Into the Darkness of Gene Regulation: Understanding the Regulation and the Function of Non-Coding RNAs from ENCODE Data

Renqiang Min
Chao Cheng
Yale University
July 26, 2011

Why Study Non-Coding RNAs

- Protein-coding genes only account for about 1.5% of the human genome (about 3 billion base pairs), and the rest is dark matter, among which there is ncRNA
- The expressions of coding genes are regulated not only by proteins (TFs), but also by some ncRNAs
- ncRNAs have been found to be associated with Cancer, Autism, and Alzheimer's disease
- ncRNAs are candidate bio-markers

Overview of ncRNAs in ENCODE Data

- We consider the following RNAs from Gencode v7 annotation as non-coding RNAs:

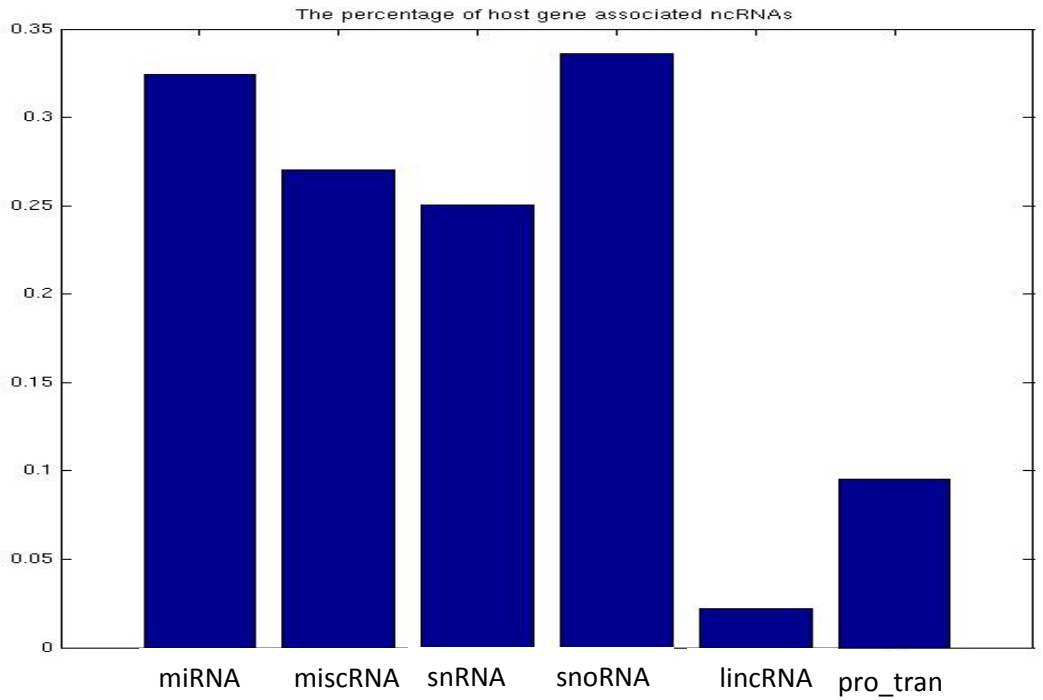
	number	avg. length (End Pos – Start Pos)
-microRNA:	1756	92
-misc-RNA:	1187	153
-snRNA:	1944	107
-snoRNA:	1521	110
-lincRNA:	1239	43970 (11828, median)
-processed transcript:	8401	26346 (6372, median)

-rRNA(excluded from regulation and functional analysis): 531

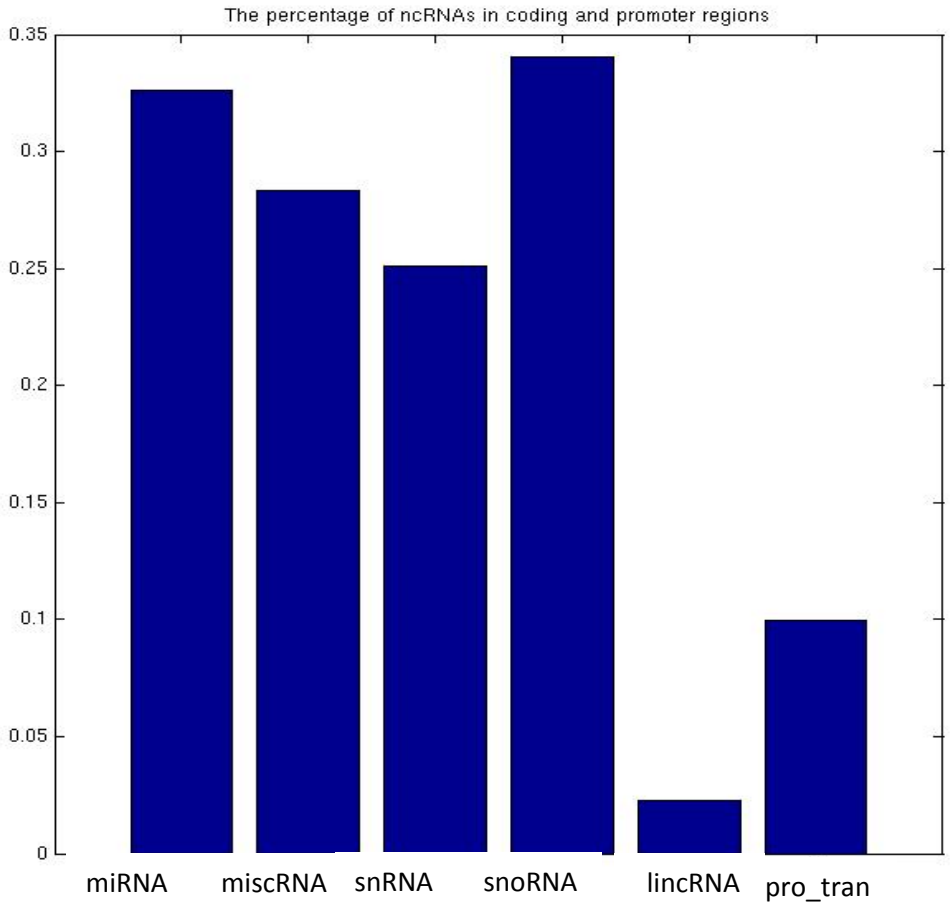
- Total Length (annotation): 276,588,160 bases
- Transcribed region: 11,884,948 bases

Host-Gene Associated ncRNAs

- The ncRNAs that lie in the protein coding regions



ncRNAs in Protein Coding and Promoter Regions



Histone Signal Peaks in Promoter Regions of ncRNAs

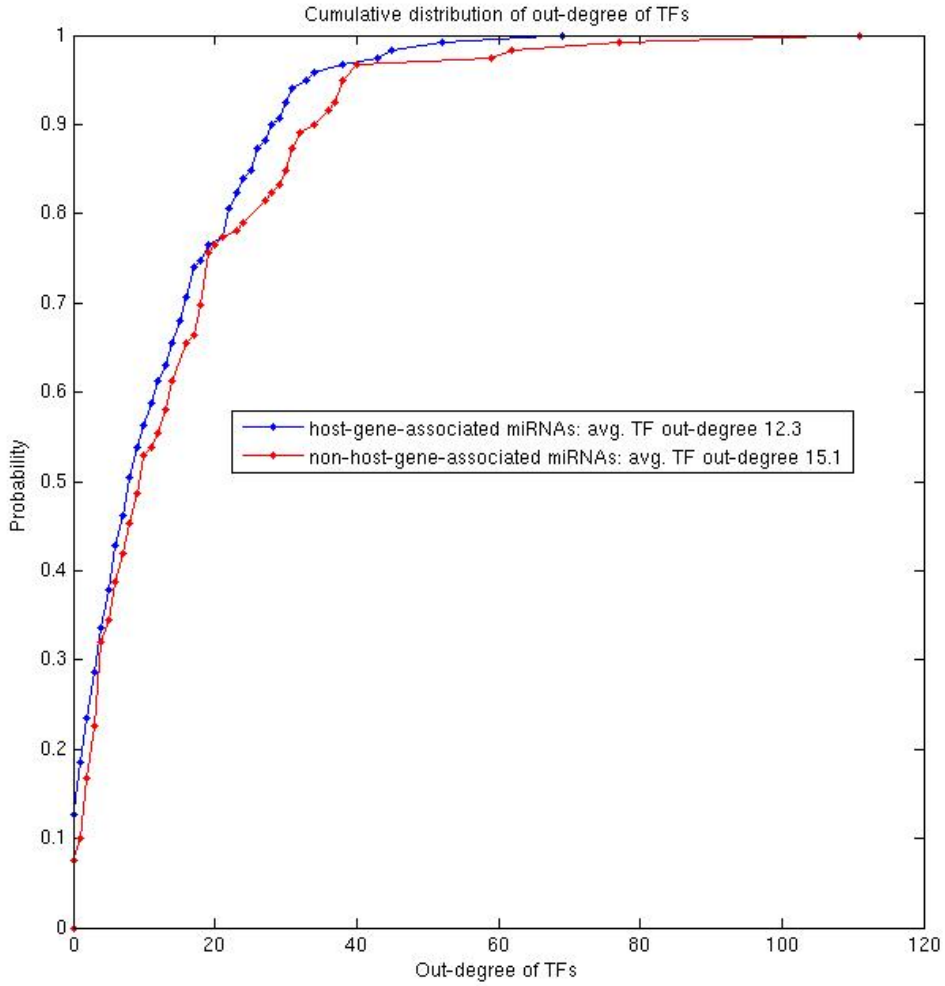
- I consider three obvious cell-line consistent histone modification experiments: 1. wgEncodeBroadHistoneGm12878CtcfStdAlnRep0 ,2.wgEncodeBroadHistoneK562CtcfStdAlnRep0,3. wgEncodeBroadHistoneK562Pol2bStdAlnRep0
- No. of Tars Host-Assoc. Non-Host-Assoc.

	Host-Assoc.			Non-Host-Assoc.		
	1	2	3	1	2	3
miRNA	23	25	25	29	33	8
miscRNA	7	9	2	22	20	8
snRNA	13	14	7	33	31	15
snoRNA	16	16	45	9	22	25

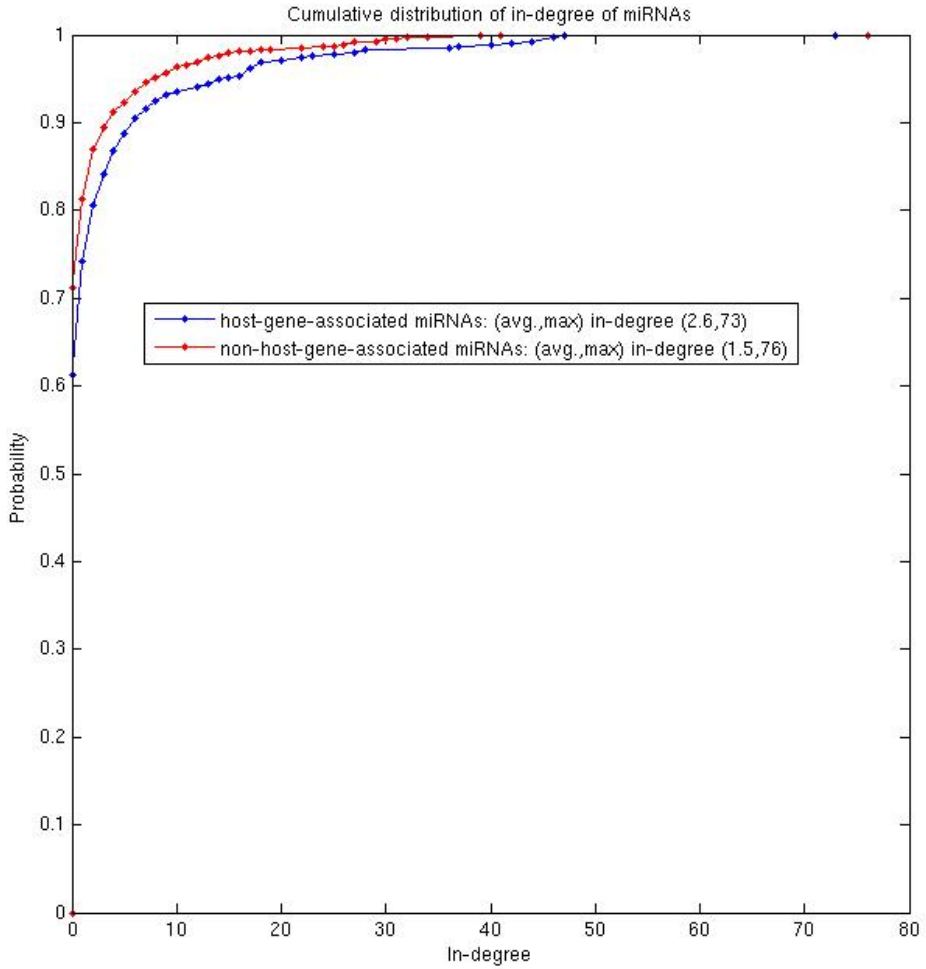
ENCODE ChIP-Seq Data for ncRNAs

- There are around 500 ChIP-Seq experiments for about 120 unique TFs
- To identify ncRNA targets of TFs, I used 1.5 KB upstream region of the starting position as promoters of miRNAs, misc_RNAs, snRNAs, and snoRNAs
- Because lincRNAs and processed transcripts are much longer, which are comparable or even longer than coding genes, I used 1.5KB upstream and 500B downstream of the starting position as promoters of lincRNAs and processed_transcript.

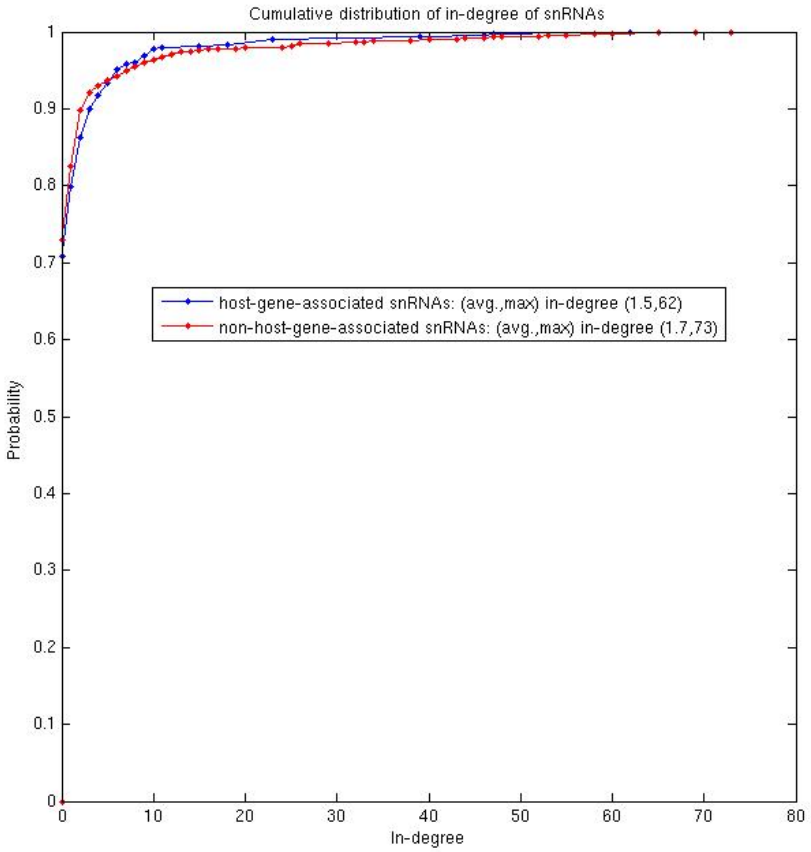
Out-degree of TFs Over miRNAs

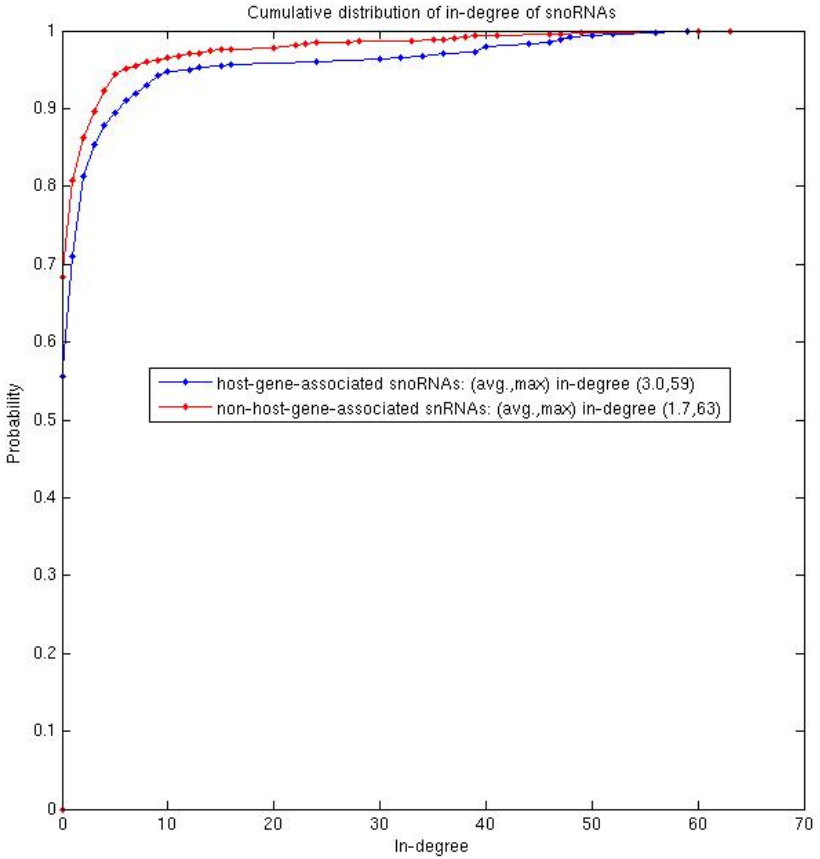


In-degree of miRNAs



In-degree of snRNAs and snoRNAs





Out-degree TF Hubs Regulating 16048 ncRNAs and 17874 Coding Genes

RAD21	1762
MAX	1756
MYC	1720
CEBPB	1302
YY1	1298
ELF1	1265
JUND	1152
HEY1	1113
STAT3	1082
FOXA1	1069
E2F6	1068
HDAC2	1062
SPI1	985
EGR1	970
SMC3	970
USF1	969
MAFK	933
TCF4	891
POU2F2	882
PAX5	867

MAX	12503
ELF1	11617
YY1	11459
E2F6	11011
HEY1	10330
MYC	9581
EGR1	9500
SIN3A	9409
E2F1	9252
HDAC2	9252
POU2F2	8564
PAX5	8440
NFKB1	8090
CHD2	7796
TCF12	7527
RAD21	7516
ZEB1	7514
TCF4	7412
SP1	7200
USF1	6812

Non-Coding RNA Associated TFs

XRCC4
BRF2
ZNF274
SMARCA4
BDP1
BRF1
POU5F1
ESR1
BATF
GATA2
JUN
MAFF
NR3C1
FOXA2
SUZ12
MAFK
FOSL1
TAL1
STAT3
POLR3A

- BRF2 subunit of RNA polymerase III
- POU5F1 embryonic development
- BRF1 subunit of RNA polymerase III
- BDP1 subunit of RNA polymerase III
- BATF Basic leucine zipper transcription factor, ATF-like
- ESR1 estrogen receptor 1

Coding Gene Associated TFs

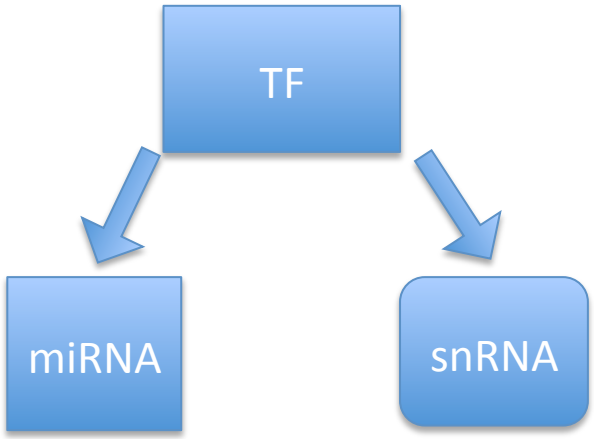
- THAP1
- IRF3
- SIX5
- IRF1
- ATF3
- E2F4
- PPARGC1A
- ELK4
- SP2
- HMGN3
- ETS1
- NRF1
- BRCA1
- SIN3A
- CCNT2
- MXI1
- ZEB1
- E2F1
- CHD2
- SMARCB1
- GABPA
- NR2C2
- ZBTB7A
- E2F6
- GTF2F1
- RFX5
- NFYA
- NFE2
- EGR1
- TCF12

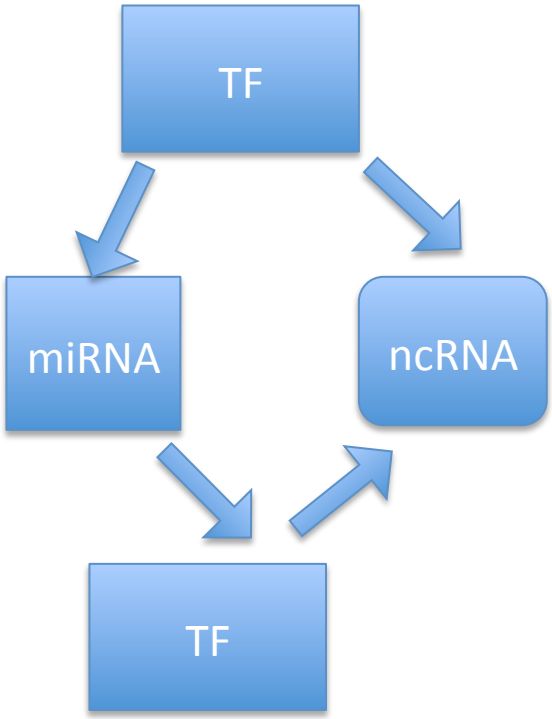
Properties of In-degree Hubs

- snoRNAs are highly enriched in-degree hubs

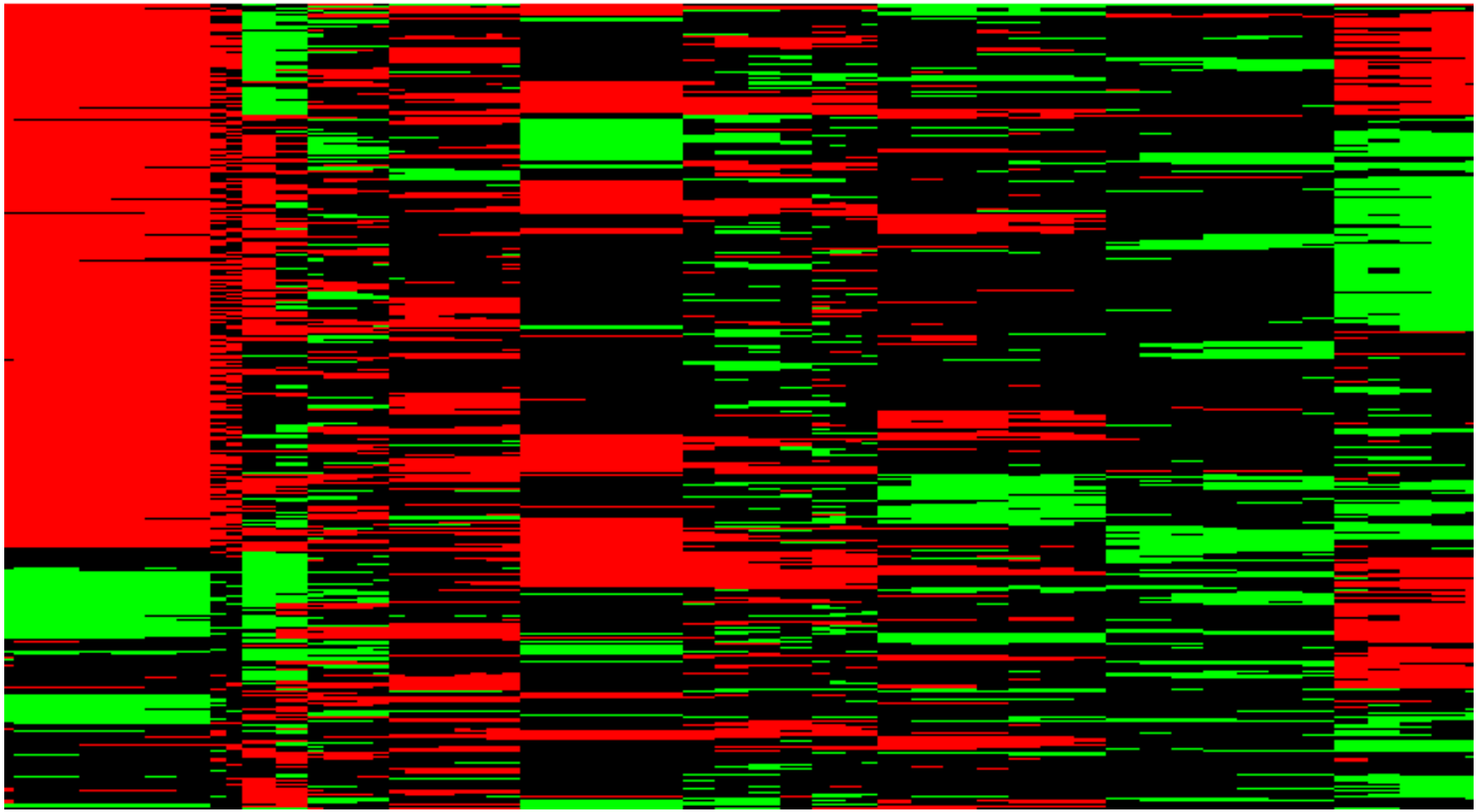
- Misc_RNAs and snRNAs tend to be under-represented in in-degree hubs

ncRNA Regulatory Network Motifs



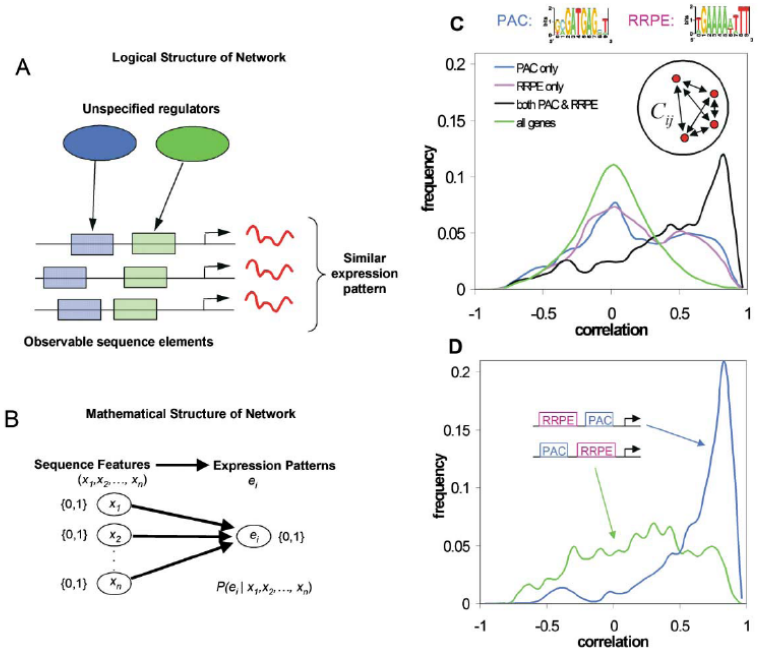


Annotating Functional Modules of ncRNAs By Clustering Expression



Predicting Interactions between Regulators for each Module

- Bayesian Networks
- Boosted Naïve Bayes Classifiers
- SVM



Performance of RBF-Kernel SVM

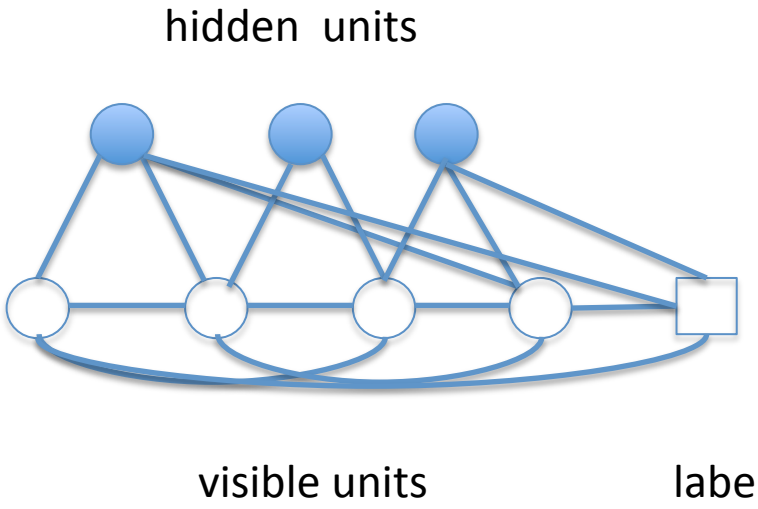
- Average One-vs-All ROC score based on kernel RBF-Kernel SVM is even above 0.80

- No tuning, no hacking, just simple testing, no cross validation yet

Why Not Bayesian Networks or SVM

- Hard to interpret the dependencies between features
- Here features are TF targetings

Semi-Supervised Semi-Restricted Boltzmann Machines (S^3RBM)



Learning Restricted Boltzmann Machines

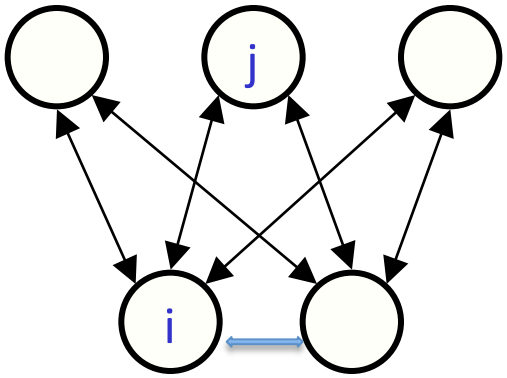
binary state of visible unit i

binary state of hidden unit j

$$E(v,h) = -\sum_{i,j} v_i h_j w_{ij} - \sum_{i,i'} v_i v_{i'} L_{ii'} - \sum_{i,y} v_i l_y L_{iy}$$

Energy with configuration v on the visible units and h on the hidden units

weight between units i and j

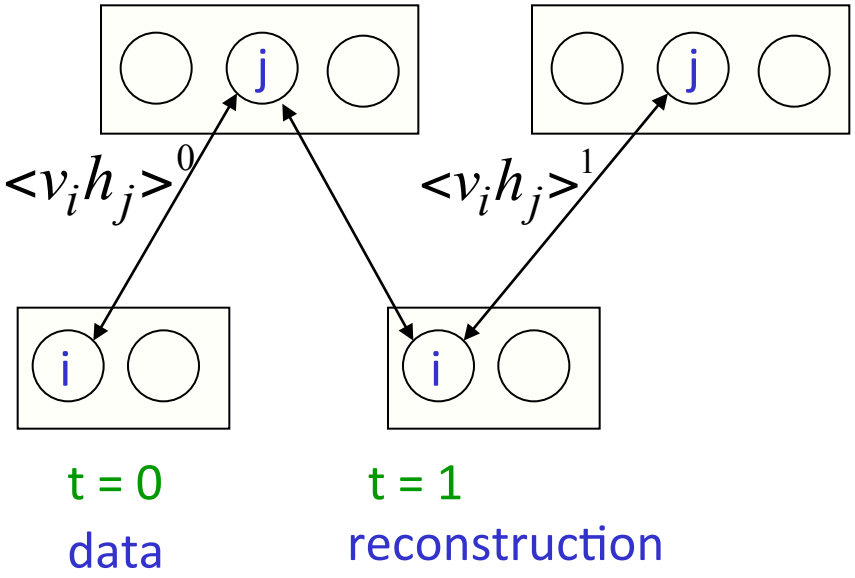


$$p(v,h) \propto e^{-E(v,h)}$$

$$\frac{\partial \log p(v)}{\partial w_{ij}} = <v_i h_j>^0 - <v_i h_j>^\infty$$

$$-\frac{\partial E(v,h)}{\partial w_{ij}} = v_i h_j$$

Learning Restricted Boltzmann Machines



Start with a training vector on the visible units.

Update all the hidden units in parallel

Update the all the visible units in parallel to get a "reconstruction".

Update the hidden units again.

$$\Delta w_{ij} = \varepsilon \left(<v_i h_j>^0 - <v_i h_j>^1 \right)$$

Learning Lateral Connections

- Mean-field Gibbs Sampling for inference

- stochastic MCMC for calculating model expectations to update weights

Predicting Expression from Chao's Histone Modification Data

- Discretize Expression and histone signals
- Learn S^3 RBM
- 5-fold cross validation accuracy 87%
- One-vs-All ROC score: above 90%
- Still in progress: interpreting dependencies between histone markers

Why S^3 RBM is Better than BN

- We can learn dependencies between features explicitly
- We can learn a dependency network for each functional module
- We can directly infer which features determine the other features with or without supervision signals

Random-Walk Models on Predicting (In)Direct, Stable, and Functional Bindings

- Utilizing all types of sources of high-throughput data and considering all the binding track data simultaneously
  - PPI
  - Binding Affinity based on PWMs
- construct a mixture model explaining which binding peak belongs to which TF

Other Projects

Conditional Random Field for predicting phenotypes from SVs

Random-walk based models for predicting gene functions by integrating network data and phenotype data

Acknowledgement

Mark Gerstein
Chao Cheng
Joel Rozowsky
Koon-Kiu Yan
Baikang Pei
Arif Harmanci
Jing Leng
Jing Wang
Xinmeng Mu