Mirsynergy: detect synergistic miRNA regulatory modules by overlapping neighbourhood expansion

Yue Li
yueli@cs.toronto.edu

October 13, 2014

1 Introduction

MicroRNAs (miRNAs) are $\sim$22 nucleotide small noncoding RNA that base-pair with mRNA primarily at the $3'$ untranslated region (UTR) to cause mRNA degradation or translational repression [1]. Aberrant miRNA expression is implicated in tumorigenesis [4]. Construction of microRNA regulatory modules (MiRM) will aid deciphering aberrant transcriptional regulatory network in cancer but is computationally challenging. Existing methods are stochastic or require a fixed number of regulatory modules. We propose *Mirsynergy*, a deterministic overlapping clustering algorithm adapted from a recently developed framework. Briefly, Mirsynergy operates in two stages that first forms MiRM based on co-occurring miRNAs and then expand the MiRM by greedily including (excluding) mRNA into (from) the MiRM to maximize the synergy score, which is a function of miRNA-mRNA and gene-gene interactions (manuscript in prep).

2 Demonstration

In the following example, we first simulate 20 mRNA and 20 mRNA and the interactions among them, and then apply `mirsynergy` to the simulated data to produce module assignments. We then visualize the module assignments in Fig.1

```
> library(Mirsynergy)
> load(system.file("extdata/toy_modules.RData", package="Mirsynergy"))
> # run mirsynergy clustering
> V <- mirsynergy(W, H, verbose=FALSE)
> summary_modules(V)

$moduleSummaryInfo
  miRNA mRNA total   synergy    density
1     4    4    12 0.1680051 0.04426190
2     2    2     6 0.1654560 0.09630038
3     6   10    22 0.1870070 0.02471431
```

```
4      8     7    23 0.1821842 0.02318249
5      2     3     7 0.1640842 0.08457176
6      3     4    10 0.1602223 0.04856618

$miRNA.internal
  modules miRNA
1       2     2
2       1     3
3       1     4
4       1     6
5       1     8

$mRNA.internal
  modules mRNA
1       1    2
2       1    3
3       2    4
4       1    7
5       1   10
```

Additionally, we can also export the module assignments in a Cytoscape-friendly format as two separate files containing the edges and nodes using the function `tabular_module` (see function manual for details).

3 Real test

In this section, we demonstrate the real utility of *Mirsynergy* in construct miRNA regulatory modules from real breast cancer tumor samples. Specifically, we downloaded the test data in the units of RPKM (read per kilobase of exon per million mapped reads) and RPM (reads per million miRNA mapped) of 13306 mRNA and 710 miRNA for the 15 individuals from TCGA (The Cancer Genome Atlas). We furhter log2-transformed and mean-centred the data. For demonstration purpose, we used 20% of the expression data containing 2661 mRNA and 142 miRNA expression. Moreover, the corresponding sequence-based miRNA-target site matrix $\mathbf{W}$ was downloaded from TargetScanHuman 6.2 database [3] and the gene-gene interaction (GGI) data matrix $\mathbf{H}$ including transcription factor binding sites (TFBS) and protein-protein interaction (PPI) data were processed from TRANSFAC [6] and BioGrid [5], respectively.

```
> load(system.file("extdata/tcga_brca_testdata.RData", package="Mirsynergy"
```

Given as input the 2661×15 mRNA and 142×15 miRNA expression matrix along with the 2661×142 target site matrix, we first construct an expression-based miRNA-mRNA interaction score (MMIS) matrix using LASSO from *glmnet* by treating mRNA as response and miRNA as input variables [2].

```
> load(system.file("extdata/toy_modules.RData", package="Mirsynergy"))
> plot_modules(V,W,H)
```

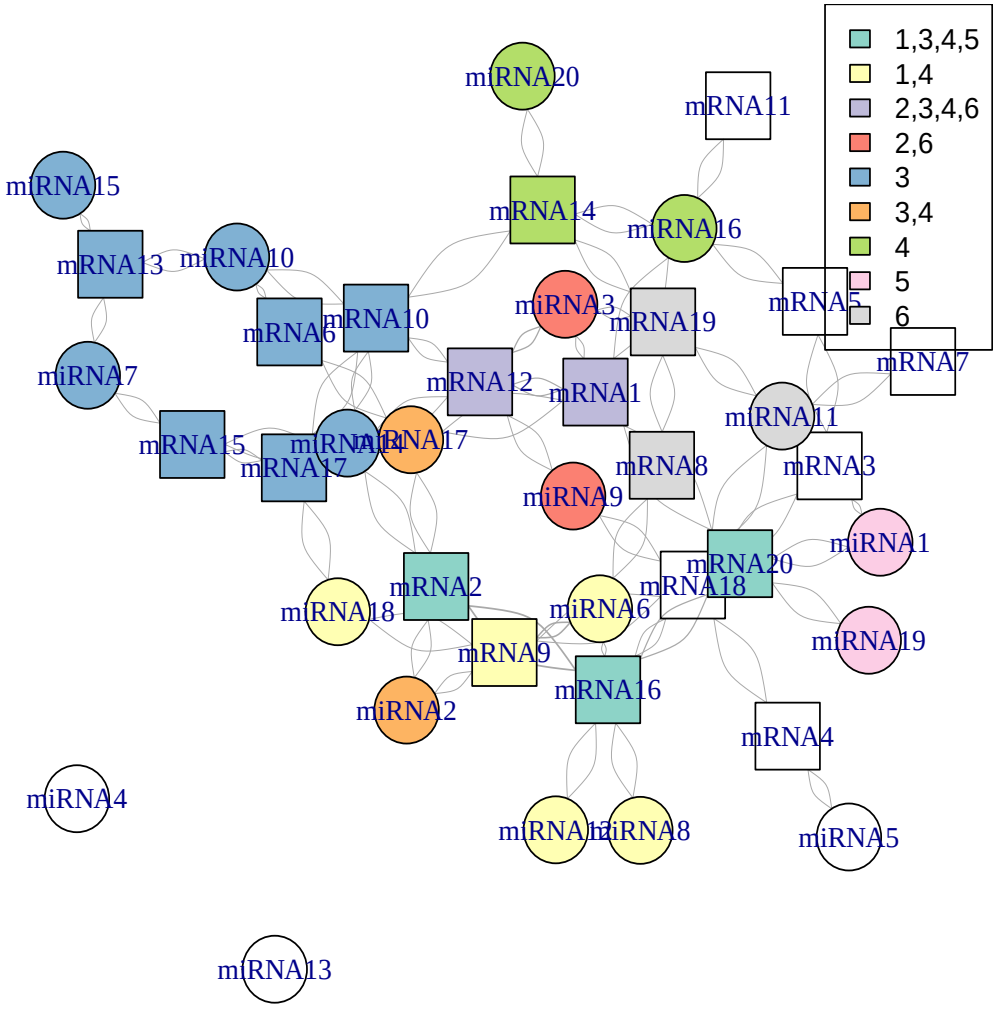

Figure 1: Module assignment on a toy example.

```
> library(glmnet)
> ptm <- proc.time()
> # lasso across all samples
> # X: N x T (input variables)
> #
> obs <- t(Z)  # T x M
> # run LASSO to construct W
> W <- lapply(1:nrow(X), function(i) {
+
+         pred <- matrix(rep(0, nrow(Z)), nrow=1,
+                 dimnames=list(rownames(X)[i], rownames(Z)))
+
+         c_i <- t(matrix(rep(C[i,,drop=FALSE], nrow(obs)), ncol=nrow(obs))
+
+         c_i <- (c_i > 0) + 0 # convert to binary matrix
+
+         inp <- obs * c_i
+
+         # use only miRNA with at least one non-zero entry across T sample
+         inp <- inp[, apply(abs(inp), 2, max)>0, drop=FALSE]
+
+         if(ncol(inp) >= 2) {
+
+                 # NOTE: negative coef means potential parget (remove inte
+                 x <- coef(cv.glmnet(inp, X[i,], nfolds=3), s="lambda.min"
+
+                 pred[, match(colnames(inp), colnames(pred))] <- x
+         }
+         pred[pred>0] <- 0
+
+         pred <- abs(pred)
+
+         pred[pred>1] <- 1
+
+         pred
+ })
> W <- do.call("rbind", W)
> dimnames(W) <- dimnames(C)
> print(sprintf("Time elapsed for LASSO: %.3f (min)",
+         (proc.time() - ptm)[3]/60))

[1] "Time elapsed for LASSO: 0.915 (min)"
```

Given the $\mathbf{W}$ and $\mathbf{H}$, we can now apply `mirsynergy` to obtain MiRM assignments.

```
> V <- mirsynergy(W, H, verbose=FALSE)
> print_modules2(V)

M1 (density=4.75e-02; synergy=2.37e-01):
hsa-miR-302a hsa-miR-520b hsa-miR-302e hsa-miR-3134
TRHDE FBXO41 MYCN SLC2A4 TRPV6 LRP8 FTSJD1 IDH1 BNC1 ZNF473
M2 (density=3.73e-02; synergy=2.16e-01):
hsa-miR-4311 hsa-miR-424 hsa-miR-1193 hsa-miR-759 hsa-miR-601
WDR43 LRRCC1 SEH1L FAM60A SLC2A14 PPP1R8 TAF7L PCDHA6
M3 (density=3.89e-02; synergy=1.91e-01):
hsa-miR-3183 hsa-miR-495 hsa-miR-519d hsa-miR-548m hsa-miR-448
ZC3HAV1L AQP4 AIF1L GFOD2 ZSCAN20 GABBR2 RFX4 PCDHA11
M4 (density=5.98e-02; synergy=1.89e-01):
hsa-miR-519e hsa-miR-4313 hsa-miR-4290
RCBTB2 TRAF4 ERP44 PNOC CD40
M5 (density=2.49e-02; synergy=2.14e-01):
hsa-miR-30b hsa-miR-4284 hsa-miR-921 hsa-miR-3125 hsa-miR-3714 hsa-miR-4252
IRAK3 RAB27B FOXM1 STAC RHPN2 UBAP2L TGIF2 TMEM194B ELFN2 YEATS2 PGM3 ABCG8
M6 (density=3.23e-02; synergy=1.44e-01):
hsa-miR-1912 hsa-miR-3201 hsa-miR-216a hsa-miR-548n hsa-miR-3125
XPO5 ERC2 KDM5A
M7 (density=2.67e-02; synergy=2.07e-01):
hsa-miR-340 hsa-miR-3161 hsa-miR-214 hsa-miR-210 hsa-miR-3135
GIPC2 SIAH1 SMAD9 CBFB ZMYND11 ACADSB HIC1 EZH1 LMO4 ITSN1 MITF PALLD AFF1 S
M8 (density=7.43e-02; synergy=1.67e-01):
hsa-miR-3692 hsa-miR-1271 hsa-miR-3929
FECH RFX4 C10orf54
M9 (density=2.82e-02; synergy=2.1e-01):
hsa-miR-626 hsa-miR-181c hsa-miR-3155 hsa-miR-214 hsa-let-7e
ATF1 CLP1 FREM2 LOR TSEN34 PROX1 RNF8 TRANK1 GALK2 SLC1A4 PLEK ANXA11 RORA N
M10 (density=5.81e-02; synergy=1.98e-01):
hsa-miR-4328 hsa-miR-216a hsa-miR-939
RRP1B LMO4 ITSN1 PAPD7 DEPDC1 KIF1B NUP210
M11 (density=9.25e-02; synergy=1.75e-01):
hsa-miR-608 hsa-miR-4293
FAM107A KCNQ4
M12 (density=8.6e-02; synergy=1.92e-01):
hsa-miR-891b hsa-miR-1322
CBFB ZNF644 CSDE1 RUNX1
M13 (density=9.31e-02; synergy=1.68e-01):
hsa-miR-185 hsa-miR-625
GEMIN8 NFIX
M14 (density=6.22e-02; synergy=2.92e-01):
hsa-miR-4271 hsa-miR-18b hsa-miR-335 hsa-miR-1301
UBE2E2 UBE2V1 TRIM23 RNF165 RNF4 RNF8 RNF114 UBE2D1 UEVLD ZNRF1 SMG5 ARIH2 N
```

```
M15 (density=6.24e-02; synergy=2.52e-01):
hsa-miR-513b hsa-miR-1915
KIAA1161 C6orf170 GPR126 BOLL CDC25A ABCA13 NUPL1 DMD FIGF KIF26A
M16 (density=2.28e-02; synergy=2.3e-01):
hsa-miR-626 hsa-miR-181c hsa-miR-4262 hsa-miR-181d hsa-miR-3155 hsa-miR-214
ATF1 CLP1 DPP6 FREM2 LOR USP6NL CNKSR3 TSEN34 PROX1 RNF8 TRANK1 GATA6 SLC1A4
M17 (density=9.11e-02; synergy=1.27e-01):
hsa-miR-1254 hsa-miR-661
GJB1
M18 (density=5.03e-02; synergy=1.02e-01):
hsa-miR-137 hsa-miR-3154
RFX5 PPPDE2 C9orf150
M19 (density=2.43e-02; synergy=1.47e-01):
hsa-miR-340 hsa-miR-3161 hsa-miR-1297 hsa-miR-210 hsa-miR-3135
IRAK3 GIPC2 ACADSB AGPAT5 ITPR2 PLXND1 PALLD FGF1 SYT1
M20 (density=6.71e-02; synergy=1.23e-01):
hsa-miR-586 hsa-miR-595
ZFP1 CNOT1
M21 (density=2.07e-02; synergy=2.07e-01):
hsa-miR-374c hsa-miR-3183 hsa-miR-3692 hsa-miR-4308 hsa-miR-3665 hsa-miR-495
DCLK2 ZC3HAV1L AQP4 AIF1L GFOD2 GABBR2 SYNM FECH RFX4 ONECUT1 PCDHA11 C10orf
```

```
> print(sprintf("Time elapsed (LASSO+Mirsynergy): %.3f (min)",
+   (proc.time() - ptm)[3]/60))

[1] "Time elapsed (LASSO+Mirsynergy): 0.998 (min)"
```

There are several convenience functions implemented in the package to generate summary information such as Fig.2. In particular, the plot depicts the m/miRNA distribution across modules (upper panels) as well as the synergy distribution by itself and as a function of the number of miRNA (bottom panels).

For more details, please refer to our paper (manuscript in prep.).

4 Session Info

```
> sessionInfo()

R version 3.1.1 Patched (2014-09-25 r66681)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
```

```
> plot_module_summary(V)
```

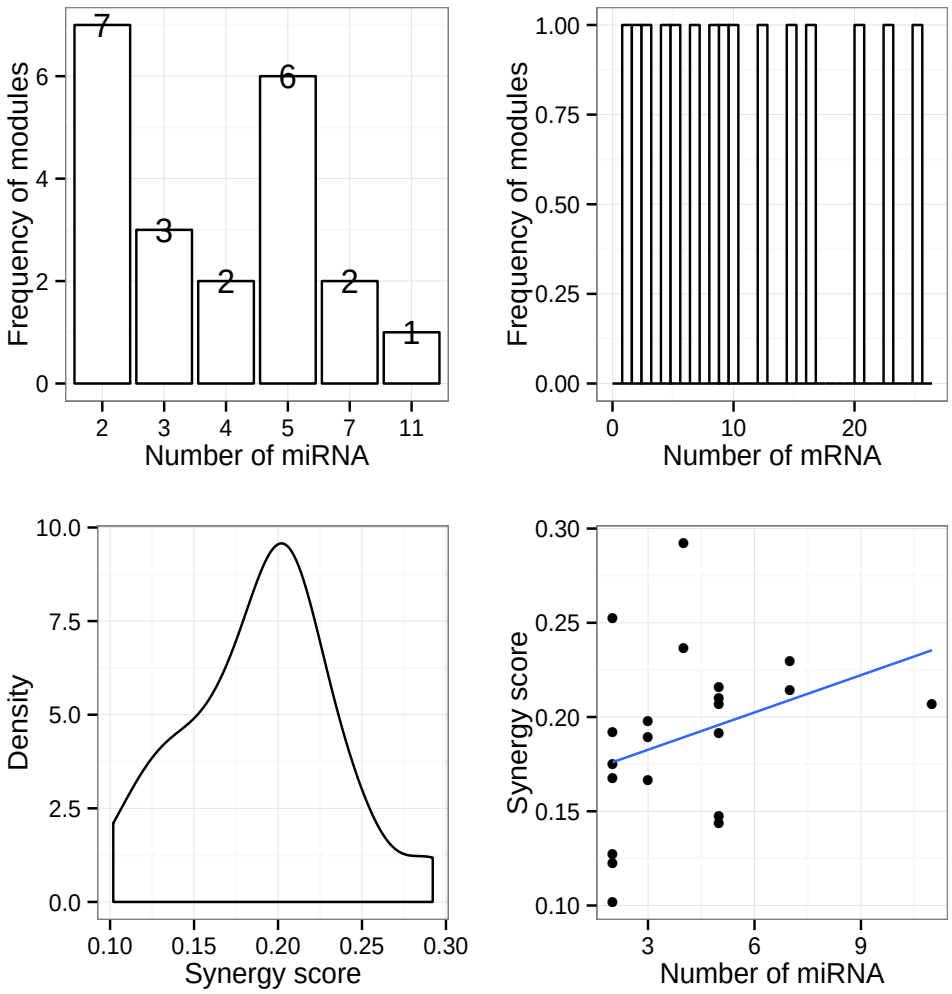

Figure 2: Summary information on MiRM using test data from TCGA-BRCA. Top panels: m/miRNA distribution across modulesas; Bottom panels: the synergy distribution by itself and as a function of the number of miRNA.

```
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] glmnet_1.9-8     Matrix_1.1-4     Mirsynergy_1.2.0 ggplot2_1.0.0
[5] igraph_0.7.1

loaded via a namespace (and not attached):
 [1] MASS_7.3-35        RColorBrewer_1.0-5 Rcpp_0.11.3        colorspace_1.2
 [5] digest_0.6.4       evaluate_0.5.5     formatR_1.0        grid_3.1.1
 [9] gridExtra_0.9.1    gtable_0.1.2       knitr_1.7          labeling_0.3
[13] lattice_0.20-29    munsell_0.4.2      parallel_3.1.1     plyr_1.8.1
[17] proto_0.3-10       reshape_0.8.5      reshape2_1.4       scales_0.2.4
[21] stringr_0.6.2      tools_3.1.1
```

References

[1] David P Bartel. MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, 136(2):215–233, January 2009.

[2] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, 33(1):1–22, 2010.

[3] Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, and David P Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105, January 2009.

[4] Riccardo Spizzo, Milena S Nicoloso, Carlo M Croce, and George A Calin. SnapShot: MicroRNAs in Cancer. *Cell*, 137(3):586–586.e1, May 2009.

[5] Chris Stark, Bobby-Joe Breitkreutz, Andrew Chatr-Aryamontri, Lorrie Boucher, Rose Oughtred, Michael S Livstone, Julie Nixon, Kimberly Van Auken, Xiaodong Wang, Xiaoqi Shi, Teresa Reguly, Jennifer M Rust, Andrew Winter, Kara Dolinski, and Mike Tyers. The BioGRID Interaction Database: 2011 update. *Nucleic acids research*, 39(Database issue):D698–704, January 2011.

[6] E Wingender, X Chen, R Hehl, H Karas, I Liebich, V Matys, T Meinhardt, M Prüss, I Reuter, and F Schacherer. TRANSFAC: an integrated system for gene expression regulation. *Nucleic acids research*, 28(1):316–319, January 2000.