Vignette for NEM-Tar:a method for cancer regulatory network inference and prioritization of potential therapeutic targets

Yuchen ZHANG,Lina ZHU,Xin WANG

January 27, 2021

Contents

1 Introduction

Cancers are not single disease entities, but comprising multiple molecularly distinct subtypes, and the heterogeneity prevents precise selection of patients for optimized therapy. Dissecting cancer subtype-specific signaling pathways is crucial to pinpointing dysregulated genes for the prioritization of novel therapeutic targets. Nested effects models (NEMs) are a group of graphical models that encode subset relations between observed downstream effects under perturbations to upstream signaling genes, providing a prototype for mapping the inner workings of the cell. In this study, we developed NEM-Tar, which extends the original NEMs to predict drug targets by incorporating causal information of (epi)genetic aberrations for signaling pathway inference. An information theory-based score, weighted information gain (WIG), was proposed to assess the impact of signaling genes on a specific downstream biological process. We will show in detail how to implement a toy example for singaling network inference as well as a real case study for prioritizing the potential therapeutic targets.

2 A brief overview of NEMs and NEM-Tar

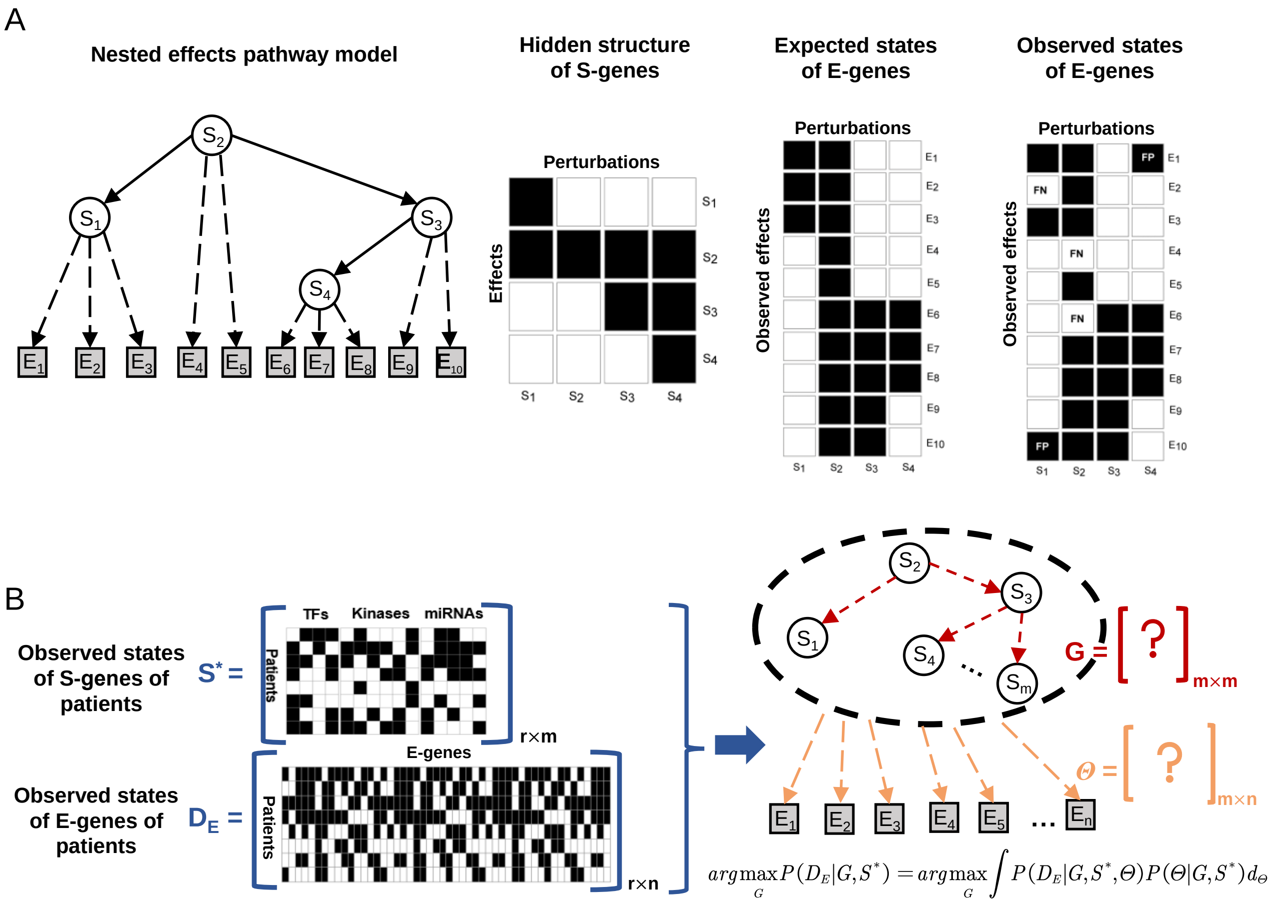

A
Nested effects pathway model
S_2
S_1
S_3
S_4
E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}
Hidden structure of S-genes
Perturbations
Effects
S1 S2 S3 S4
Expected states of E-genes
Perturbations
Observed effects
E1 E2 E3 E4 E5 E6 E7 E8 E9 E10
Observed states of E-genes
Perturbations
Observed effects
FP
FN
B
Observed states of S-genes of patients
$S^* =$
TFs Kinases miRNAs
Patients
r×m
Observed states of E-genes of patients
$D_E =$
E-genes
Patients
r×n
S_1 S_2 S_3 S_4 ... S_m
$G = [\,?\,]_{m\times m}$
$\Theta = [\,?\,]_{m\times n}$
E_1 E_2 E_3 E_4 E_5 ... E_n
$\underset{G}{argmax} P(D_E|G,S^*) = \underset{G}{argmax} \int P(D_E|G,S^*,\Theta)P(\Theta|G,S^*)d_\Theta$

Figure 1. Comparison between the structures of (A) classic NEM and (B) our proposed NEM-Tar

As shown in Figure 1A, in classic nested effects models, the S-genes are modeled as hidden variables, and their signaling interaction graph G (solid arrows) is the target to infer. In experiments with perturbations to individual S-genes, differential expression of downstream genes could be observed and considered as effect reporter genes (E-genes). Assuming that each E-gene is directly regulated by at most one S-gene in G, the maximum a posteriori attachment Θ (dashed arrows) of effect genes to S-genes could be computed. The goal is to search for the signaling graph G, which yields the most likely probabilistic nested effects. Illustrated in Figure 1B, an extra observational dimension (the real patients) is the factor that NEM-Tar should deal with.The necessary adjustment should be conducted on the design and inference strategies of classic NEM. However, the information needs to infer is also the hidden interaction between S-genes and the attachment relationship of E-genes to S-genes.And the likelihood function of NEM-Tar is similar to that of NEMs, except the state matrix of regulators (S-genes) S* in our model.

3 Network inference on a toy example

3.1 Introduction of the in-silico data

The applicaitons on real case studies of NEM-Tar require a lot of data preprocssing and integrative selection of singaling genes(S-genes) and effect reporter genes(E-genes).For the purpose of interpreting the main work flow and contribution of NEM-Tar.At first, we will introduce the employed in-silico data.

```
library(nemTar)
```

```
## Loading required package: nem
```

```
## Loading required package: dplyr
```

```
## Warning: replacing previous import 'vctrs::data_frame' by 'tibble::data_frame'
## when loading 'dplyr'
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(dplyr)
data("example")
```

The toy example contains four different elements.The S-gene number was applied in medium size(12), thus the dimension of Sgene_hidden, storing the unobserved S-gene states was 12×12; Edata is the E-gene profiles with the dimension 804×100, S_obs is the observed profile of S-gene states ‘after’ perturbations,with the dimension of 100×12.

```
dim(Edata)
```

```
## [1] 804 100
```

```
dim(S_obs)
```

```
## [1] 100  12
```

```
dim(Sgene_hidden)
```

```
## [1] 12 12
```

```
para
```

```
## [1] 0.05 0.05
```

3.2 Network Inference Using Greedy hill-climbing

```
control<-nem::set.default.parameters(Sgenes=rownames(Sgene_hidden),type="mLL",para=para)
nemTar_rslt<-nem_Tar_greedy(Edata,Sgenes=control$Sgenes,S_obs,control=control)
```

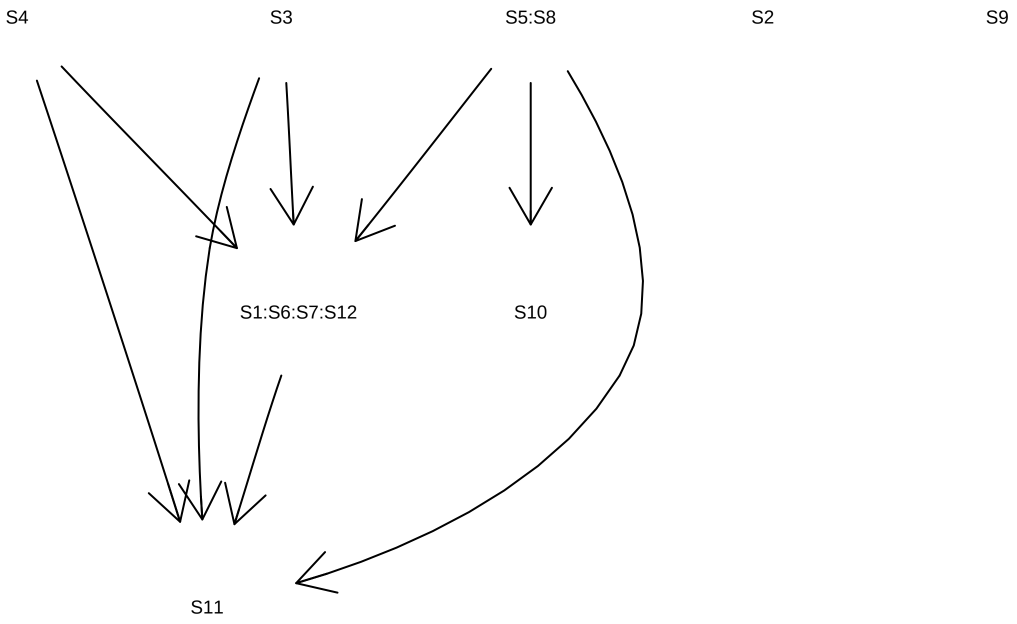
```
## Greedy hillclimber for 12 S-genes (lambda = 0 )...
##
## 132  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7
## --> Edge added, removed or reversed
## 132  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7
## --> Edge added, removed or reversed
## 130  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7
## --> Edge added, removed or reversed
## 129  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7
## --> Edge added, removed or reversed
## 128  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7
## --> Edge added, removed or reversed
## 123  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7
## --> Edge added, removed or reversed
## 117  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7
## --> Edge added, removed or reversed
## 115  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7
## --> Edge added, removed or reversed
## 108  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7
## --> Edge added, removed or reversed
## 66  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7
## --> Edge added, removed or reversed
## 53  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7
## --> Edge added, removed or reversed
## 49  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  S1 S2 S3 S4 S5 S6 S7
## log-likelihood of model =  -36204.99
```

This process takes a few seconds.

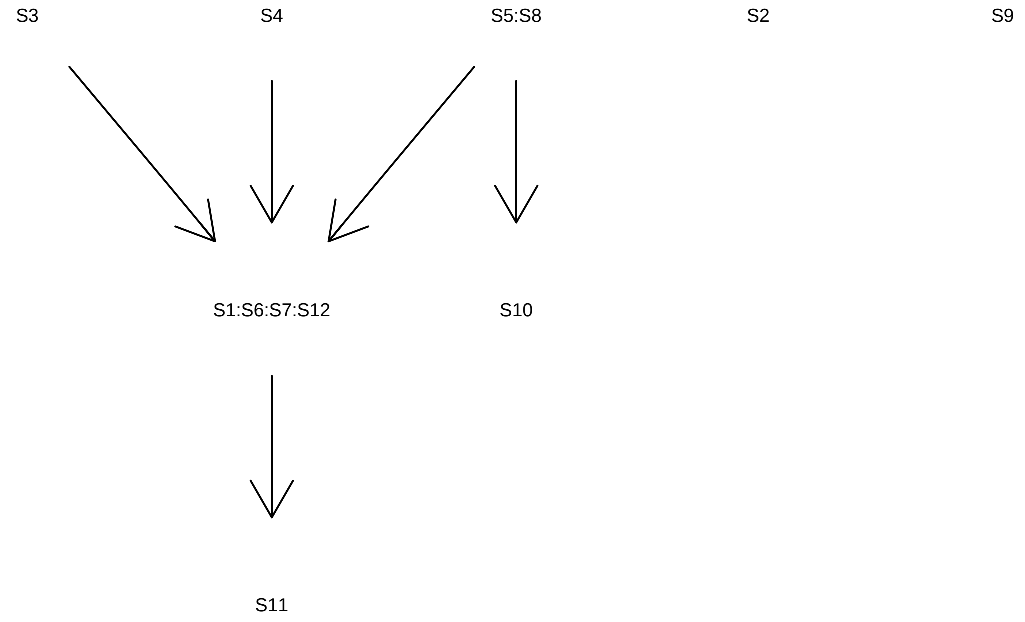
3.3 Visulization of the inferred S-gene network

To visualize the inference results, an R package RedeR is suggested.However, the function plot.nem could give a more quickly visulization.The error matrix is the difference between the inferred S-gene matrix and the generated S-gene matrix, the result that all the entries is 0 indicates that the inference is perfect.

```
plot.nem(nemTar_rslt,what="graph")
```



```
plot.nem(nemTar_rslt,what="graph",transitiveReduction=T)
```



```
adj<-as(nemTar_rslt$graph, "matrix")
error<-Sgene_hidden-adj
print(error)
```

```
##     S1 S2 S3 S4 S5 S6 S7 S8 S9 S10 S11 S12
## S1   0  0  0  0  0  0  0  0  0   0   0   0
## S2   0  0  0  0  0  0  0  0  0   0   0   0
## S3   0  0  0  0  0  0  0  0  0   0   0   0
## S4   0  0  0  0  0  0  0  0  0   0   0   0
## S5   0  0  0  0  0  0  0  0  0   0   0   0
## S6   0  0  0  0  0  0  0  0  0   0   0   0
## S7   0  0  0  0  0  0  0  0  0   0   0   0
## S8   0  0  0  0  0  0  0  0  0   0   0   0
## S9   0  0  0  0  0  0  0  0  0   0   0   0
## S10  0  0  0  0  0  0  0  0  0   0   0   0
## S11  0  0  0  0  0  0  0  0  0   0   0   0
## S12  0  0  0  0  0  0  0  0  0   0   0   0
```

4 Case study I-Inferring the Signaling Network Driving the EMT Subtype of Gastric Cancer and Prioritization of Potential Drug Targets

4.1 Introduction of the input profiles of GC

The profiles of GC contains three different elements.The S-gene number was determined as 14, thus the dimension of adjacency matrix of S-gene interacction network was 14×14; Edata_GC_ori and D are the E-gene profiles before and after the transformation to binary variable,with the dimension of 1194×40 and 818×27, respectively; Sdata_GC is the observed profile of 'natural' perturbation states of Sgenes with the dimension of 27×14.

```
# load the input profiles of GC
library(nemTar)
data("case_GC")
data("EMT_list")
```

4.2 Transforming the E-gene profiles into binary variable and network inference

```
res.disc <- nem.discretize(Edata_GC_ori,neg.control=1:10,pos.control=11:13,nfold=2,cutoff= 0.5)
```

```
## discretizing with respect to POS and NEG controls
```

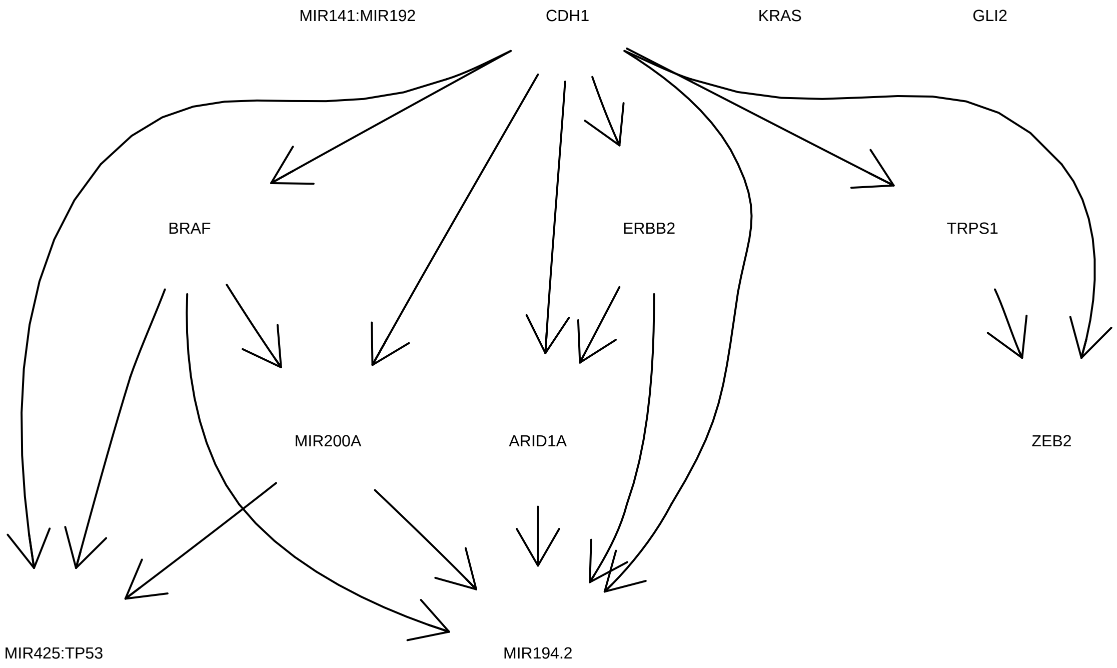
```
D<-res.disc$dat
para<-res.disc$para
control<-set.default.parameters(Sgenes=Sgenes_GC,type="mLL",para=para)
nemTar_GC<-nem_Tar_greedy(D=D,Sgenes=Sgenes_GC,S_pattern=Sdata_GC,control=control)
```

```
## Greedy hillclimber for 14 S-genes (lambda = 0 )...
##
## 182  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 182  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 180  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 179  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 177  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 176  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 174  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 173  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 170  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 169  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 163  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 161  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 139  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## --> Edge added, removed or reversed
## 116  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200A MIR192
## log-likelihood of model =  -13956.07
```

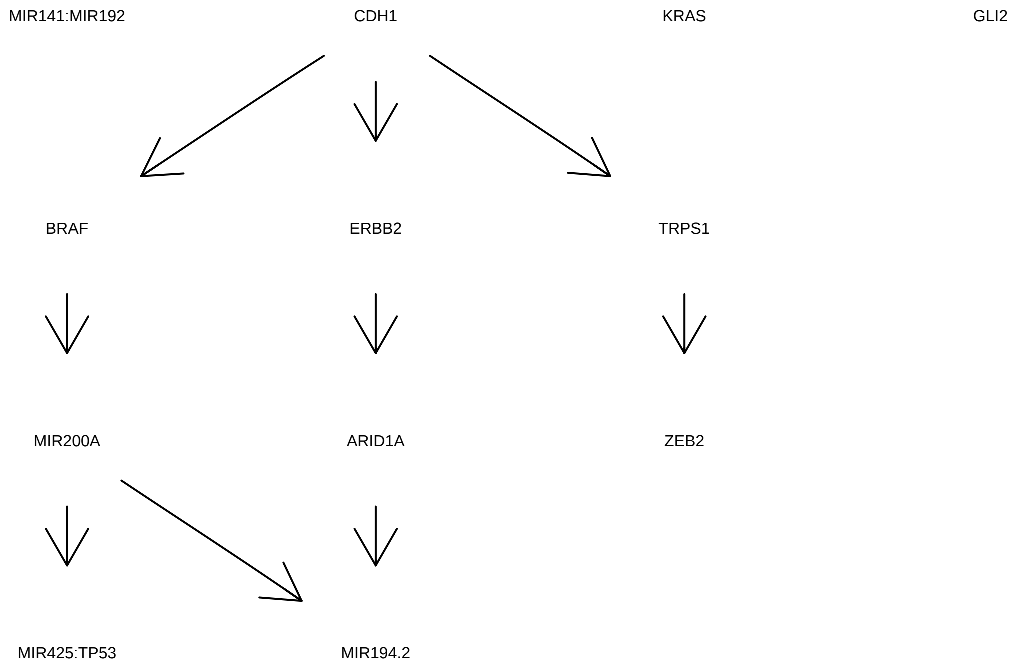
4.3 Visulization of the inferred S-gene network

4.3.1 Quick visulization

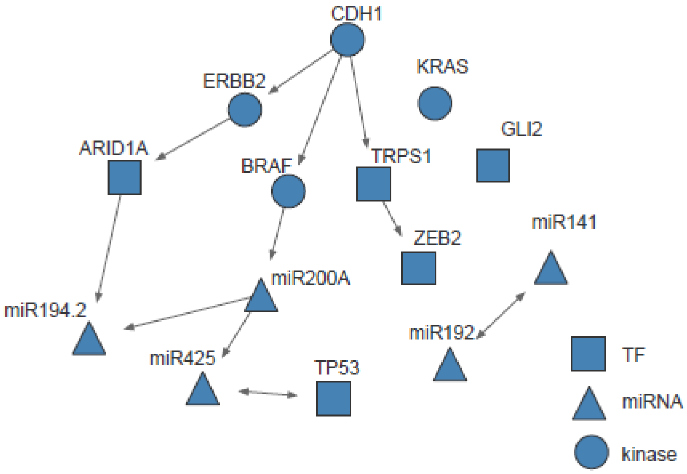
```
plot.nem(nemTar_GC,what="graph")
```



```r
plot.nem(nemTar_GC,what="graph",transitiveReduction=T)
```



4.3.2 Employing the 'RedeR' package to visualize the network(after adjustment of the layout,color etc.)



4.4 Assessment of the influence of S-gene perturbations on EMT process

The statistical significance of the causal effect that the S-genes exert on downstream pathways(EMT pathway here) is quantified by permutation tests, i.e.,random sampling of E-genes with the same number of EMT signature genes in the regulon of a S-gene, and calculating the frequency of observing a same or higher WIG from the sampled E-gene sequences.(The value 0 of the adjusted P stands for the corresponding P value is less than the resolution of current sampling times,in the following case P<1e-05.)

```
EMT_post<-path_post(nemTar_GC,EMT_list,Sgenes_GC)
WIG<-compute_WIG(EMT_post$post_affected,EMT_post$path_affected,14)
sample_WIG0<-WIG_sample(EMT_post,nemTar_GC,EMT_list,14,1e05)
sig_test<-WIGsig_test(sample_WIG0,WIG,1e05)
### results summary
sig_rslt<-data.frame(S_genes=names(EMT_post$Sig_affected)[which(lengths(EMT_post$path_affected)!=0)],
                     WIG=round(WIG$WIG,2),adjusted_P=format(round(sig_test,6),
                                                            scientific = TRUE,digits = 3))
sig_rslt[which(sig_rslt$adjusted_P==0),3]<-'<1e-05'
colnames(sig_rslt)<-c("S genes","WIG","Adjusted P")
sig_rslt<-sig_rslt[order(sig_rslt$WIG,decreasing=TRUE),]
knitr::kable(sig_rslt,align="l",row.names=F,caption="Assessment of single S-gene perturbation
             on EMT in GC")
```

Table 1: Assessment of single S-gene perturbationon EMT in GC

S genes	WIG	Adjusted P
CDH1	64.21	0.00e+00
GLI2	44.34	1.71e-01
ERBB2	20.90	4.50e-05
ARID1A	15.32	2.97e-01
BRAF	10.72	3.09e-03
TRPS1	8.29	1.28e-02
MIR200A	8.20	4.02e-01
ZEB2	4.98	8.99e-01
KRAS	0.80	5.63e-01

Also, the weigheted information gain(WIG) of combinational perturbation of 2 S-genes could be calcualted.The kinases of double perturbation with higher WIGs could be the candidate combinational therapeutic targets.

```
Sgene_double_WIG<-WIG_double(EMT_post,14)
### results summary
sig_rslt<-data.frame(S_genes=names(Sgene_double_WIG$WIG),
                     WIG=round(Sgene_double_WIG$WIG,2))
colnames(sig_rslt)<-c("S genes","WIG")
sig_rslt<-sig_rslt[order(sig_rslt$WIG,decreasing=TRUE),]
knitr::kable(sig_rslt[match(c("CDH1/ERBB2","KRAS/CDH1","BRAF/CDH1","BRAF/ERBB2",
                              "KRAS/ERBB2","KRAS/BRAF"),rownames(sig_rslt)),],
             align="l",row.names=F,caption="Assessment of double perturbations
             (kinase only) on EMT in GC")
```

Table 2: Assessment of double perturbations (kinase only) on EMT in GC

S genes	WIG
CDH1/ERBB2	66.37
KRAS/CDH1	65.01
BRAF/CDH1	64.21
BRAF/ERBB2	28.81
KRAS/ERBB2	21.71
KRAS/BRAF	11.52

5 Case study II-Inferring the Signaling Network Driving the CMS4-mesenchymal Subtype of Colorectal Cancer(CRC) and Prioritization of Potential Drug Targets

5.1 Introduction of the input profiles of CRC

The profiles of GC contains three different elements.The S-gene number was prioritized as 15, thus the dimension of adjacency matrix of S-gene interacction network was 15×15; Edata_GC_ori and D are the E-gene profiles before and after the transformation to binary variable,with the dimension of 1337×93 and 1292×54,respectively; Sdata_CRC is the observed profile of 'natural' perturbation states of S-genes with the dimension 54×15.

```
library(nemTar)
data("case_CRC")
data("EMT_list")
```

5.2 Transforming the E-gene profiles into binary variable and network inference

```
res.disc <- nem.discretize(Edata_CRC_ori,neg.control=1:27,pos.control=28:39,nfold=2,cutoff= 0.6)
```

```
## discretizing with respect to POS and NEG controls
```

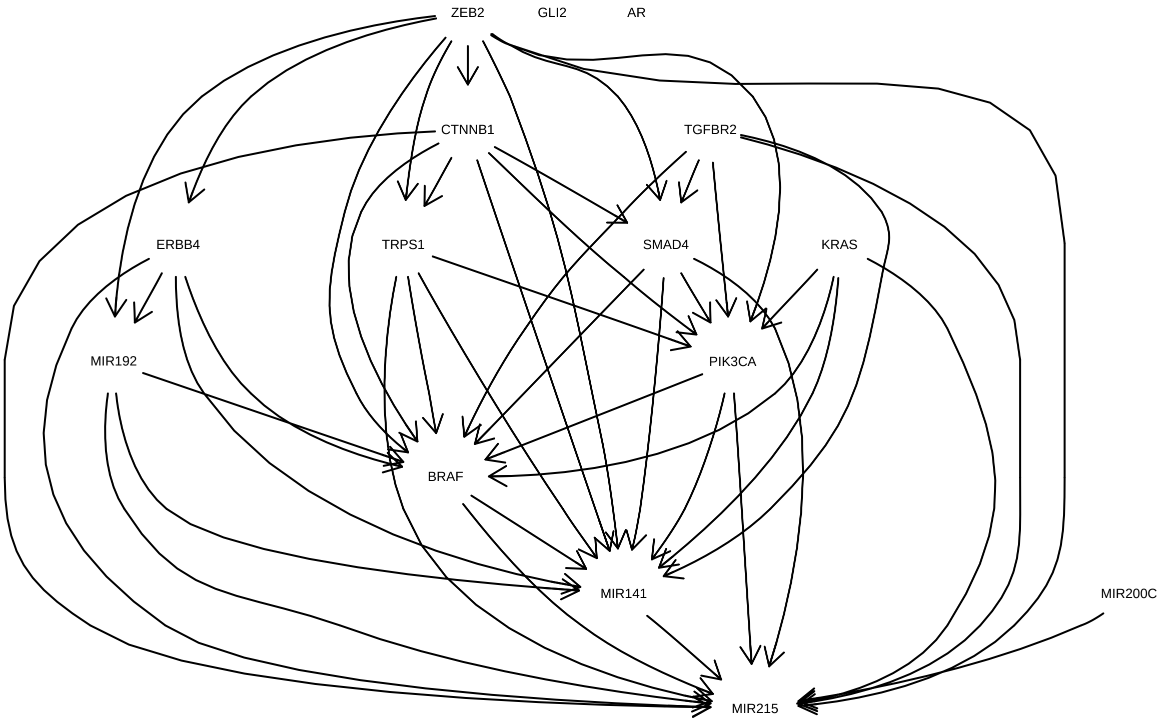
```
D<-res.disc$dat
para<-res.disc$para
control<-set.default.parameters(Sgenes=Sgenes_CRC,type="mLL",para=para)
nemTar_CRC<-nem_Tar_greedy(D,Sgenes=control$Sgenes,Sdata_CRC,control=control)
```

```
## Greedy hillclimber for 15 S-genes (lambda = 0 )...
##
## 210  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200C MIR192
## --> Edge added, removed or reversed
## 210  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200C MIR192
## --> Edge added, removed or reversed
## 209  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200C MIR192
## --> Edge added, removed or reversed
## 207  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200C MIR192
## --> Edge added, removed or reversed
## 205  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200C MIR192
## --> Edge added, removed or reversed
## 204  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200C MIR192
## --> Edge added, removed or reversed
## 200  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200C MIR192
## --> Edge added, removed or reversed
## 194  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200C MIR192
## --> Edge added, removed or reversed
## 193  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200C MIR192
## --> Edge added, removed or reversed
## 191  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200C MIR192
## --> Edge added, removed or reversed
## 189  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200C MIR192
## --> Edge added, removed or reversed
## 179  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200C MIR192
## --> Edge added, removed or reversed
## 168  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200C MIR192
## --> Edge added, removed or reversed
## 167  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200C MIR192
## --> Edge added, removed or reversed
## 165  local models to test ...
## ((Marginal) posterior likelihood difference of best vs. second best model for  MIR141 MIR200C MIR192
## log-likelihood of model =  -28772.92
```

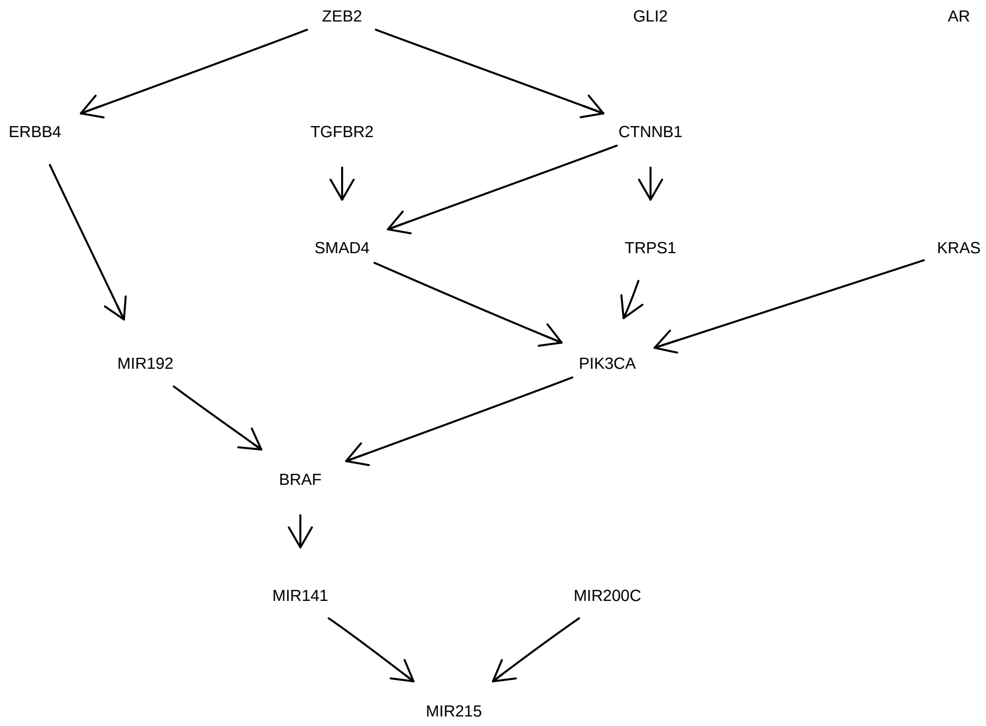
5.3 Visulization of the inferred S-gene network

5.3.1 Quick visulization

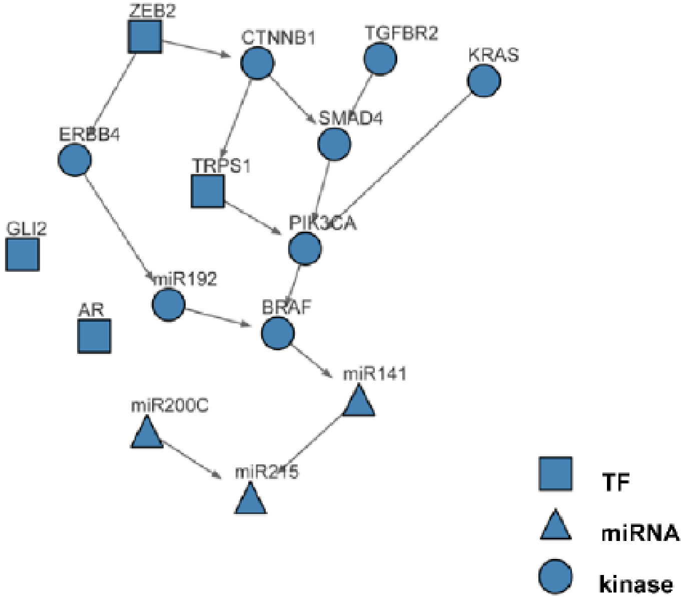
```
plot.nem(nemTar_CRC,what="graph")
```



```
plot.nem(nemTar_CRC,what="graph",transitiveReduction=T)
```



5.3.2 Employing the 'RedeR' package to visualize the network(after adjustment of the layout,color etc.)



5.4 Assessment of the influence of S-gene perturbations on EMT process

The statistical significance of the causal effect that the S-genes exert on downstream pathways(EMT pathway here) is quantified by permutation tests, i.e.,random sampling of E-genes with the same number of EMT signature genes in the regulon of a S-gene, and calculating the frequency of observing a same or higher WIG from the sampled E-gene sequences.(The value 0 of the adjusted P stands for the corresponding P value is less than the resolution of current sampling times,in the following case P<1e-05.)

```r
EMT_post<-path_post(nemTar_CRC,EMT_list,Sgenes_CRC)
WIG<-compute_WIG(EMT_post$post_affected,EMT_post$path_affected,15)
sample_WIG0<-WIG_sample(EMT_post,nemTar_CRC,EMT_list,15,1e05)
sig_test<-WIGsig_test(sample_WIG0,WIG,1e05)
### results summary
sig_rslt<-data.frame(S_genes=names(EMT_post$Sig_affected)[which(lengths(EMT_post$path_affected)!=0)],
                     WIG=round(WIG$WIG,2),adjusted_P=format(round(sig_test,6),
                                                           scientific = TRUE,digits = 3))
colnames(sig_rslt)<-c("Sgenes","WIG","Adjusted P")
sig_rslt[which(sig_rslt[3]==0),3]<-"<1e-05"
sig_rslt<-sig_rslt[order(sig_rslt$WIG,decreasing=TRUE),]
knitr::kable(sig_rslt,align="l",row.names=F,caption="Assessment of the impact of single S-gene
             perturbation on EMT in CRC")
```

Table 3: Assessment of the impact of single S-gene perturbation on EMT in CRC

Sgenes	WIG	Adjusted P
ZEB2	48.80	0.00e+00
TGFBR2	43.55	2.00e-04
CTNNB1	17.89	0.00e+00
AR	15.37	5.70e-01
KRAS	12.59	5.71e-03
TRPS1	12.23	3.02e-02
SMAD4	10.98	2.07e-03
GLI2	5.35	2.58e-01
PIK3CA	5.32	3.54e-01
ERBB4	2.54	2.63e-01
MIR200C	1.53	2.98e-01
MIR192	0.73	4.10e-01

Similar to study in GC, the weigheted information gain(WIG) of combinational perturbation of 2 Sgenes could be calcualted.The kinases of double perturbation with higher WIGs could be the candidate combinational therapeutic targets.

```r
Sgene_double_WIG<-WIG_double(EMT_post,15)
### results summary
sig_rslt<-data.frame(S_genes=names(Sgene_double_WIG$WIG),
                     WIG=round(Sgene_double_WIG$WIG,2))
colnames(sig_rslt)<-c("S genes","WIG")
sig_rslt<-sig_rslt[order(sig_rslt$WIG,decreasing=TRUE),]
knitr::kable(sig_rslt[match(c("KRAS/TGFBR2","TGFBR2/CTNNB1","TGFBR2/ERBB4","PIK3CA/TGFBR2",
                              "SMAD4/TGFBR2","KRAS/CTNNB1","ERBB4/CTNNB1","KRAS/SMAD4",
                              "PIK3CA/CTNNB1","SMAD4/CTNNB1"),rownames(sig_rslt)),],
             align="l",row.names=F,caption="Assessment of double perturbations
             (kinase only) on EMT in CRC")
```

Table 4: Assessment of double perturbations (kinase only) on EMT in CRC

S genes	WIG
KRAS/TGFBR2	50.81
TGFBR2/CTNNB1	50.46
TGFBR2/ERBB4	46.09
PIK3CA/TGFBR2	43.55
SMAD4/TGFBR2	43.55
KRAS/CTNNB1	25.15
ERBB4/CTNNB1	19.70
KRAS/SMAD4	18.24
PIK3CA/CTNNB1	17.89
SMAD4/CTNNB1	17.89