysis with the Clutr-Free software [1], as described in detail in [2]. The new method for estimation of the number of factors from [3] can be used as well, and it will appear at http://www.biostat.jhsph.edu/∼jleek.

```
> nPattern <- 5
```

We also specify the number of MCMC iterations to be used for the GAPS matrix decomposition.

```
> nIter <- 50000000
```

With these preliminary data sets loaded and parameters specified, we can finally perform the corresponding CoGAPS analysis with a single call to the `CoGAPS` function. This single function call will first perform the GAPS matrix decomposition and then compute the corresponding Z-score indicating the transcription factor activity for each inferred pattern.

```
> results <- CoGAPS(data=GIST.D, unc=GIST.S,
+                   GStoGenes=tf2ugFC,
+                   numPatterns=5,
+                   SAIter = 2*nIter, iter = nIter,
+                   outputDir='GISTResults',
+                   plot=FALSE)
```

The `results` variable contains the following variables:

meanChi2 χ^2 value for fit of mean **A** and **P** matrices to the data

D gene expression data matrix

Sigma uncertainty in gene expression data matrix

Amean sample mean value of the MCMC **A** matrix estimates

Asd sample standard deviation value of the MCMC **A** matrix estimates

Pmean sample mean value of the MCMC **P** matrix estimates

Psd sample standard deviation value of the MCMC **P** matrix estimates

meanMock mock data formed by Amean $\times$ Pmean

GSUpreg Z-score derived p values for upregulation of each transcription factor for each pattern

GSDownreg Z-score derived p values for downregulation of each transcription factor for each pattern

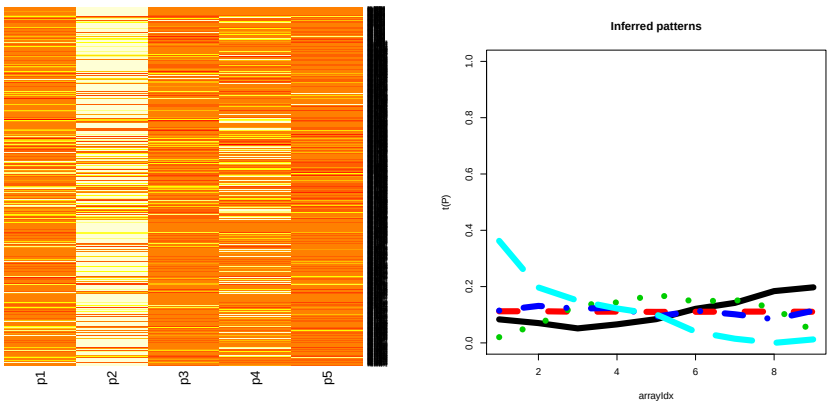
GSActEst Z-score derived estimates for the activity of each transcription factor for each pattern scaled from -1 (low activity) to $+1$ (high activity)

Generated with the `plotGAPS` function, Figure 1 displays the resulting estimates for the mean **A** and **P** matrices, which in this example have a fit to **D** of $\chi^2 = 23009.5483206579$.

```
> plotGAPS(results$Amean, results$Pmean,
+                   outputPDF='GISTResults/GIST_GAPS_Figs')
```

Consistent with [4], Figure 1 reveals two patterns relatively constant across the time-course samples, one falling with time, one transiently rising, and one continuously rising. The p values stored in `results$GSUpreg`, `results$GSDownreg`, and `results$GSActEst` indicate the statistical significance of transcription factor activity for each of these patterns computed using the Z-score statistic of [4]. This statistic is generated by first obtaining the mean Z-score for all genes in a set (i.e., all targets of a transcription factor),

$$Z_{t,p} = \frac{1}{R} \sum_{r \in G} \frac{A_{rp}}{\sigma_{rp}} \tag{1}$$



(a) Inferred amplitude matrix (b) Inferred patterns

Figure 1: Visualization of the (a) **A** and (b) **P** matrices inferred by GAPS for the gene expression measurements of [4].

where r indexes the genes, p indexes the patterns, G indicates the gene set, A_{rp} is the element of the **A** matrix for gene r and pattern p, and σ_{rp} is the corresponding element of the σ matrix. This is then a threshold independent statistic for the gene sets.

While $Z_{t,p}$ provides a statistic for all transcription factors t in all patterns p, it does not provide an estimate of significance. For that, we calculate Eqn. 1 for permuted labels of genes in gene sets (i.e., $r \in G_{permuted}$). This provides a null distribution for each pattern and for each number of genes in a gene set, as the number of elements in $G_{permuted}$ matches the number of elements in G. We then calculate a p-value for each transcription factor in each pattern. This can be done for both upregulation (i.e., higher than expected $Z_{t,p}$) and downregulation (i.e., lower than expected $Z_{t,p}$), which are stored in `GSUpreg` and `GSDownreg` respectively.

For transcription factors, it can also be useful to convert the p-value to an estimation of activity on a scale from $-1 \rightarrow 1$, with -1 suggesting low activity and $+1$ suggesting high activity. This is a simple rescaling of the p-values and is stored in `GSActEst`.

Table 1 lists the p-values for upregulation (`GSUpreg`) of each transcription factor in each pattern from pathway activity in GIST (Figure 3; [4]) and Table 2 for p-values of downregulation (`GSDownreg`). Finally, Table 3 lists the activity estimate of `GSActEst` for these transcription factors.

```
> EGFRTF <- c("c.Jun", "NF.kappaB", "Smad4", "Sp1", "STAT3", "Elk.1",
+     "c.Myc", "E2F.1", "AP.1", "CREB", "FOXO", "p53")
> EGFRSigUp <- matrix(0, nrow=length(EGFRTF), ncol=5)
> row.names(EGFRSigUp) <- EGFRTF
```

```
> colnames(EGFRSigUp) <- paste("p",1:5,sep="")
> EGFRSigDown <- matrix(0, nrow=length(EGFRTF), ncol=5)
> row.names(EGFRSigDown) <- EGFRTF
> colnames(EGFRSigDown) <- paste("p",1:5,sep="")
> EGFRSigAct <- matrix(0, nrow=length(EGFRTF), ncol=5)
> row.names(EGFRSigAct) <- EGFRTF
> colnames(EGFRSigAct) <- paste("p",1:5,sep="")
> for (i in 1:5) {
+   EGFRSigUp[,i] <- results$GSUpreg[i,EGFRTF]
+   EGFRSigDown[,i] <- results$GSDownreg[i,EGFRTF]
+   EGFRSigAct[,i] <- results$GSActEst[i,EGFRTF]
+ }
```

	p1	p2	p3	p4	p5
c.Jun	0.33	0.00	0.02	0.80	0.04
NF.kappaB	0.06	0.07	0.44	0.09	0.14
Smad4	0.65	0.47	0.02	0.92	0.30
Sp1	0.32	0.85	0.13	0.08	0.13
STAT3	0.08	0.05	0.19	0.80	0.38
Elk.1	0.01	0.97	0.12	0.37	0.00
c.Myc	0.26	0.99	0.52	0.07	0.00
E2F.1	0.47	0.98	0.27	0.56	0.10
AP.1	0.95	0.24	0.18	0.30	0.09
CREB	0.14	0.25	0.02	0.53	0.11
FOXO	0.31	0.52	0.27	0.02	0.92
p53	0.28	0.64	0.00	0.74	0.05

Table 1: p-values from `GSUpreg` for upregulation of each transcription factor in each pattern from pathway activity in GIST (Figure 3; [4])

All files containing results and diagnostics from the CoGAPS analysis are given in the directory `'GISTResults'` specified in the `outputDir` argument. Specifically, CoGAPS creates the following files

```
> list.files('GISTResults')

 [1] "Amean.0.2593600142.txt"
 [2] "AResults0.2593600142.Diagnostics.txt"
 [3] "Asd.0.2593600142.txt"
 [4] "GIST_GAPS_Figs_Amplitude.pdf"
 [5] "GIST_GAPS_Figs_Patterns.pdf"
 [6] "GSActEst.txt"
 [7] "GSDownStat.txt"
 [8] "GSUpStat.txt"
 [9] "Pmean.0.2593600142.txt"
```

	p1	p2	p3	p4	p5
c.Jun	0.67	1.00	0.98	0.20	0.96
NF.kappaB	0.94	0.93	0.56	0.91	0.86
Smad4	0.35	0.53	0.98	0.08	0.70
Sp1	0.68	0.15	0.87	0.92	0.87
STAT3	0.92	0.95	0.81	0.20	0.62
Elk.1	0.99	0.03	0.88	0.63	1.00
c.Myc	0.74	0.01	0.48	0.93	1.00
E2F.1	0.53	0.02	0.73	0.44	0.90
AP.1	0.05	0.76	0.82	0.70	0.91
CREB	0.86	0.75	0.98	0.47	0.89
FOXO	0.69	0.48	0.73	0.98	0.08
p53	0.72	0.36	1.00	0.26	0.95

Table 2: p-values from `GSDownreg` for downregulation of each transcription factor in each pattern from pathway activity in GIST (Figure 3; [4])

```
[10] "PResults0.2593600142.Diagnostics.txt"
[11] "Psd.0.2593600142.txt"
```

The files with the suffix `Diagnostics.txt` contain diagnostics from the GAPS analysis for the **A** and **P** matrices, based on the corresponding file prefix. We note each file name also contains a random identifier to link files from **A** and **P** matrices from a single GAPS decomposition. These files contain the parameters of the GAPS decomposition and a list of diagnostics at a subset of the MCMC iterations. This diagnostics notably include the χ^2 value for the MCMC fit at that iteration and number of atoms in the atomic domains for the corresponding **A** and **P** matrices. The file also includes summary statistics for the χ^2 value (23009.5483206579) for mean chain estimates of the **A** and **P** matrices and average number of atoms in each of these matrices after the burn-in iterations. Additionally, the results of the Z-score statistic indicating upregulation, downregulation, and activity of each transcription factor are provided in the files `GSUpStat.txt`, `GSDownStat.txt`, and `GSActEst.txt`, respectively. Specifically, these files contain the values from `results$GSUpReg`, `results$GSDownreg`, and `results$GSActEst` variables. Although not provided here, CoGAPS may also retain all the values of **A** and **P** with file names containing the same random identifier created during the MCMC sampling with the suffix `ChainValues.txt` as described in the User's Manual.

	p1	p2	p3	p4	p5
c.Jun	0.34	0.99	0.96	-0.60	0.92
NF.kappaB	0.88	0.86	0.13	0.82	0.73
Smad4	-0.29	0.06	0.95	-0.85	0.39
Sp1	0.37	-0.69	0.75	0.84	0.73
STAT3	0.83	0.90	0.63	-0.61	0.25
Elk.1	0.99	-0.94	0.77	0.27	0.99
c.Myc	0.49	-0.99	-0.04	0.86	0.99
E2F.1	0.05	-0.96	0.46	-0.12	0.79
AP.1	-0.89	0.53	0.64	0.39	0.82
CREB	0.72	0.50	0.96	-0.06	0.78
FOXO	0.39	-0.04	0.47	0.95	-0.84
p53	0.44	-0.27	1.00	-0.48	0.89

Table 3: Activity of each transcription factor from `GSActEst` in each pattern from pathway activity in GIST (Figure 3; [4])

Bibliography

[1] G. Bidaut and M.F. Ochs. Clutrfree: cluster tree visualization and interpretation. *Bioinformatics*, 20(16):2869–2871, 2004.

[2] G. Bidaut, K. Suhre, J.-M. Claverie, and M.F. Ochs. Determination of strongly overlapping signaling activity from microarray data. *BMC Bioinformatics*, 7:99, 2006.

[3] JT Leek. Asymptotic conditional singular value decomposition for high-dimensional genomic data. *Biometrics*, In Press, 2010.

[4] M.F. Ochs, L. Rink, C. Tarn, S. Mburu, T. Taguchi, B. Eisenberg, and A.K. Godwin. Detection of treatment-induced changes in signaling pathways in gastrointestinal stromal tumors using transcriptomic data. *Cancer Res*, 69(23):9125–9132, 2009.