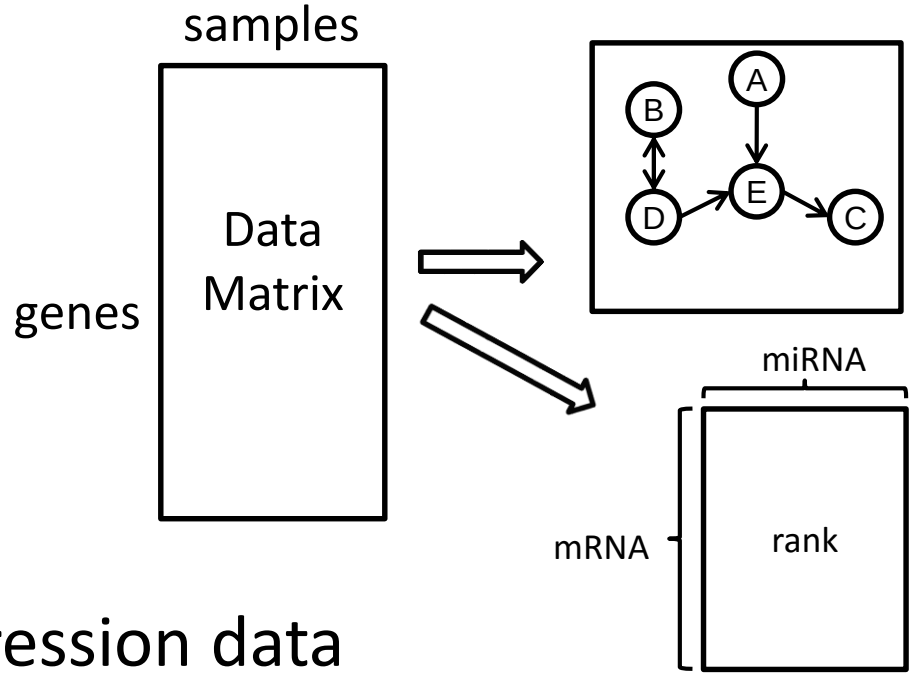
Inference of gene regulation from expression datasets

Phd Work of: Yiqian Zhou

Advisor: Ahmet Sacan

Overview

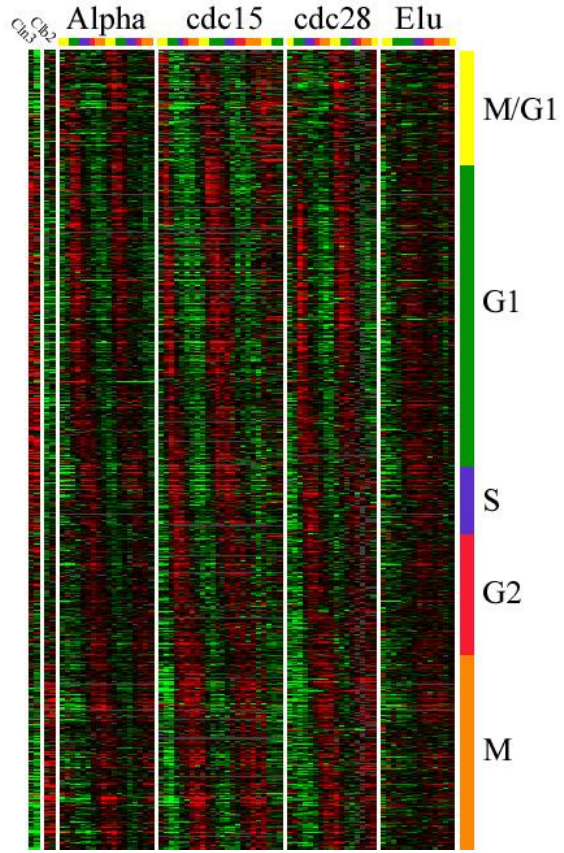


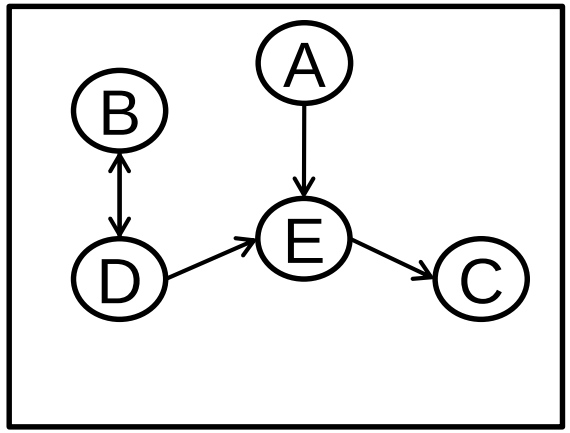
- Time-series gene expression data
  - ➔Aim1: Gene regulatory network & simulation

- Paired miRNA-mRNA expression data
  - ➔ Aim2: miRNA-mRNA interaction
  - ➔ Aim3: miRNA functional annotation

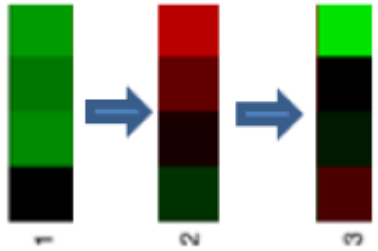
Aim1: Time-series gene expression data

➔Gene regulatory network & Simulation





Network Reconstruction

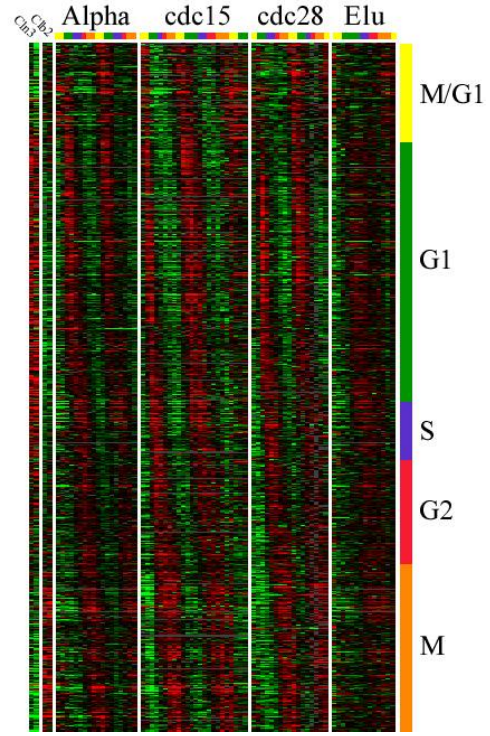


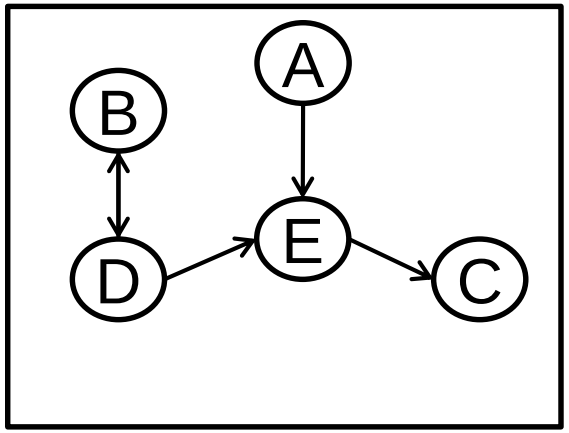
Prediction and simulation

(Y Zhou, R Qureshi, A Sacan, 2012)

Reconstruction from microarray data -- Related Work:

- Correlation networks
- Differential equation models
- Boolean network
- Bayesian network

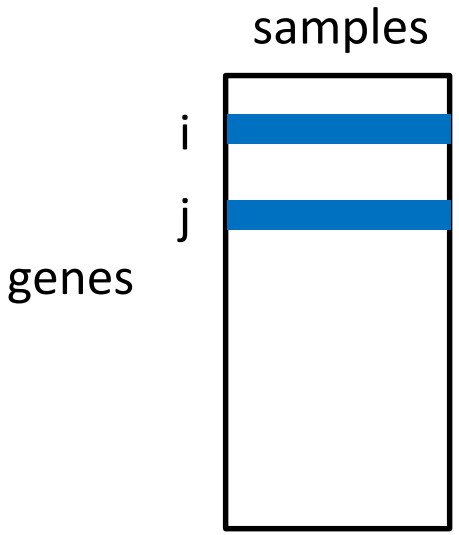


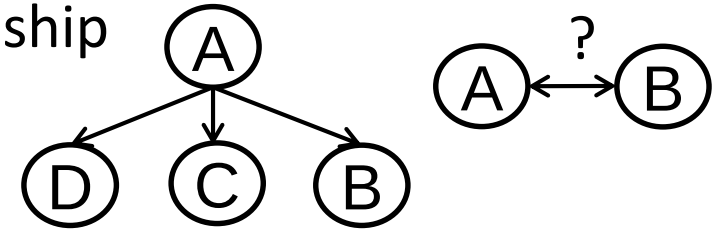


(Y Zhou, R Qureshi, F Bell, A Sacan, 2014)

Correlation networks

- Similarity between transcript pairs
  - Pearson correlation
  - Mutual information
- Advantages:
  - Simple
  - Computationally cost-effective
  - Low data requirement
- Disadvantages:
  - Co-regulation and causal relationship
  - Only pairwise interaction





(Marbach et al. 2009)

Boolean network

- binary variables
  - 1: transcripts is expressed;
  - 0: not expressed
- Boolean function

samples

t			t+1		
v_1	v_2	v_3	v_1'	v_2'	v_3'
0	0	0	0	0	1
0	0	1	0	0	1
0	1	0	1	0	1
0	1	1	1	0	1
1	0	0	0	0	0
1	0	1	0	1	0
1	1	0	1	0	0
1	1	1	1	1	0

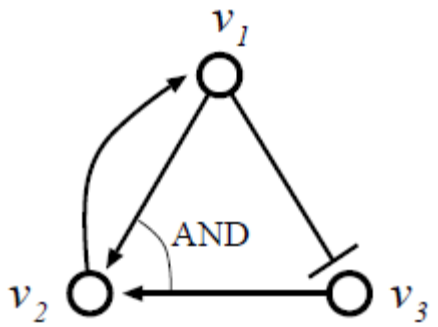
$$x_i(t+1) = f_i^{\mathrm{B}}\left(x_1(t), \ldots, x_N(t)\right)$$

⇨

$$v_1' = v_2$$

$$v_2' = v_1 \ \mathbf{AND} \ v_3$$

$$v_3' = \mathbf{NOT} \ v_1$$

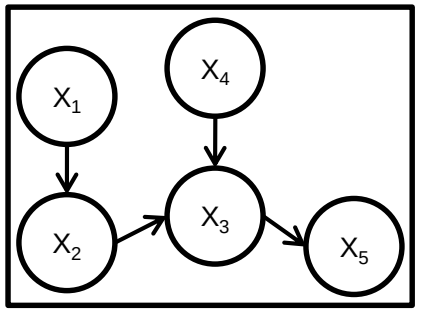


- Difficulties:
  - Discretization of data
  - Requirement of large amount of data: n regulator, 2^n combination

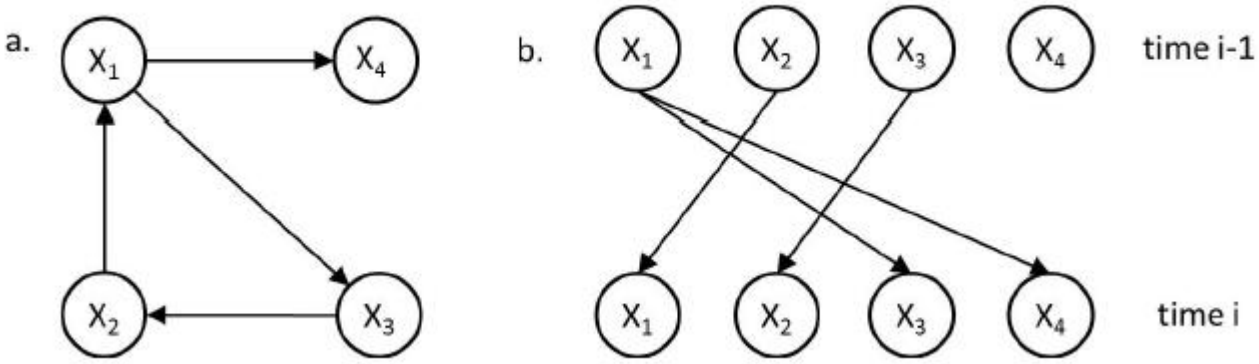
Bayesian network

- Each gene is a random variable, determined by probability distribution function that is expressed as a product of conditional probabilities

$$P(X_1, X_2, X_3, ..., X_n) = \prod_{i=1}^{n} P(X_i | Pt(X_i))$$



- Find a network that best represents the data.
  - NP-hard problem
  - heuristic search methods: greedy-hill climbing, Markov Chain Monte Carlo, and simulated annealing



DBN (Kim, Imoto et al. 2003)

Differential equation models

- Model $$\frac{\mathrm{d}x_i}{\mathrm{d}t} = f_i(x_1, \ldots, x_N),$$
- Influence function
  - Linear
  - non-linear
- Advantages
  - Complex relation
  - Quantitative
- Disadvantages
  - Requires large volume of high quality data

Multiple linear regression model

$$y_i = \beta_0 + \sum_{j=1}^{N} \beta_{ij} x_j + \varepsilon_i$$

- linearity assumption
- Interpretability
- Ease of computation
- Many-to-one regulation
- Prediction of expression values

Multiple linear regression model for time series

Time series data for each gene

	0 min	10 min	20 min	…
Gene A	-0.46	0.15	0.71	
Gene B	-0.21	0.19	0.86	
Gene C	-0.73	-0.42	0.01	
Gene D	-0.16	-0.3	-0.33	
Gene E	0.28	-0.36	-0.03	

$$y_i = \beta_0 + \sum_{j=1}^{N} \beta_{ij} x_j + \varepsilon_i$$

$$g_t^A = w^0 + \sum_{i=1..N} w^i g_{t-1}^i$$

Predictor genes

	0 min	10 min	20 min	…
Gene B	-0.21	0.19	0.86	
Gene C	-0.73	-0.42	0.01	
Gene D	-0.16	-0.3	-0.33	
Gene E	0.28	-0.36	-0.03	

Response gene

	10 min	20 min	30 min	…
Gene A	0.15	0.71	0.06	

Problem:

Underdetermined system: fewer samples than variables

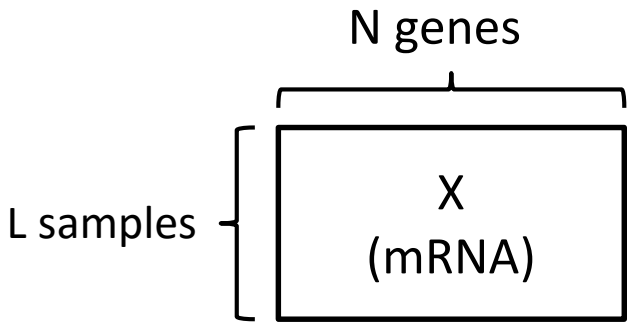
$$y_i = \beta_0 + \sum_{j=1}^{N} \beta_{ij} x_j + \varepsilon_i$$

$$\boldsymbol{y}_i = \boldsymbol{X}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i$$

Least squares

$$\min_{\boldsymbol{\beta}} \{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \}$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$



Multiple Linear Stepwise Regression (SMLR)

- Forward Selection

  1. Test each predictor not selected so far.
  2. Add the predictor with the best SSE to the model, repeat Step 1.

$$g_t^A = w^0 + ?$$

$$F = \frac{SSE^* - SSE}{SSE/(n-p-1)}$$

$$SSE = \sum (y_i - \hat{y}_i)^2$$

$$SSE^* = \sum (y_i - \hat{y}_i^*)^2$$

$$g_t^A = w^0 + w^B g_{t-1}^B$$
$$g_t^A = w^0 + w^C g_{t-1}^C$$
$$\dots$$
$$g_t^A = w^0 + w^B g_{t-1}^B + ?$$

SSE^*: sum of squared error of reduced model using only p predictor
SSE: sum of squared error of the expanded model using $p + 1$ predictor variables

(Hadi 2006)

Scoring the predictors

	10 min	20 min	30 min	…
Gene A	0.15	0.71	0.06	

(Repeat with B, C, D, E)

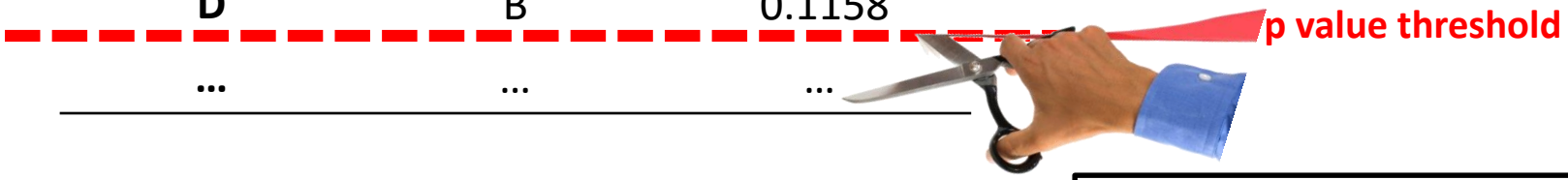
	0 min	10 min	20 min	…
Gene B	-0.21	0.19	0.86	
Gene C	-0.73	-0.42	0.01	
Gene D	-0.16	-0.3	-0.33	
Gene E	0.28	-0.36	-0.03	

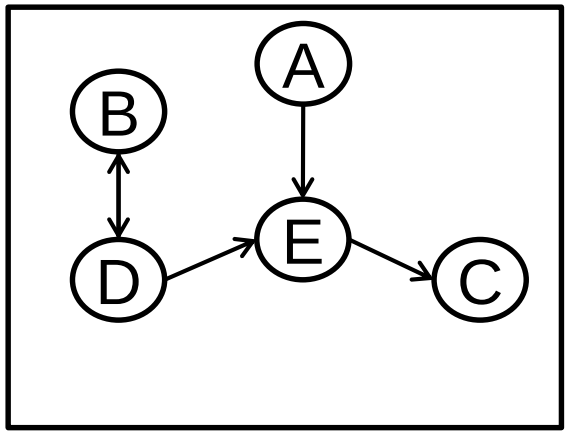
Predictor Gene	Response Gene	P value
D	E	6e-004
E	C	0.0036
A	E	0.1013
B	D	0.1102
D	B	0.1158
…	…	…

Table of predicted interactions

Network Reconstruction

Predictor Gene	Response Gene	P value
D	E	6e-004
E	C	0.0036
A	E	0.1013
B	D	0.1102
D	B	0.1158
…	…	…





Experiments on Yeast Cell Cycle Data

- Cell cycle is synchronized
- 4 datasets used, each named after the synchronization method.
  - Temperature sensitive mutants: cdc15, cdc28.
  - Alpha factor arrest.
  - Centrifugal elutriation (ELU).

TABLE 1. CHO/SPELLMAN DATA SETS AND THEIR PROPERTIES[a]

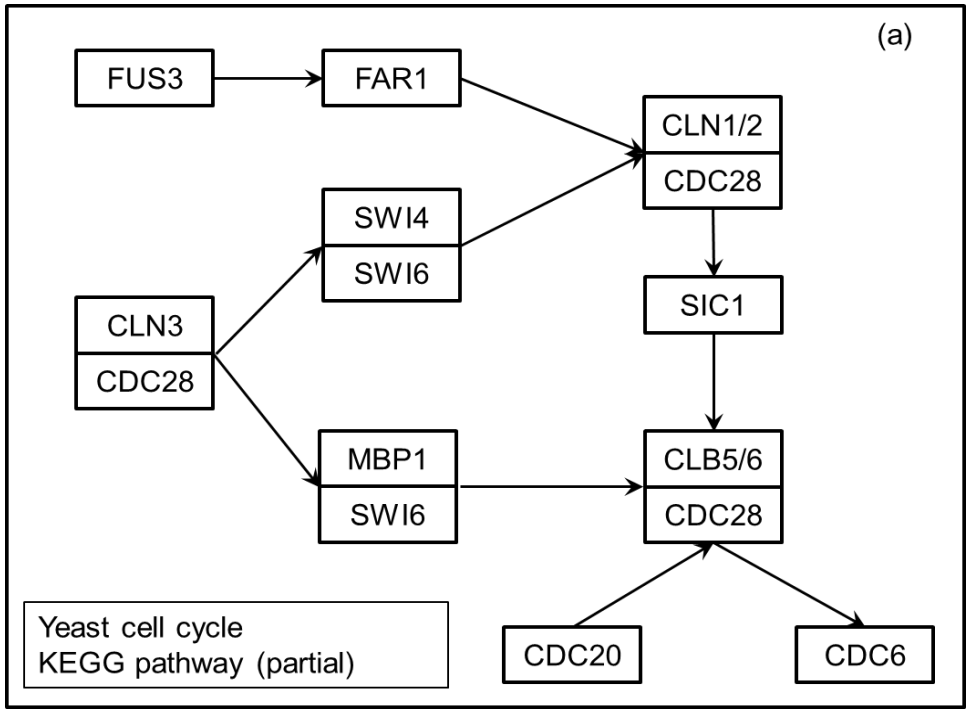
Data set	*Period obs.*	*Period det.*	*δt*	*# samples*	*# full orfs*
alpha	66 ± 11 min.	70 ± 7 min.	7	18	3361
cdc28	90 ± 10 min.	100 ± 10 min.	10	17	1188
cdc15	70 ± 10 min.	90 ± 10 min.	10/20	24	3453
elu	—	—	30	14	4753

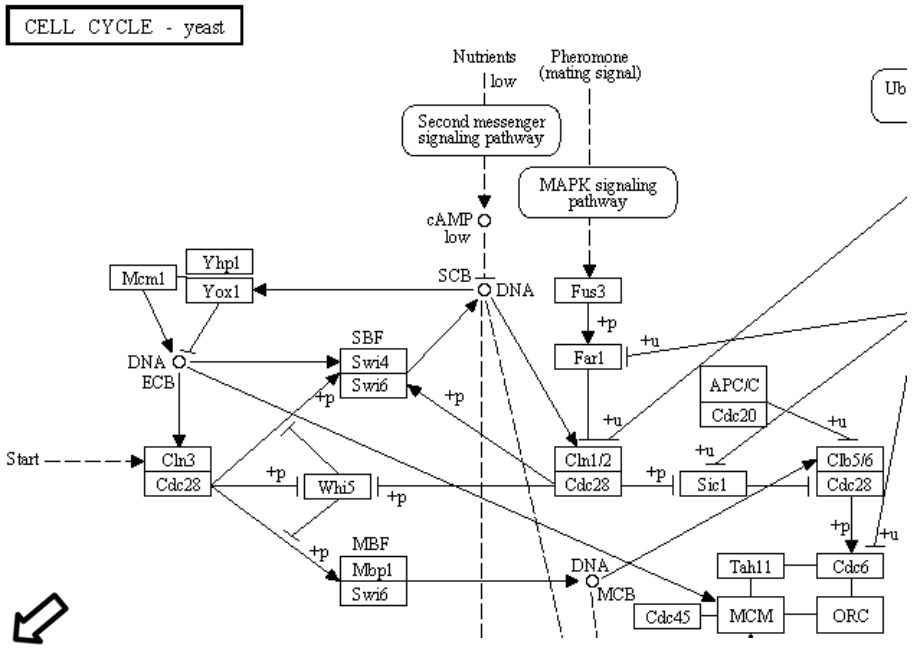
Total # of unique orf: 6178

(Cho et al. 1998, Spellman et al. 1998)

15

Target Pathway: Cell Cycle Network





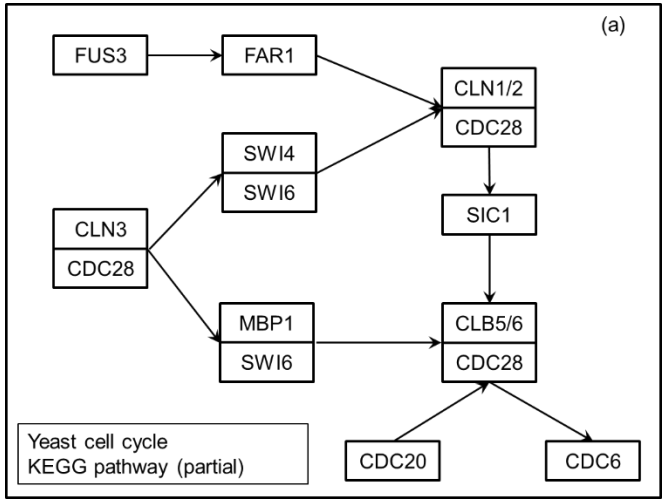
KEGG

- 14 genes
- 51 edges

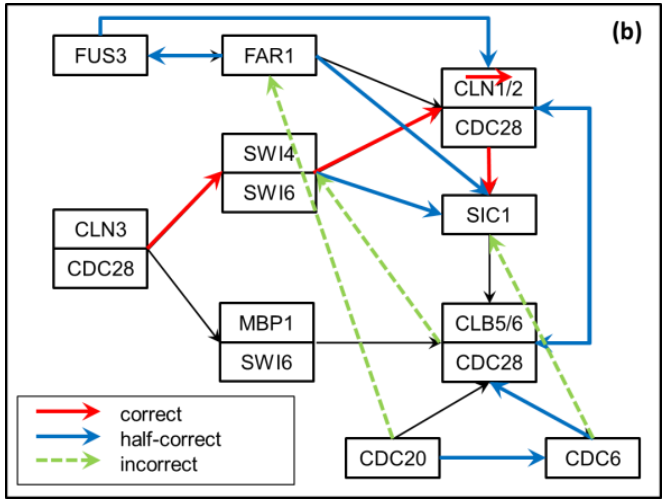
16

Comparison of reconstructed network

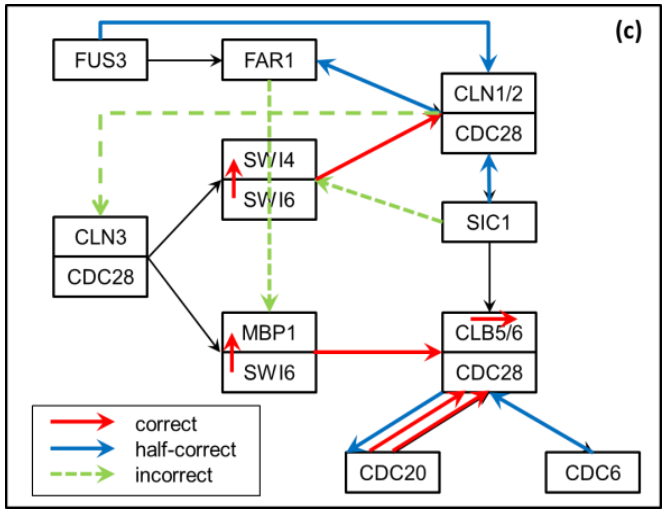
KEGG



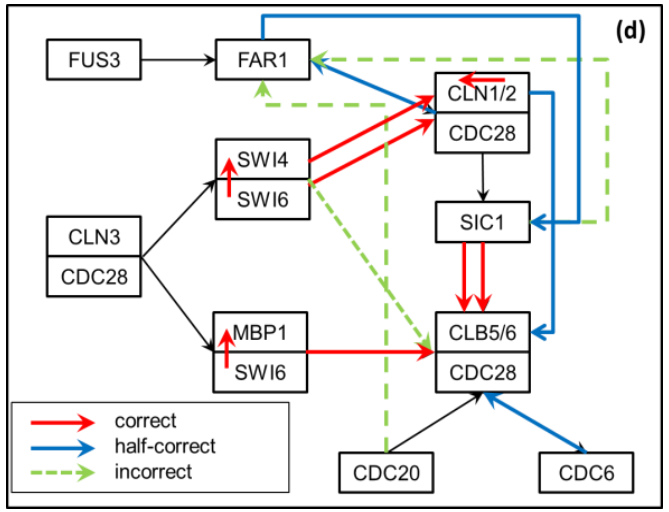
DBN



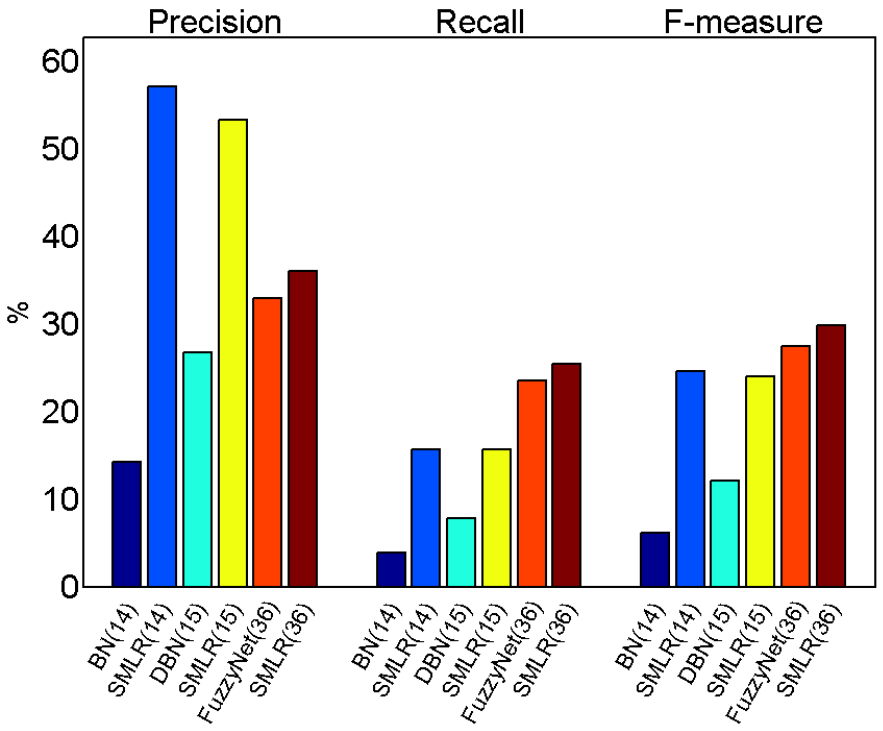
SMLR (CDC28)

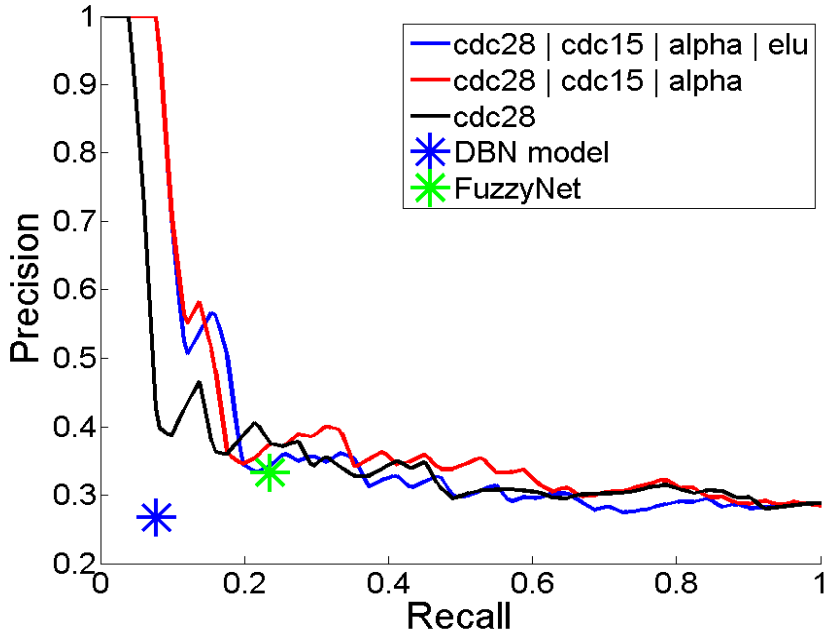


SMLR (integrated)



Reconstruction Performance





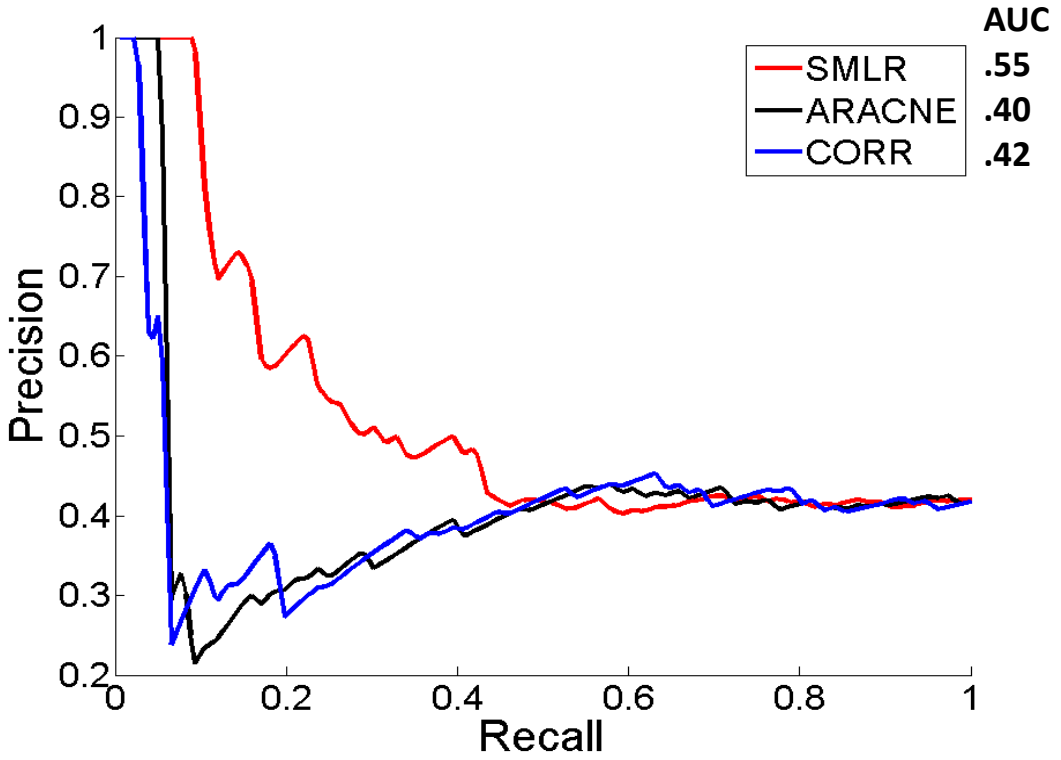
- $precision \stackrel{\text{def}}{=} \frac{\#\ of\ correctly\ predicted\ edges}{\#\ of\ predicted\ edges}$
- $recall \stackrel{\text{def}}{=} \frac{\#\ of\ correctly\ predicted\ edges}{\#\ of\ edges\ in\ the\ known\ network}$
- $F-measure \stackrel{\text{def}}{=} 2 * \frac{precision * recall}{precision + recall}$

BN: 14 edges
DBN: 15 edges (Kim, Imoto et al. 2003)
FuzzyNet: 36 edges
(Maraziotis, Dragomir et al. 2005)
recurrent neuro-fuzzy method

Comparison to “Static Learners”

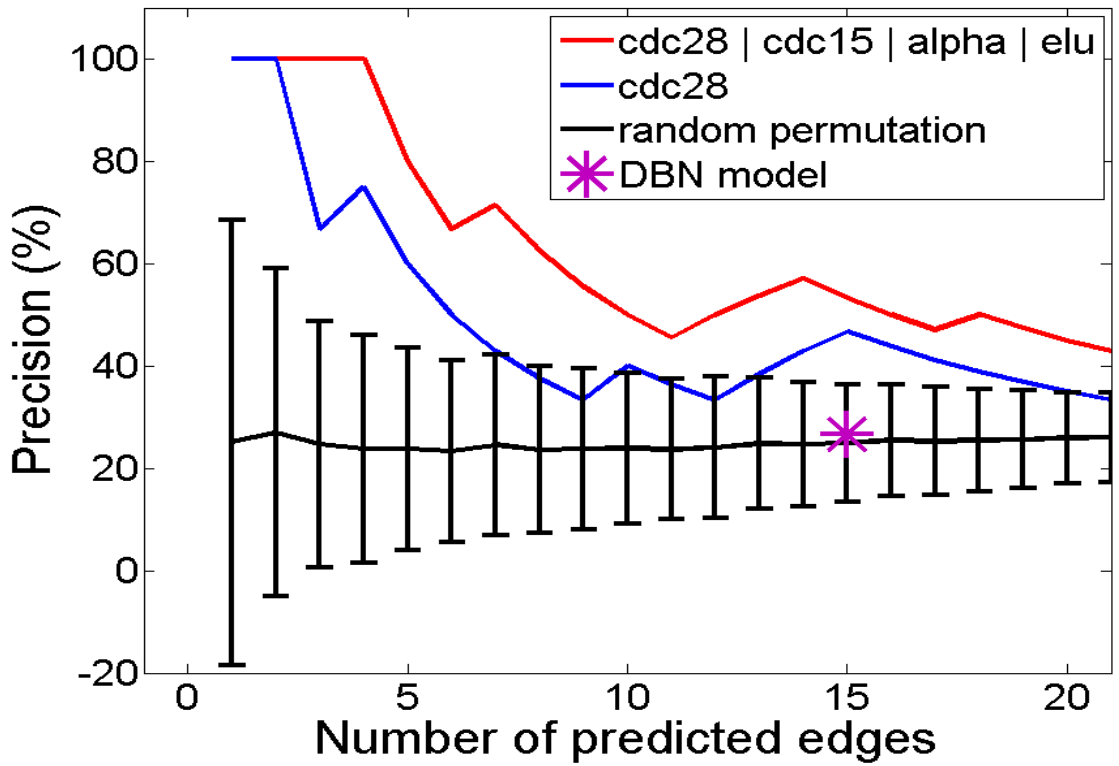
mutual information learner

correlation learner



ARACNE (Margolin, Nemenman et al. 2006)
based on mutual information (MI) calculation

Comparison to predictions from randomized data



DBN (Kim, Imoto et al. 2003)

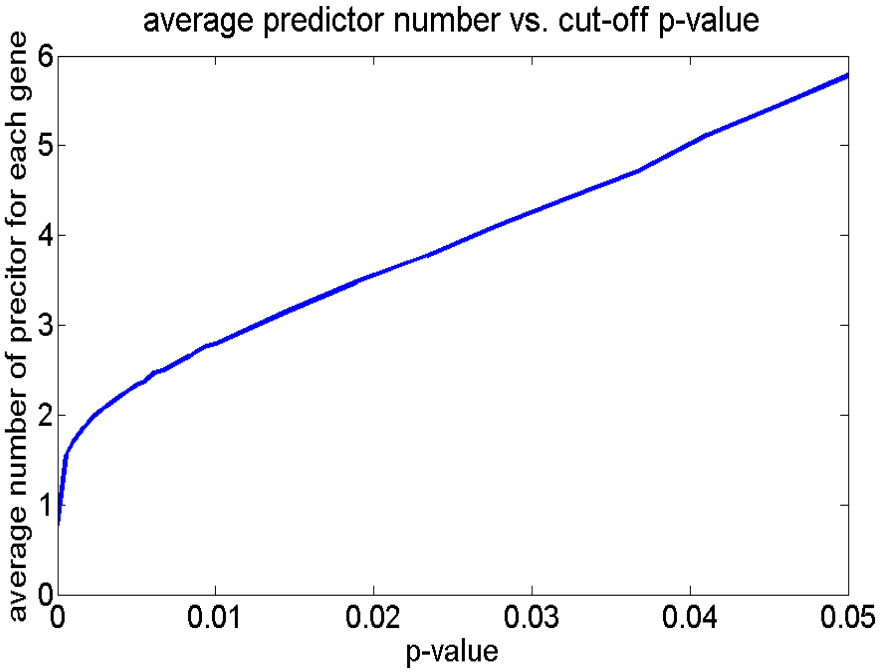
Prediction & Simulation

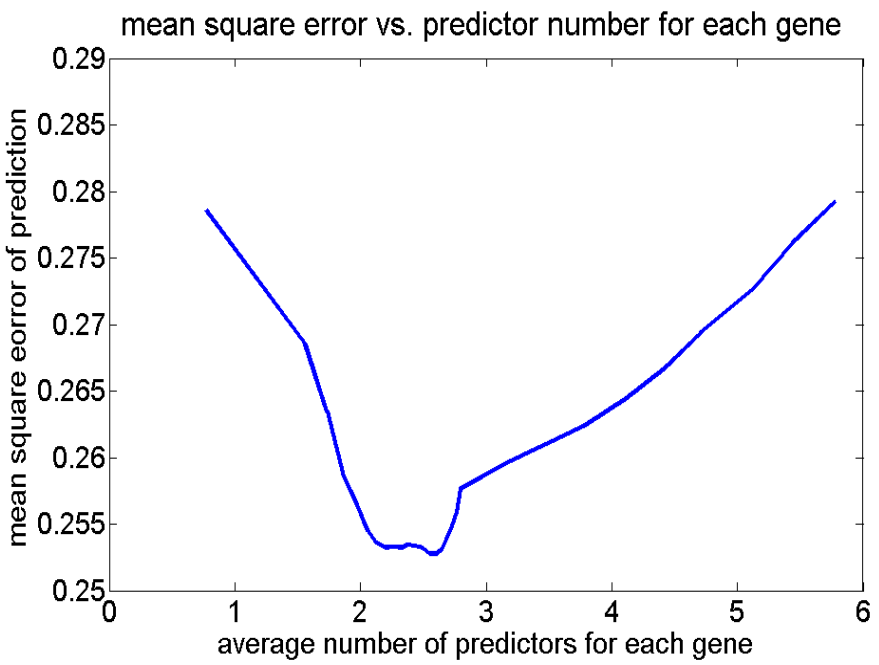
- Prediction
  - Predict the expression values in the next time step:

$$g_t^A = w^0 + \sum_{i=1..N} w^i g_{t-1}^i$$
$$G_t = M \times G_{t-1}$$

- Simulation
  - Given the first (few) time point(s), incrementally predict the rest of the time points.

p-value threshold





- 2~3 predictors per gene on average

(Thieffry D et al. 1998, M. Andrecut SH, and S.A. Kauffman, 2008)

Extended Model

- Consider multiple previous time points

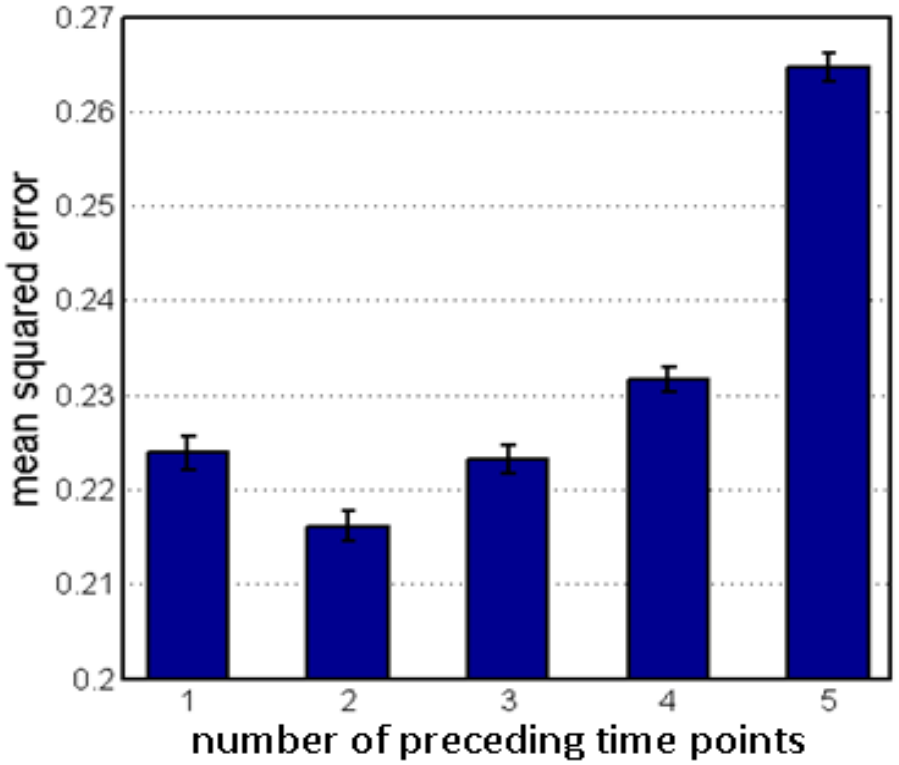
$$g_t^j = w^0 + \sum_{i=1..N} w^i g_{t-1}^i$$

$$\Downarrow$$

$$g_t^j = w^0 + \sum_{q=t-\tau \ldots t-1} \sum_{i=1..N} w_q^i g_q^i$$

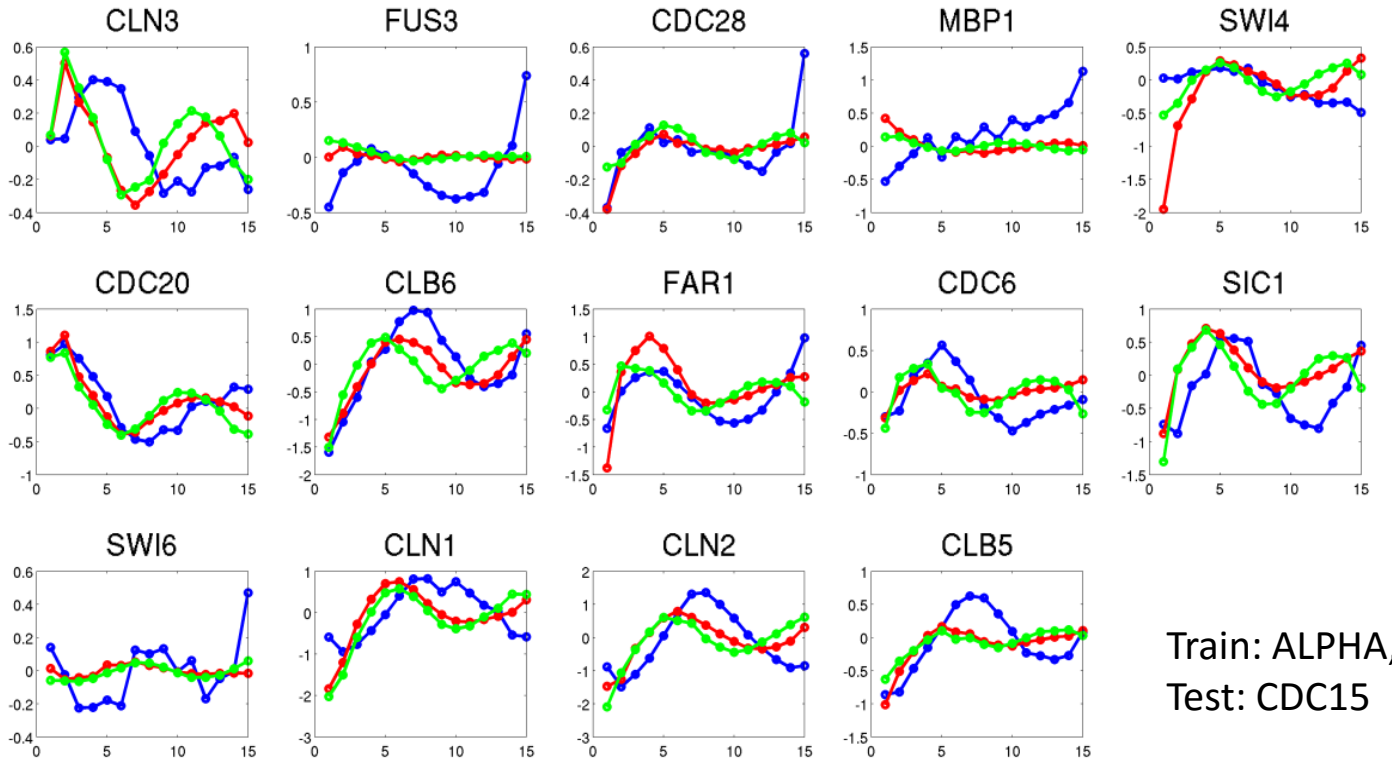
- Prediction:
  - $G_t = M_\tau \times [G_{t-\tau}; G_{t-\tau+1}; \ldots G_{t-2}; G_{t-1}]$

Number of previous time points (τ)



- 4-fold cross-validation, x1000 repeats
- τ=2 gives better MSE than others.

Simulation Results



Train: ALPHA, CDC28
Test: CDC15

blue: real

red: one previous time point

green: two previous time points

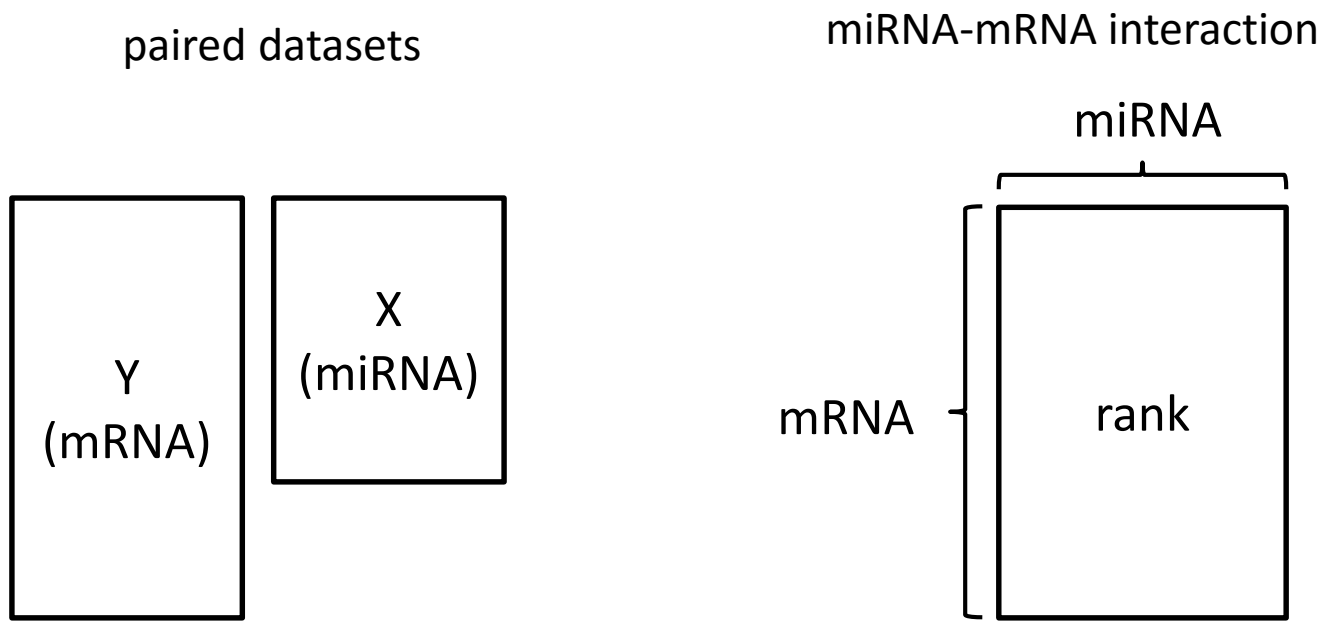
Summary of Aim I

- Time series data is modeled using a linear model. SMLR is used to fit data.
- The model solves the prediction, simulation, and network reconstruction problems.
- Our performance is better than other methods.

Aim two

paired miRNA-mRNA expression data

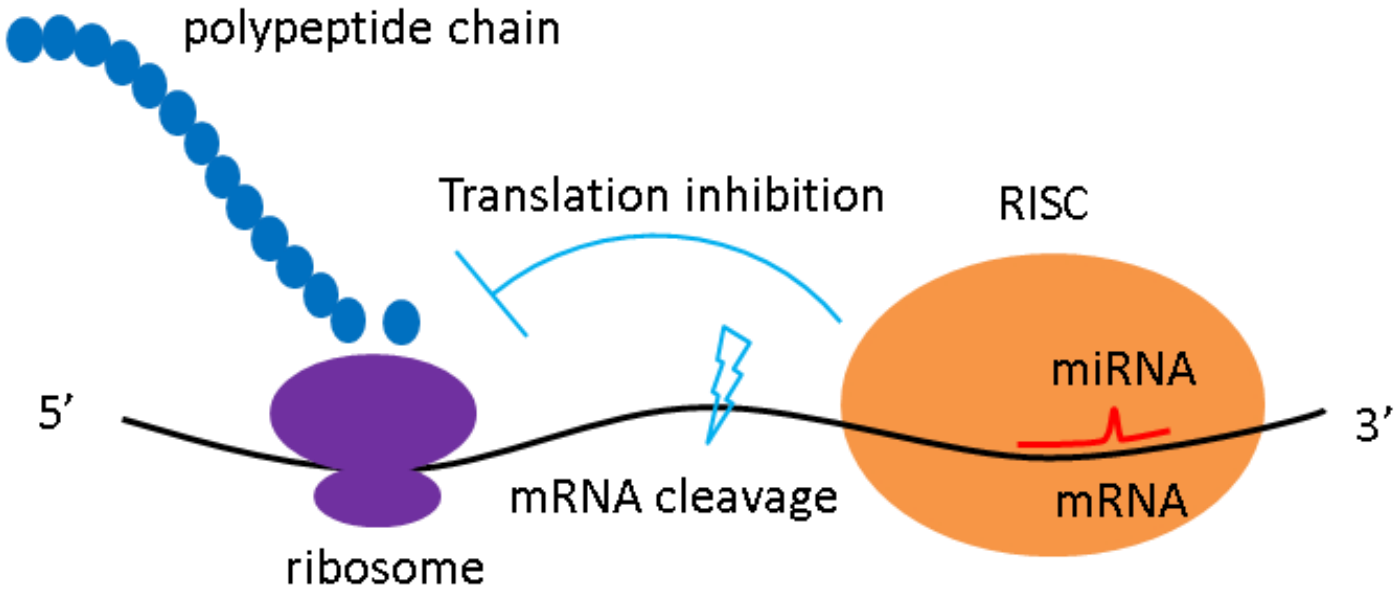
➔ miRNA-mRNA interaction



microRNA

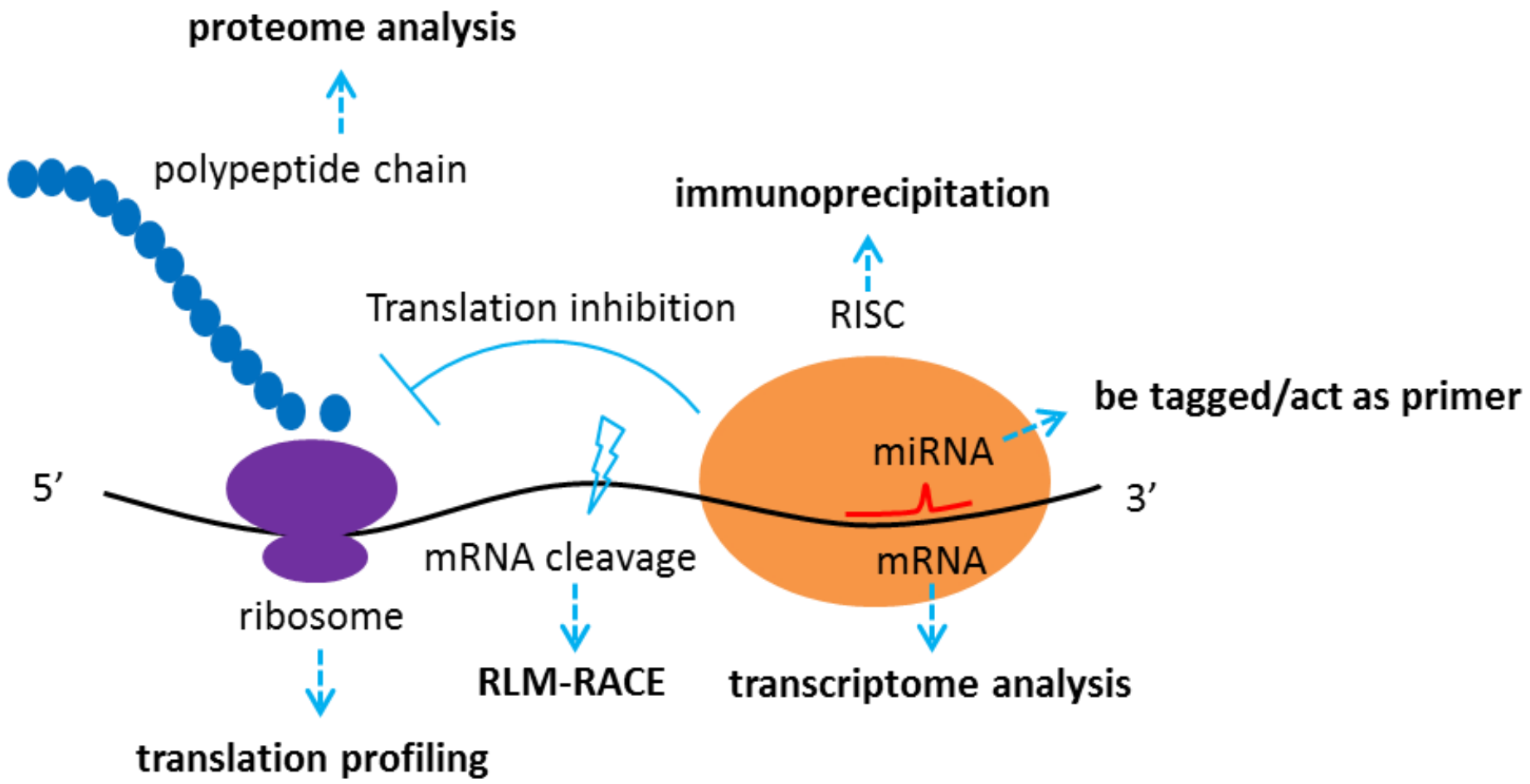
- small, ~22 nucleotide, non-coding endogenous RNA
- by pairing the messenger RNA, miRNAs repress the expression of their target gene, but not always
  (Vasudevan, Tong et al. 2007)
- participate in a wide range of biological process
  (He and Hannon 2004)
- affect over 60% of mammalian gene
  (Friedman, Farh et al. 2009),
- act as fine-tuners
  (Baek, Villen et al. 2008, Bartel 2009)
- contribute to tumor formation and progression
  (Croce 2009, Lujambio and Lowe 2012)

miRNA repress expression of target gene



(1) degrades mRNA molecules with Ago-family proteins
(2) decrease the translational efficiency

Identification of miRNA-target interaction



1872 human miRNA sequence annotated in mirBase (v20, June 2013)
(Kozomara and Griffiths-Jones 2014).

sequence-based algorithm

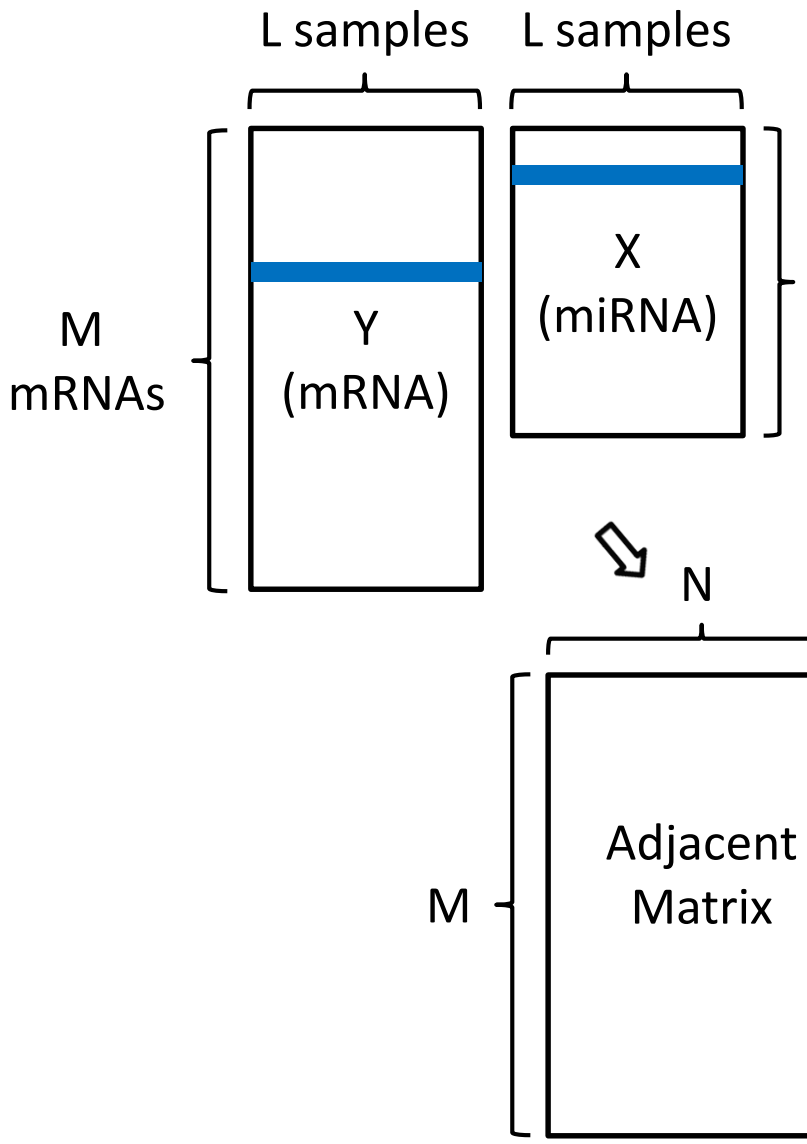
- Rules:
  - sequence complementarity
  - energetically favorable hybridization
  - evolutionary conservation
  - RNA secondary structure accessibility
  - multiple target sites
- Prediction methods and Databases
  - TargetScan, miRanda, PicTar, TargetScanS, PITA, DIANA-microT.
- Problem
  - false positive
  - false negative
- One solution:
  - Analysis of paired miRNA-mRNA expression datasets

Example datasets of paired miRNA-mRNA expression profiling for various cancer types

GEO ID	miRNA platform	mRNA platform	Number of samples	Sample type
GSE22220	GPL8178	GPL6098	426	breast cancer
GSE19536	GPL8227	GPL6480	215	breast cancer
GSE35602	GPL8227	GPL6480	59	colorectal cancer
GSE32688	GPL7723	GPL570, GPL6801	96	pancreatic cancer
GSE40355	GPL8227	GPL13497	48	bladder cancer
GSE19783	GPL8227	GPL6480	216	breast cancer
GSE28544	GPL10850	GPL6244	56	breast cancer
GSE35982	GPL14767	GPL4133	32	colorectal cancer
GSE20161	GPL8178	GPL6102	215	prostate cancer
GSE21032	GPL8227	GPL4091, GPL5188, GPL10264	743	prostate cancer
GSE25692	GPL9081	GPL7363	43	prostate cancer
…	…	…	…	…

And the TCGA project

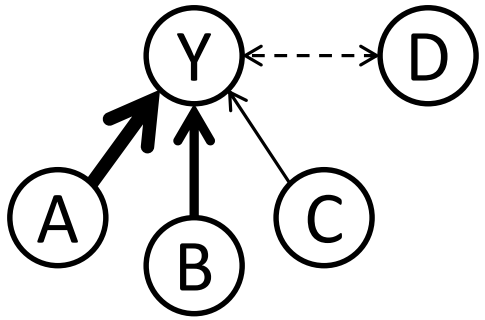
Expression-based methods



pairwise simple correlation
mutual information

miRGator
(Nam, Kim et al. 2008)
MMIA
(Nam, Li et al. 2009)
mimiRNA
(Ritchie, Flamant et al. 2010)
MAGIA
(Sales, Coppe et al. 2010)
ExprTarget
(Gamazon, Im et al. 2010)
mirConnX
(Huang, Athanassiou et al. 2011)

Problem with Correlation-based approach
Example:

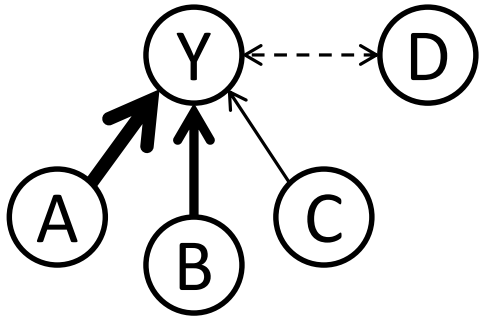


$$Y = 20*A + 10*B - C$$

- simple correlation methods may fail to give Y-C proper rank
- shadowed by stronger co-regulators A and B

Forward-correlation (forwardCorr)

Hybrid correlation and multiple linear model



step 1:
Calculate correlation: Y-A, Y-B, Y-C, Y-D
Select A

Step 2:
Remove effect of A, Y’ = Y – effect(A), B’ = … C’ = … D’ = …
Calculate correlation: Y’-B’, Y’-C’, Y’-D’
Select B
…

$\boldsymbol{X}$: be the training matrix of selected potential predictors

$\boldsymbol{X}^{out}$ be the training matrix of remaining potential predictors

QR factorization

$$\boldsymbol{X} = \boldsymbol{QR} = [\boldsymbol{Q}_1 \quad \boldsymbol{Q}_2]\begin{bmatrix}\boldsymbol{R}_1 \\ \boldsymbol{0}\end{bmatrix} = \boldsymbol{Q}_1\boldsymbol{R}_1$$

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = (\boldsymbol{R}_1^T\boldsymbol{R}_1)^{-1}\boldsymbol{R}_1^T\boldsymbol{Q}_1^T\boldsymbol{y}$$

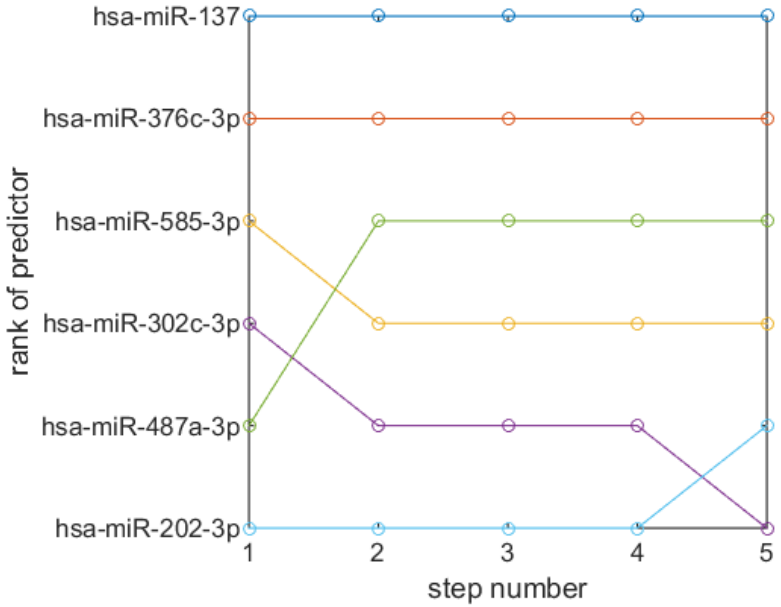
$$= \boldsymbol{R}_1^{-1}(\boldsymbol{R}_1^T)^{-1}\boldsymbol{R}_1^T\boldsymbol{Q}_1^T\boldsymbol{y} = \boldsymbol{R}_1^{-1}\boldsymbol{Q}_1^T\boldsymbol{y}$$

$$\boldsymbol{y}_r = \boldsymbol{y} - \widehat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{Q}_1\boldsymbol{Q}_1^T\boldsymbol{y}$$

$$\boldsymbol{X}_r^{out} = \boldsymbol{X}^{out} - \boldsymbol{Q}_1\boldsymbol{Q}_1^T\boldsymbol{X}^{out}$$

Example of forwardCorr algorithm

Predictor	Step 1	Step 2	Step 3	Step 4	Step 5
hsa-miR-137	**0.02**	0.03	0.16	0.21	0.35
hsa-miR-376c-3p	**0.39**	0.73	0.89	0.88	0.66
hsa-miR-585-3p	0.45	0.58	**0.44**	0.68	0.8
hsa-miR-302c-3p	**0.59**	0.99	0.83	0.9	0.76
hsa-miR-487a-3p	0.59	**0.39**	0.51	0.67	0.89
hsa-miR-202-3p	0.77	0.92	0.8	0.95	**0.58**



Learn each dataset with ForwardCorr and combine the results

GEO ID	miRNA platform	mRNA platform	Number of samples	Sample type
GSE22220	GPL8178	GPL6098	426	breast cancer
GSE19536	GPL8227	GPL6480	215	breast cancer
GSE35602	GPL8227	GPL6480	59	colorectal cancer
GSE32688	GPL7723	GPL570, GPL6801	96	pancreatic cancer
GSE40355	GPL8227	GPL13497	48	bladder cancer
GSE19783	GPL8227	GPL6480	216	breast cancer
GSE28544	GPL10850	GPL6244	56	breast cancer
GSE35982	GPL14767	GPL4133	32	colorectal cancer
GSE20161	GPL8178	GPL6102	215	prostate cancer
GSE21032	GPL8227	GPL4091, GPL5188, GPL10264	743	prostate cancer
GSE25692	GPL9081	GPL7363	43	prostate cancer
…	…	…	…	…

And the TCGA project

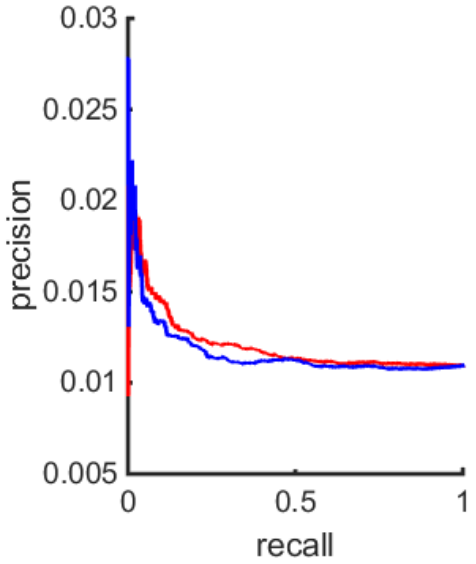
forwardCorr vs. simple correlation

Combined results

(16 paired miRNA-mRNA datasets, rank-product)

MirTarBase as true positive

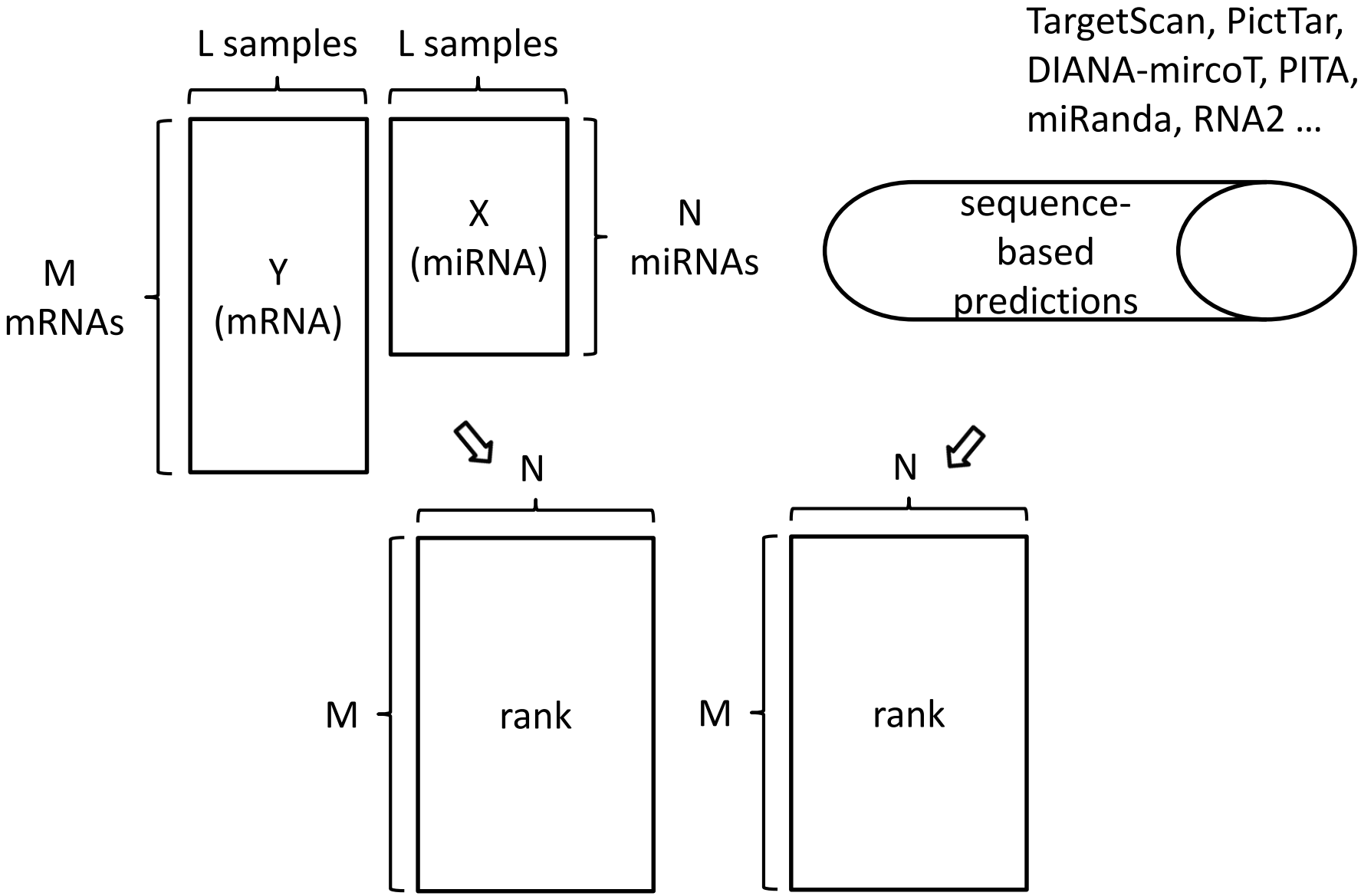
All pairs



Area under curve

Precision-recall	All pairs
Negative correlation	0.0117
Forward selection	0.0121
Improvement	+3.5%

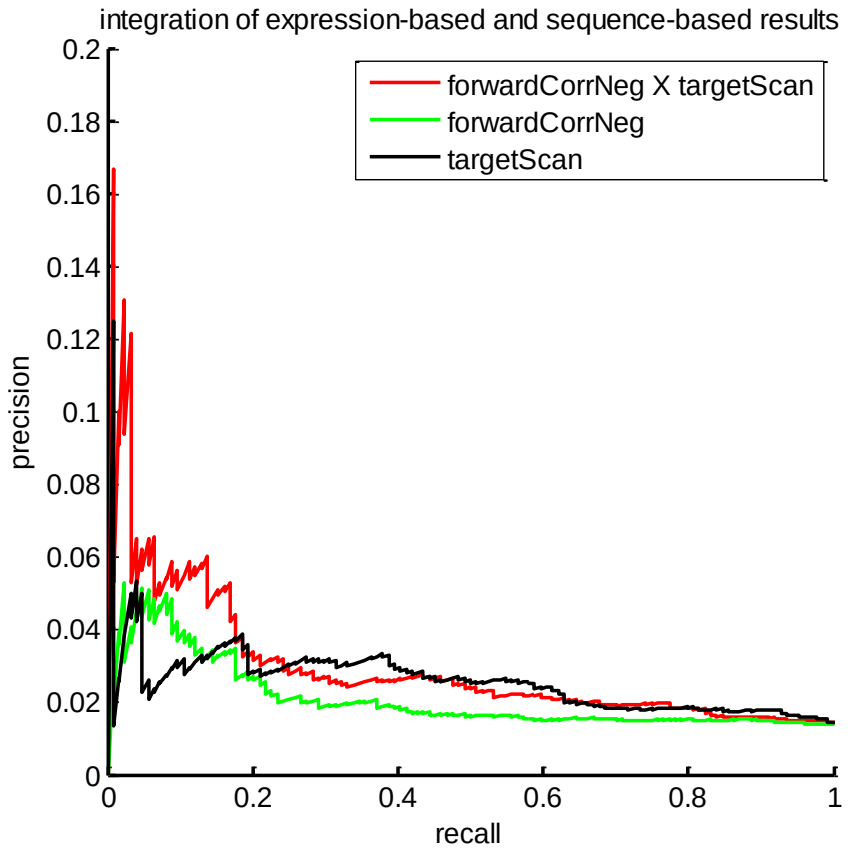
Combine expression-based prediction and sequence-based prediction



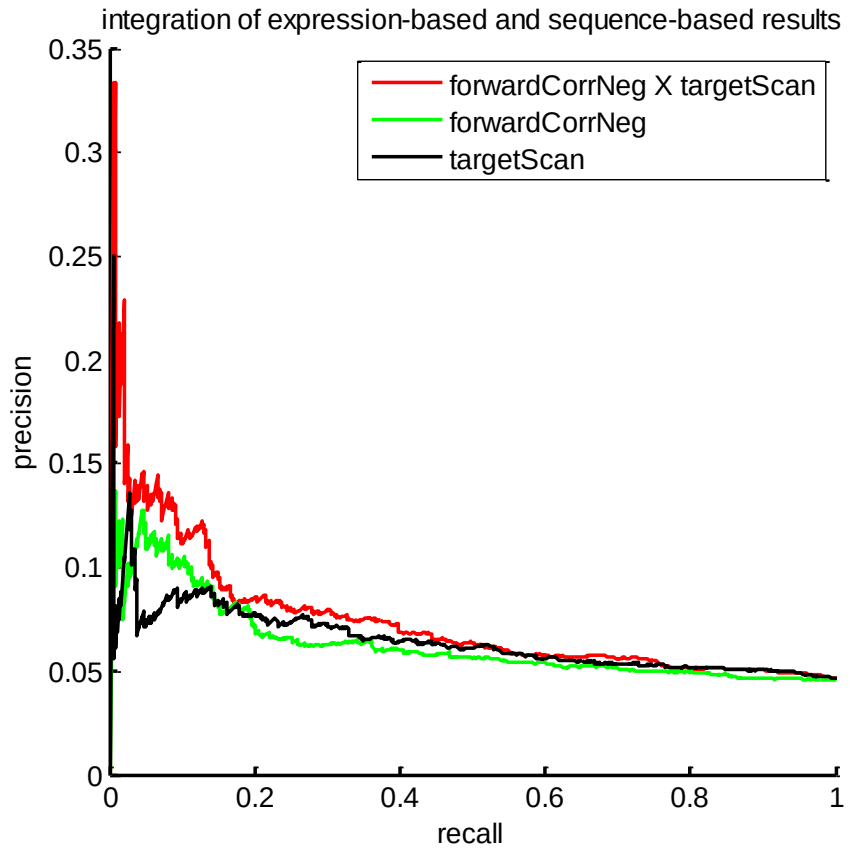
Combine sequence-based and expression-based methods

rank-product

Strong evidence



all



* Strong evidence in MirTarBase: Reporter assay or Western blot

Area under curve for integrative analysis of miRNA-target interaction

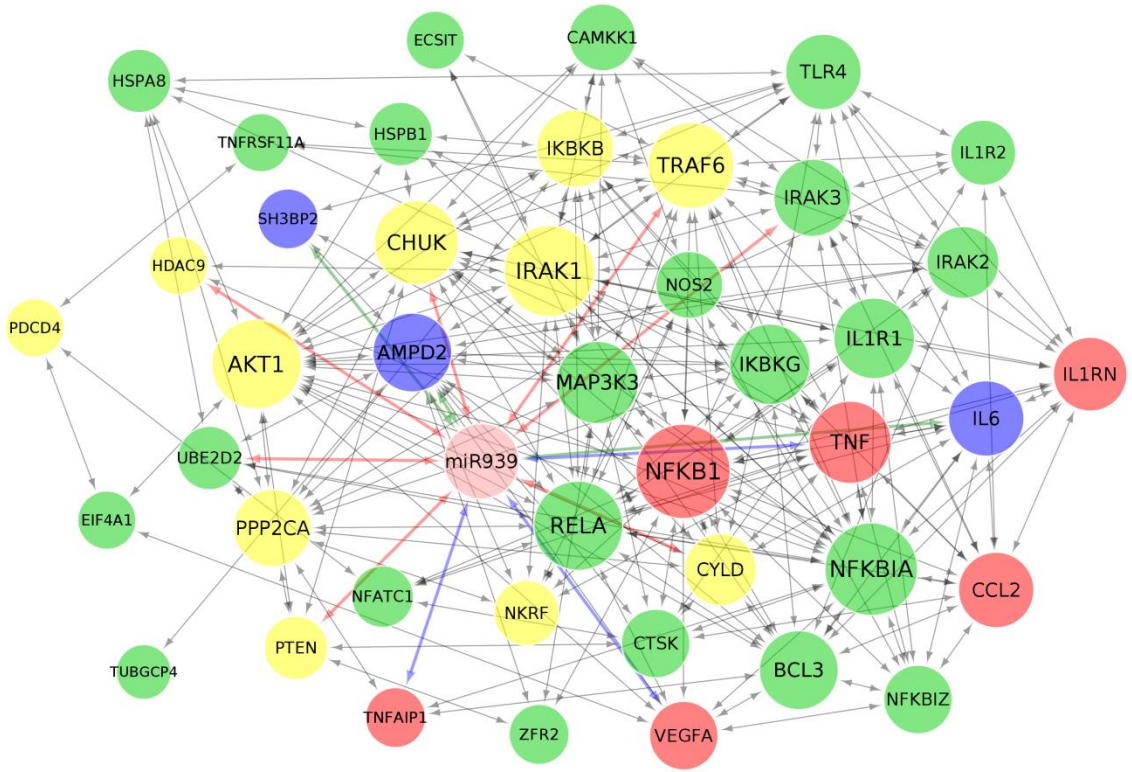
Area under curve	MTB strong evidence	MTB
TargetScan	0.0253	0.0641
ForwardCorrNeg	0.0206	0.0628
ForwardCorrNeg X TargetScan	0.0291	0.0738
Improvement over Targetscan	15%	15%

Case Study: hsa-mir-939

- miR-939 was previously found to be significantly altered in samples from patients with complex regional pain syndrome (CRPS) versus control samples

Orlova et al., 2011

hsa-mir-939 and inflammatory network



178 Predicted by targetScan
1002 Predicted by expression

geneMANIA web server
(Warde-Farley et al., 2010)

NFκB, play a central role in inflammation
Yellow: known neighbor of NFκB
Green: from geneMANIA
Blue: from MirTarBase
Red: experimentally validated proinflammatory genes

Summary of Aim II

- A hybrid method is proposed for inference of miRNA-mRNA interaction
- Better than simple correlation
- Combining sequence-based and expression-based prediction methods improves inference performance

Gene expression prediction

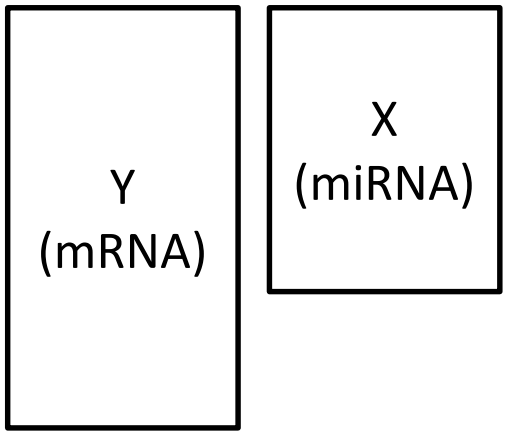
- SMLR model
  - mRNA expression prediction and Time series simulation
  - miRNA expression → mRNA expression ???

Aim three

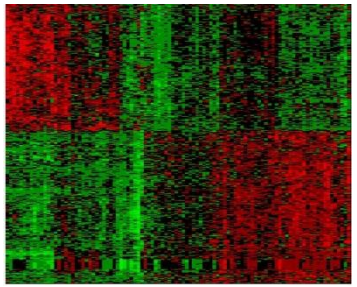
Paired miRNA-mRNA expression data

➔ miRNA functional annotation

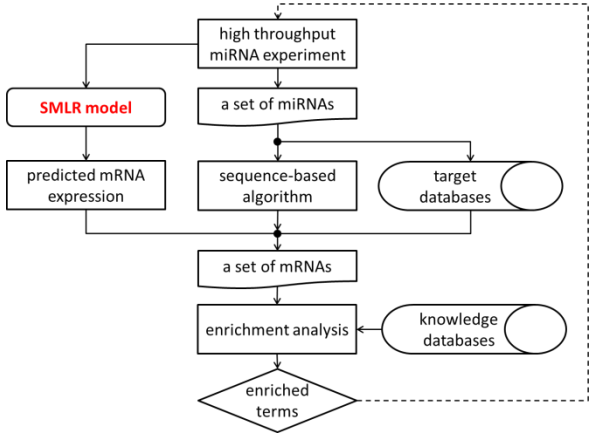
paired datasets



mRNA expression prediction



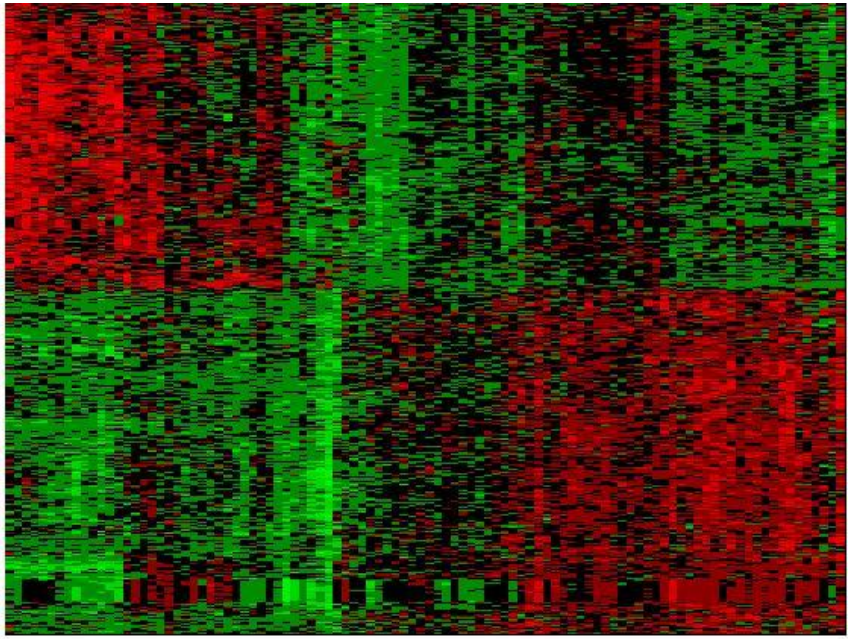
miRNA functional annotation



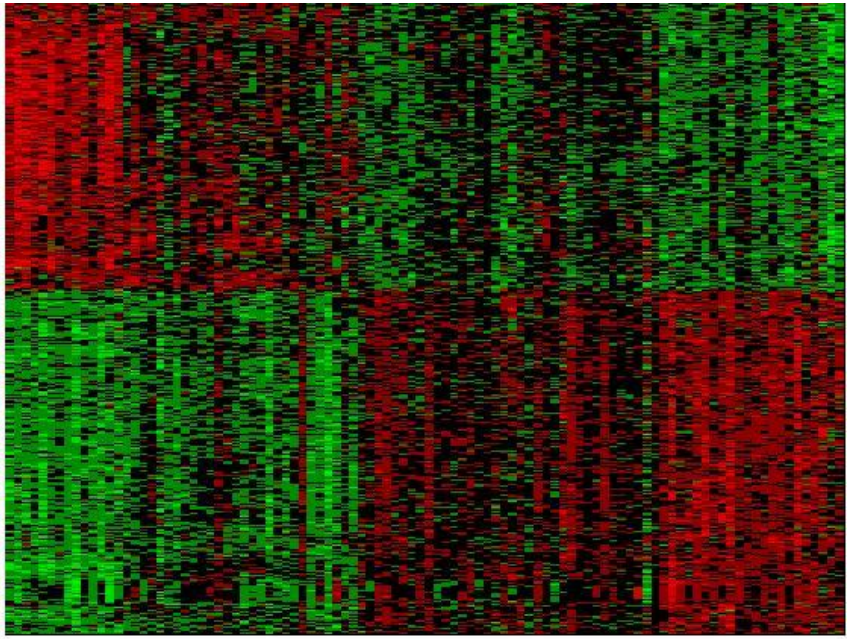
Prediction within one dataset

leave one out cross validation

true mRNA expression of GSE19536 datasets



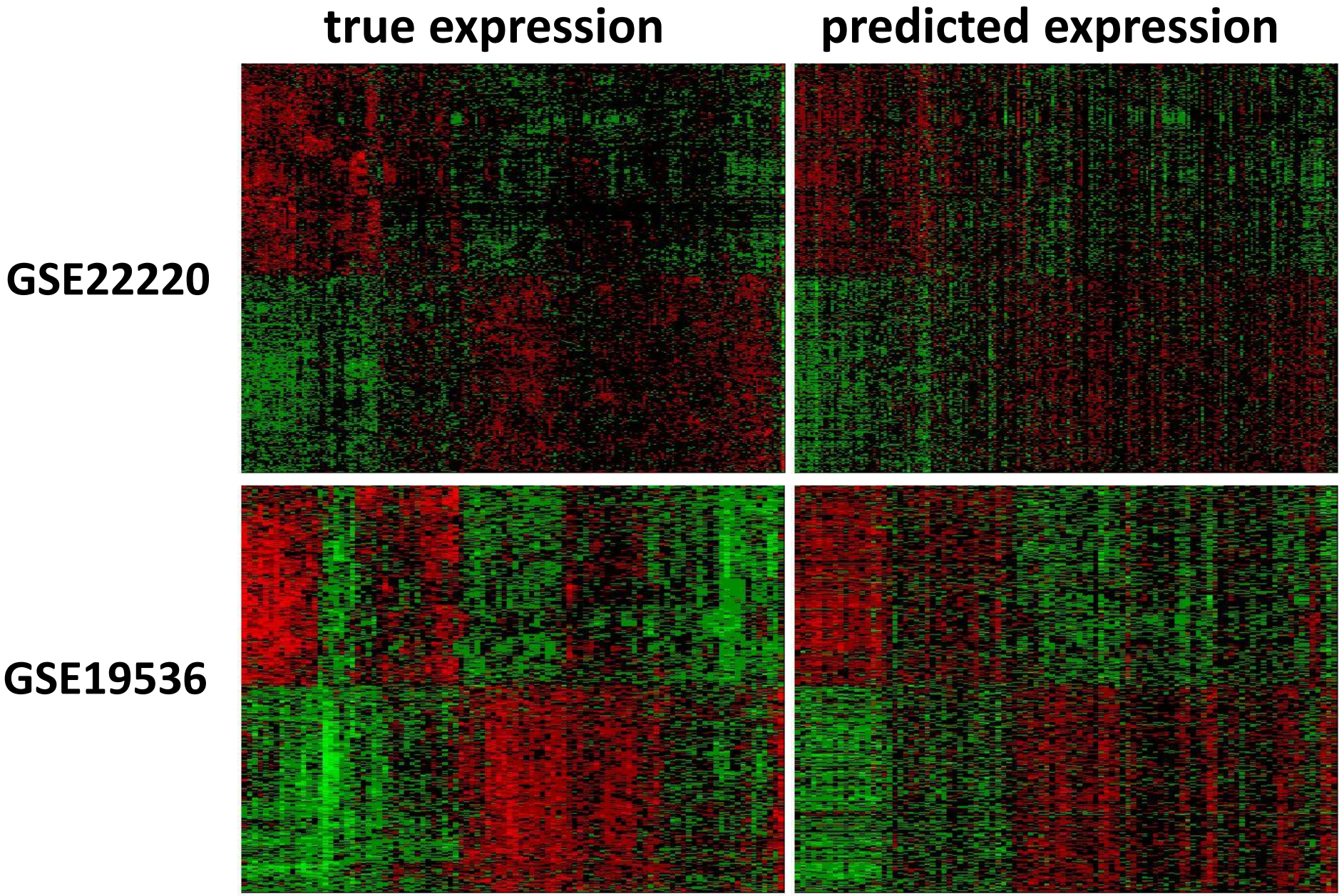
predicted expression by leave-one-out-cross-validation strategy



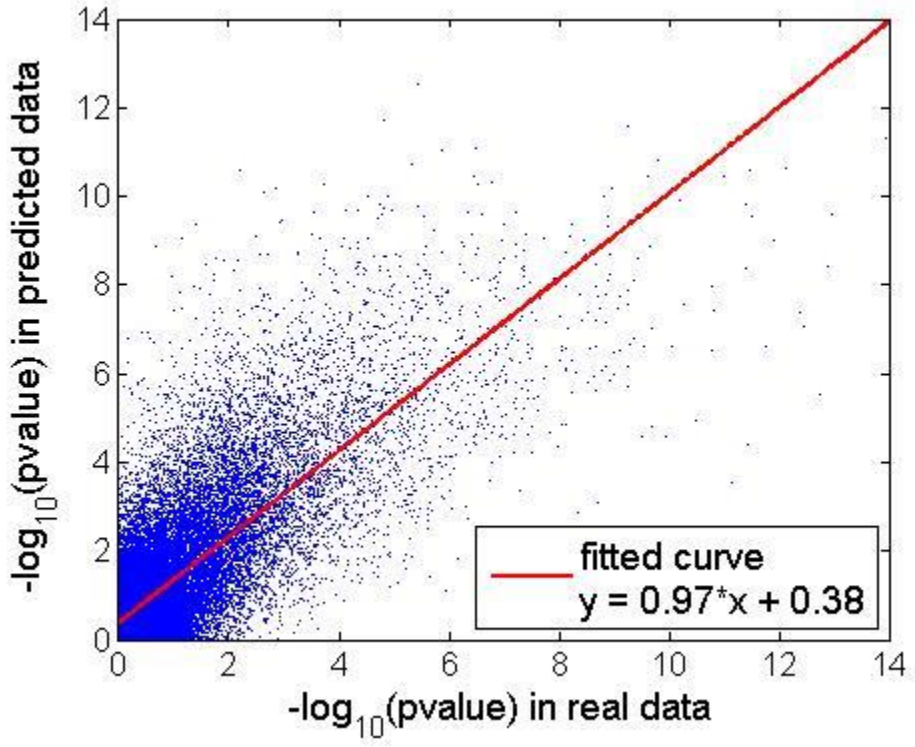
Two breast cancer datasets

	GSE22220	GSE19536
reference	(Buffa, Camps et al. 2011)	(Enerly, Steinfeld et al. 2011)
sample number	207	101
cohort	Oxford	Oslo region
miRNA platform	Illumina Human v1 MicroRNA expression beadchip (GPL8178)	Agilent-019118 Human miRNA Microarray 2.0 G4470B (miRNA ID version) (GPL8227)
mRNA platform	Illumina humanRef-8 v1.0 expression beadchip (GPL6098)	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F (Probe Name version) (GPL6480)
miRNA	735	489
mRNA	24332	40989
common miRNA	248	248
common mRNA	14873	14873

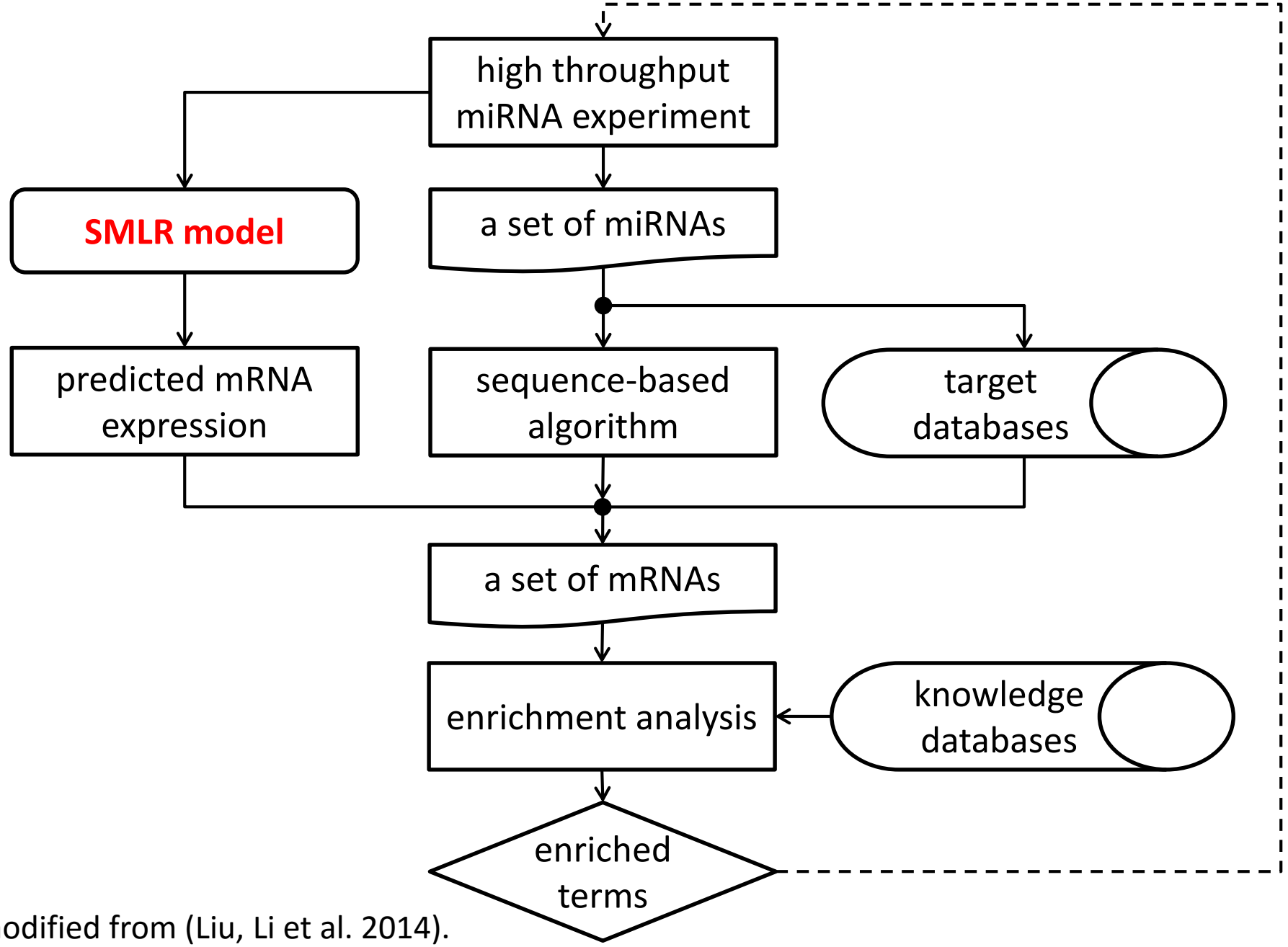
Prediction across two datasets



Comparison of differentially expressed mRNAs identified from the real and predicted expression data

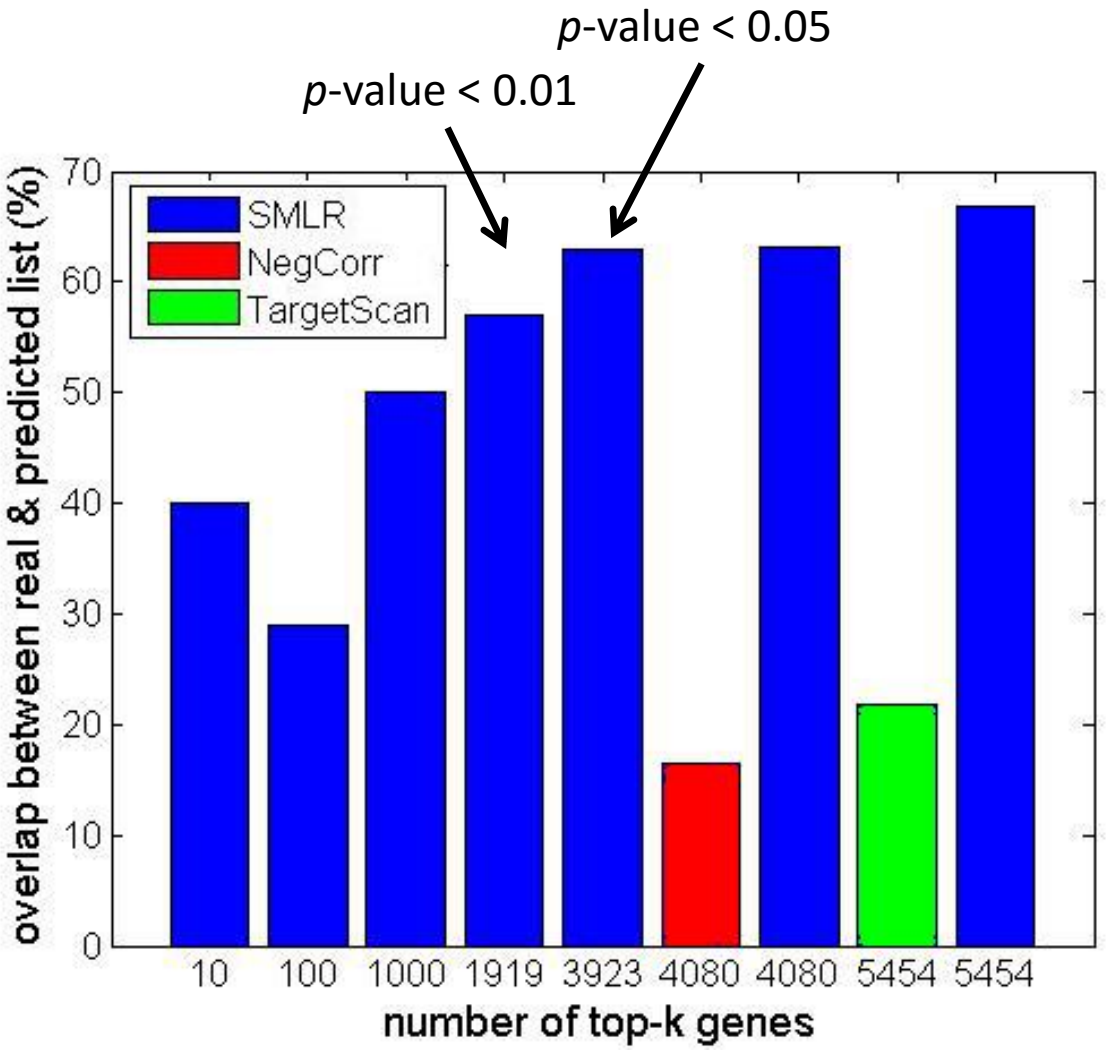


A framework of miRNA functional annotation.

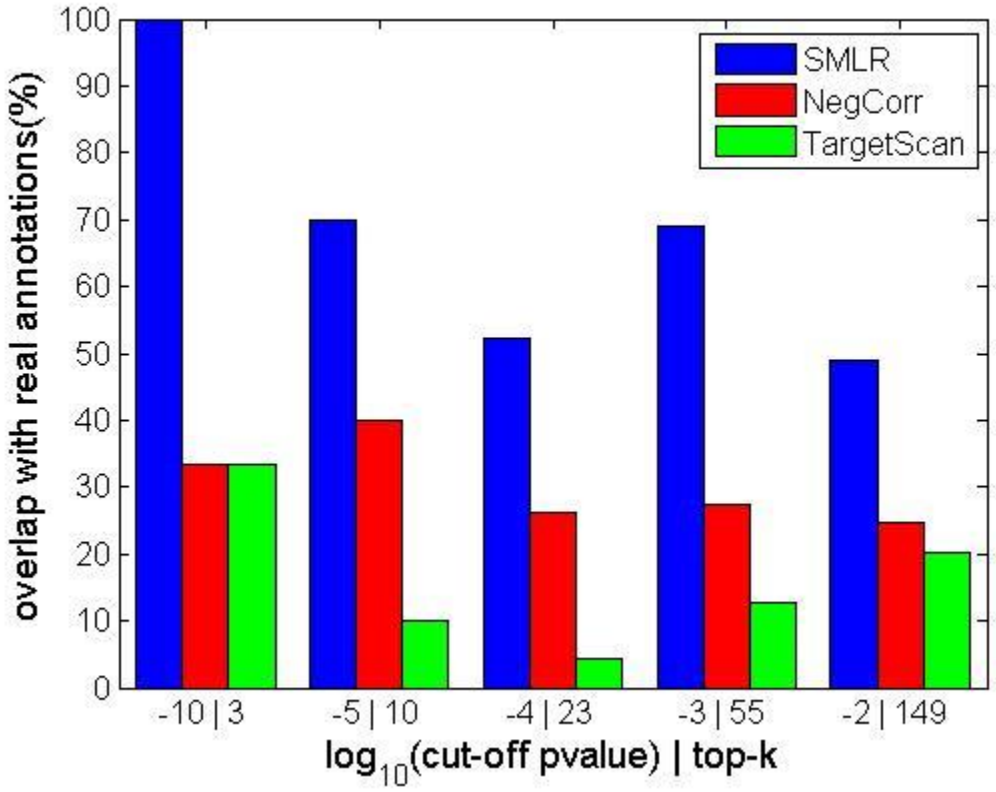


modified from (Liu, Li et al. 2014).

Amount of overlap between the lists of differentially expressed genes in real and predicted data



Enrichment of functional categories (GO & KEGG)



Summary of Aim III

- SMLR model is able to prediction mRNA expression from miRNA expression data.
- A novel strategy is proposed for functional annotation of miRNA using predicted gene expression.

Summary

- I. Time series gene expression data
  - Reconstruction of gene regulatory network
  - Simulation of time-series by SMLR
- II. Paired miRNA-mRNA expression data
  - miRNA-target inference based on expression
  - Integration of sequence-based and expression-based target prediction methods
- III. Functional annotation of miRNA
  - mRNA expression prediction from miRNA profile
  - Improved functional annotation by predicted profiles

Future work

- I. Time series gene expression data
  - Integrating data from different time periods
  - Comparing models from normal tissue simulation to detect abnormalities from disease
- II. Paired miRNA-mRNA expression data
  - Increase number of datasets utilized as more studies are published
  - Develop tissue specific models
  - Integration with other prediction methods (literature mining, multiple sequence based methods, etc)
- III. Functional annotation of miRNA
  - Build tissue specific SMLR models
  - A software or web service to take miRNAs and give gene expression values and enriched pathways

Publications

- Yiqian Zhou, Rehman Qureshi, Francis Bell, and Ahmet Sacan, Reconstruction of Regulatory Networks from Microarray Data. Microarray Image and Data Analysis. Mar 2014 , 401-429
- Yiqian Zhou, Rehman Qureshi, and Ahmet Sacan, Data Simulation and Regulatory Network Reconstruction from Time-series Microarray Data Using Stepwise Multiple Linear Regression. Network Modeling Analysis in Health Informatics and Bioinformatics. 1(1): 3-17, 2012.
- Yiqian Zhou, Rehman Qureshi, and Ahmet Sacan, Reconstruction Of Gene Regulatory Networks By Stepwise Multiple Linear Regression From Time-Series Microarray Data. *IEEE Transactions on International Symposium on Health Informatics and Bioinformatics (HIBIT)*. 2012.
- Yiqian Zhou, Jacqueline Gerhart, and Ahmet Sacan, Gene Regulatory Networks Reconstruction by Multiple Linear Regressions from Time-series Microarray Data.
  IEEE International Conference Bioinformatics & Biomedicine, 2011. [*Best poster award.*]

- Yiqian Zhou, Rehman Qureshi, and Ahmet Sacan, Analysis of paired miRNA-mRNA microarray expression data using a stepwise multiple linear regression model (in preparation)

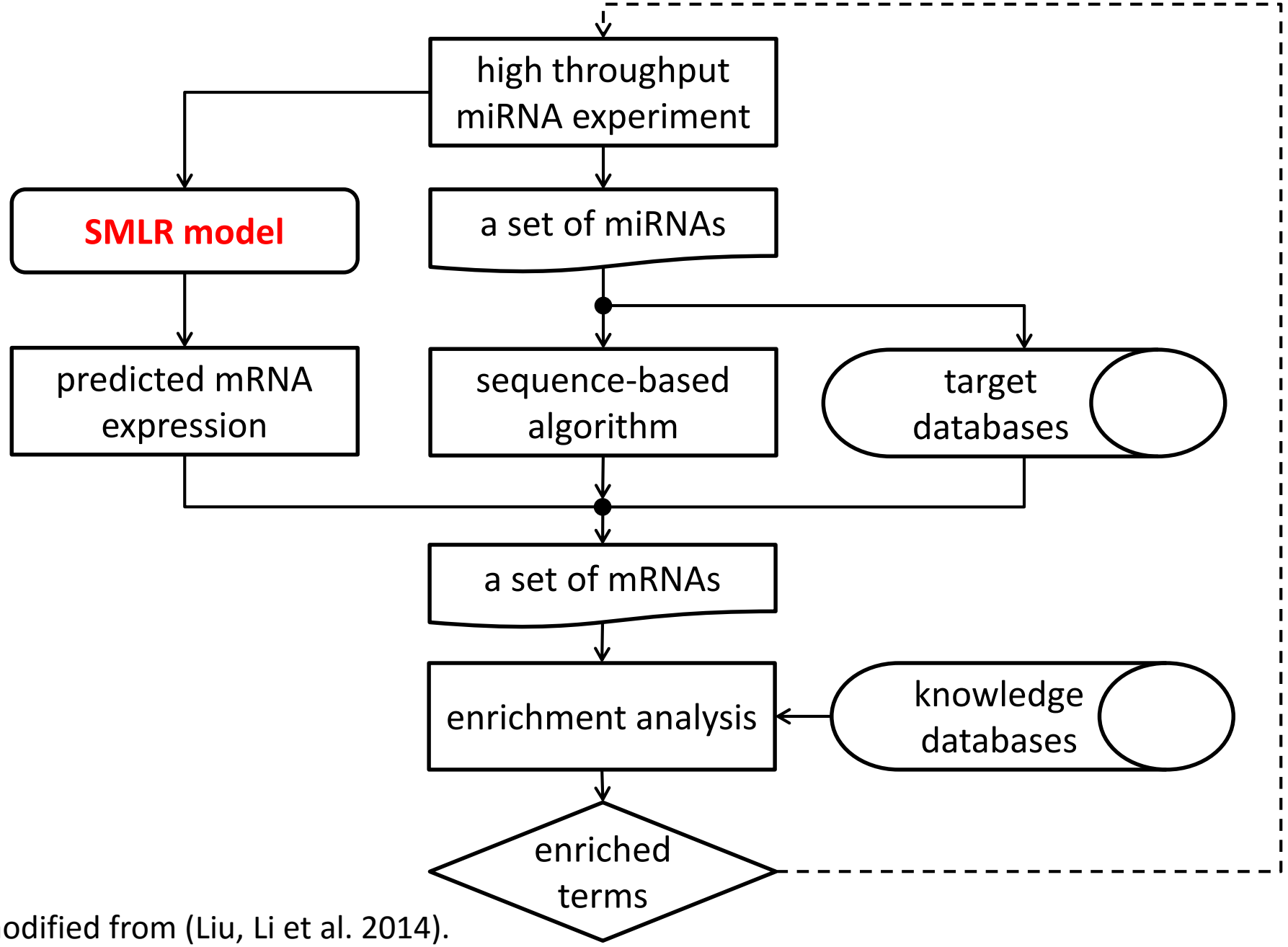
Acknowledgements

- **Faculty Advisor**
  - Ahmet Sacan, Ph.D.
- **Committee Members**
  - Seena Ajit, Ph. D.
  - Lin Han, Ph.D.
  - Uri Hershberg, Ph.D.
  - Andres Kriete, Ph.D.
- **Group Members**
  - Francis Bell
  - Rehman Qureshi
  - Daisy Yang
  - Chunyu Zhao

Questions?

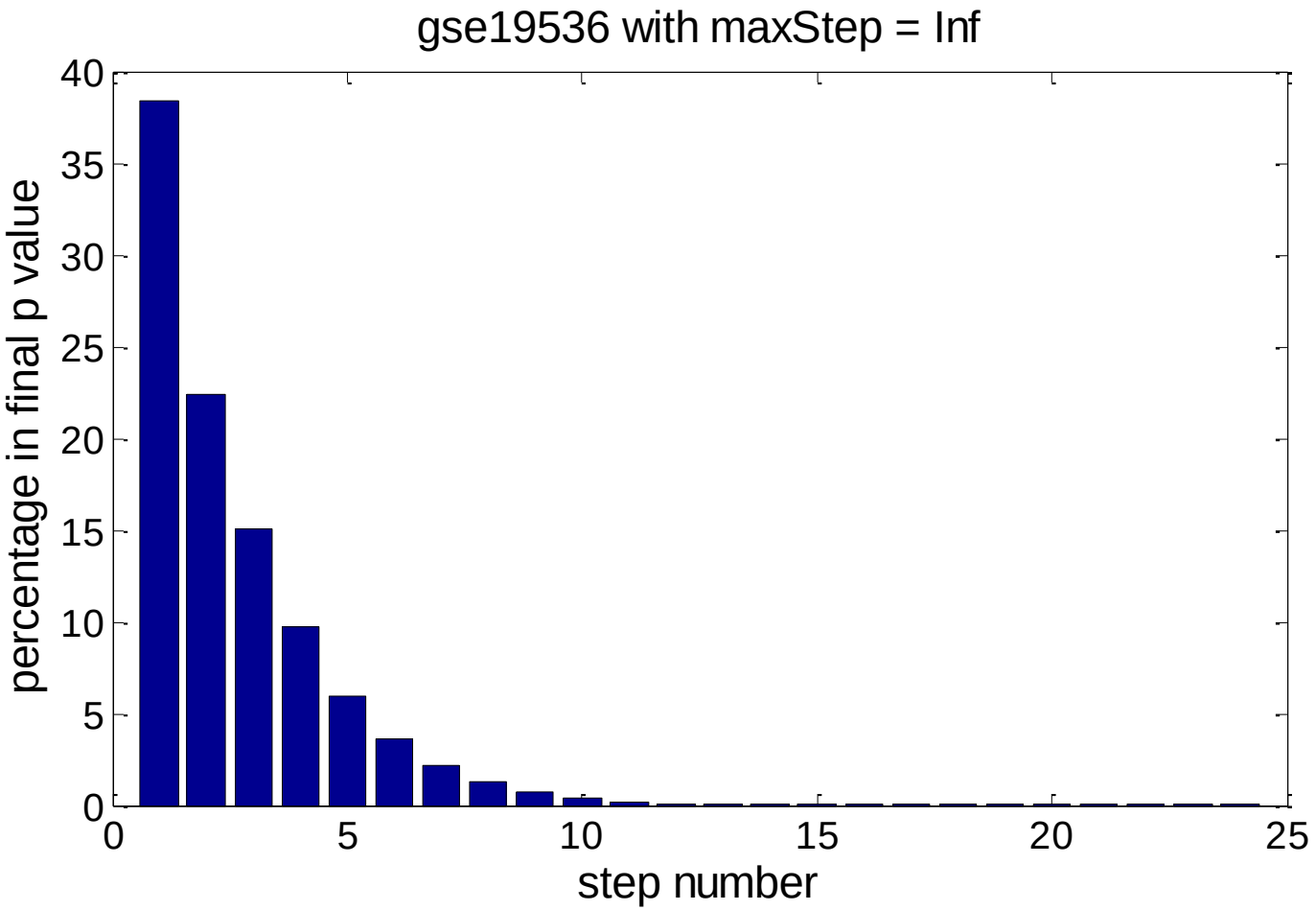
extra slides

A framework of miRNA functional annotation.



modified from (Liu, Li et al. 2014).

Area under curve	MTB strong evidence	MTB
targetScan	0.0253	0.0641
forwardCorrNeg	0.0206	0.0628
forwardCorrNeg * targetScan	**0.0291**	**0.0738**
corrNeg * targetScan	0.0283	0.0724


gse19536 with maxStep = Inf
percentage in final p value
step number
0
5
10
15
20
25
30
35
40

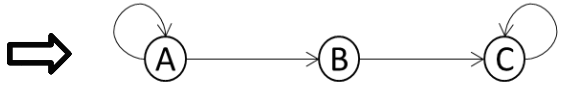
Three specific aims:

samples

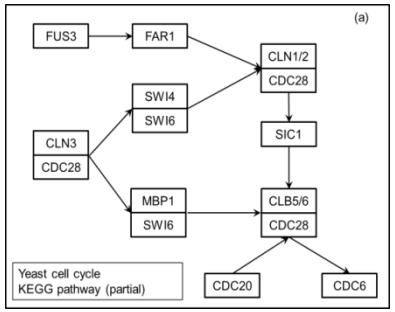
genes

Data Matrix

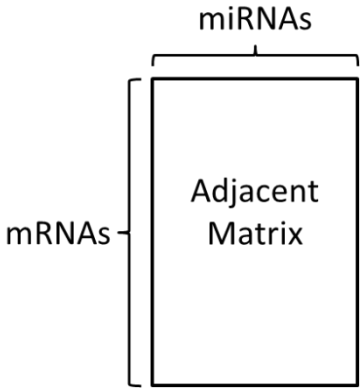
tall



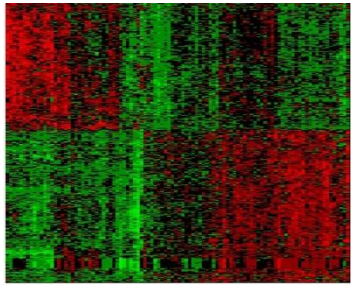
Gene regulatory network



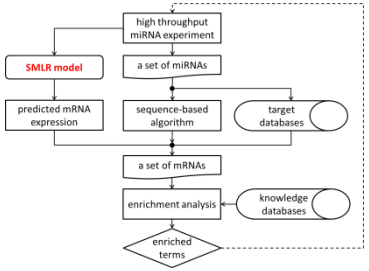
miRNA-mRNA interaction



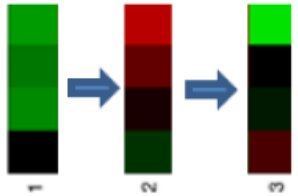
miRNA➔ mRNA



miRNA functional annotation



simulation



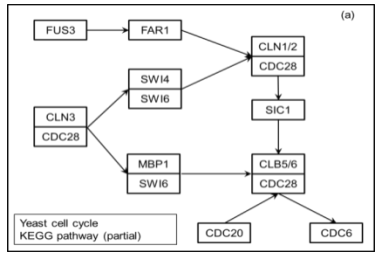
Three specific aims:

1. Time-series gene expression data ➔ Gene regulatory network & simulation

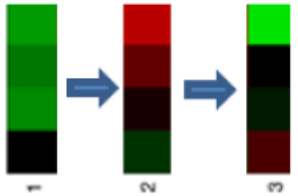
2. Paired miRNA-mRNA expression data ➔ miRNA-mRNA interaction

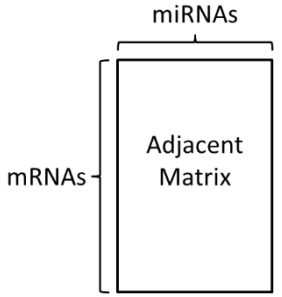
3. Paired miRNA-mRNA expression data ➔ miRNA functional annotation

Gene regulatory network



Simulation

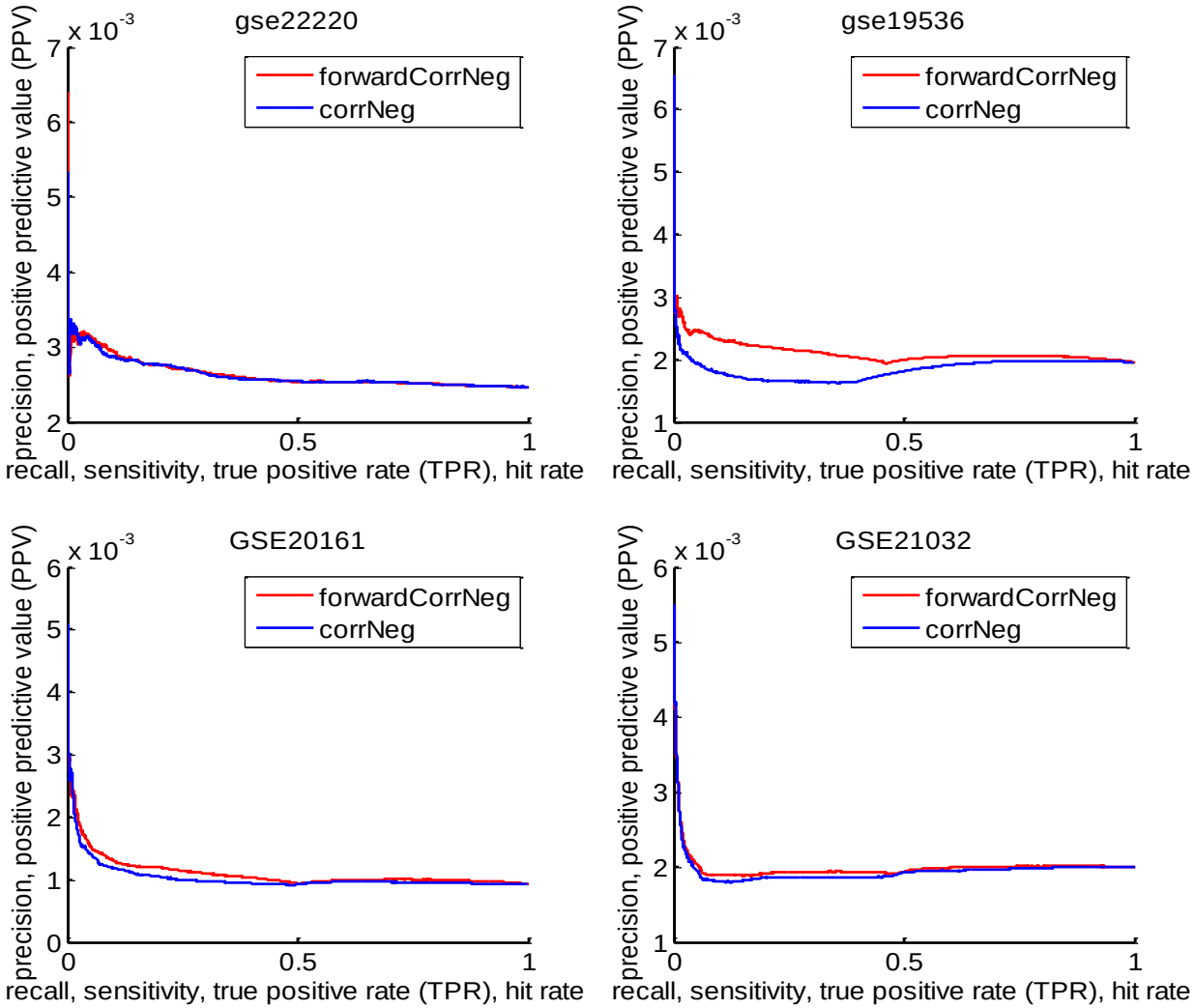




miRNA-mRNA interaction

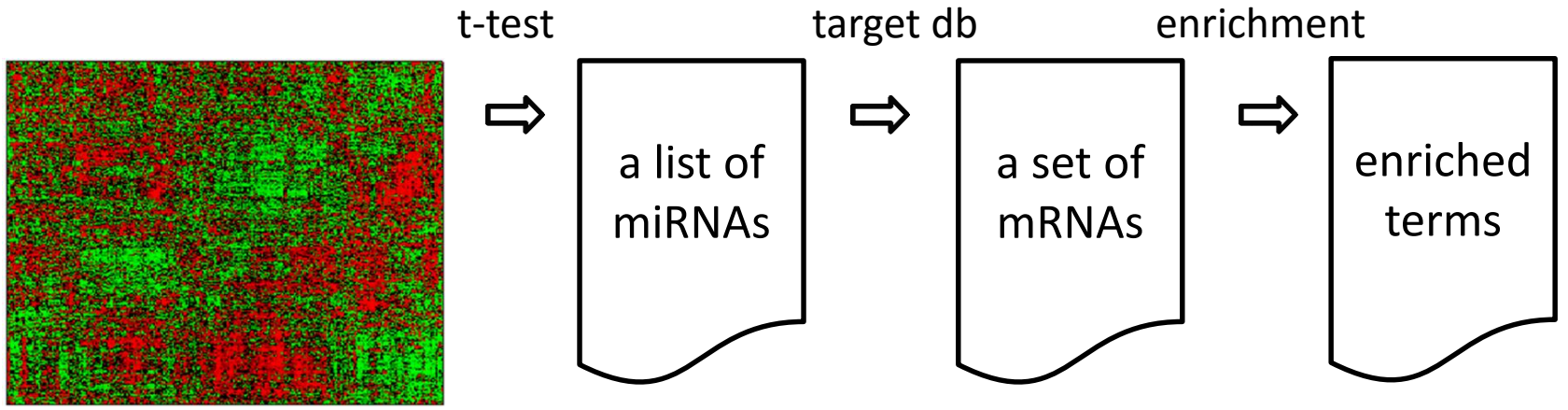
Precision-recall.

Comparison between ForwardCorr and simple correlation



True Positive: MirTarBase

Annotation of miRNA function



(Liu, Li et al. 2014).

Annotation of miRNA function