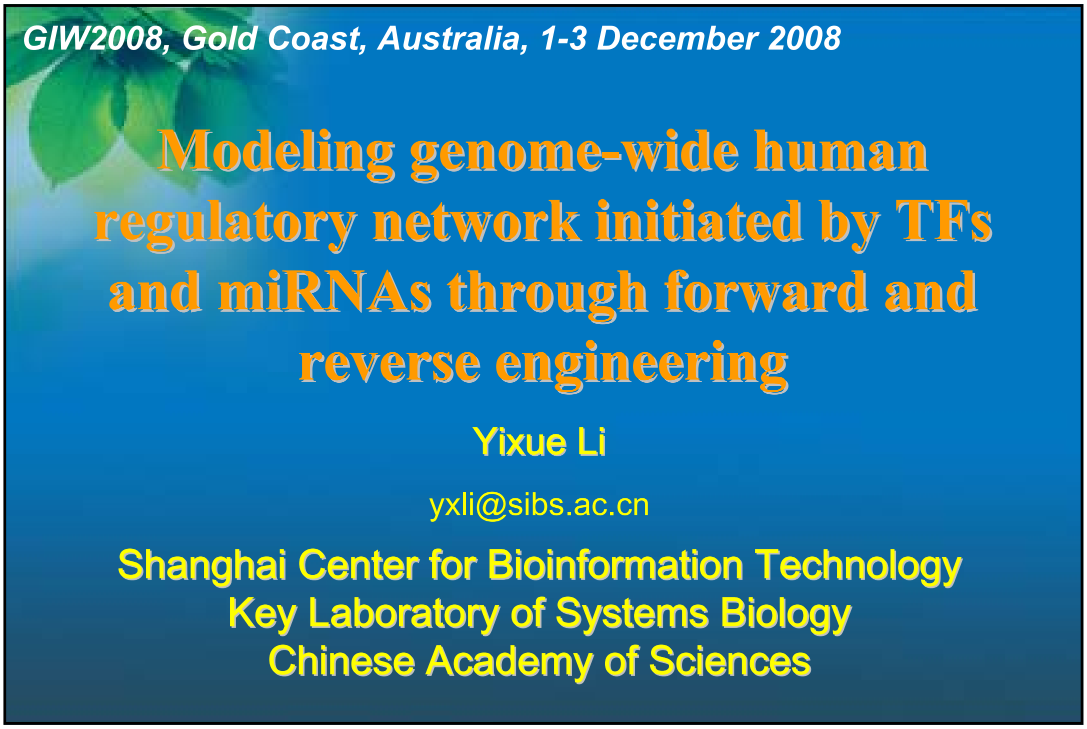


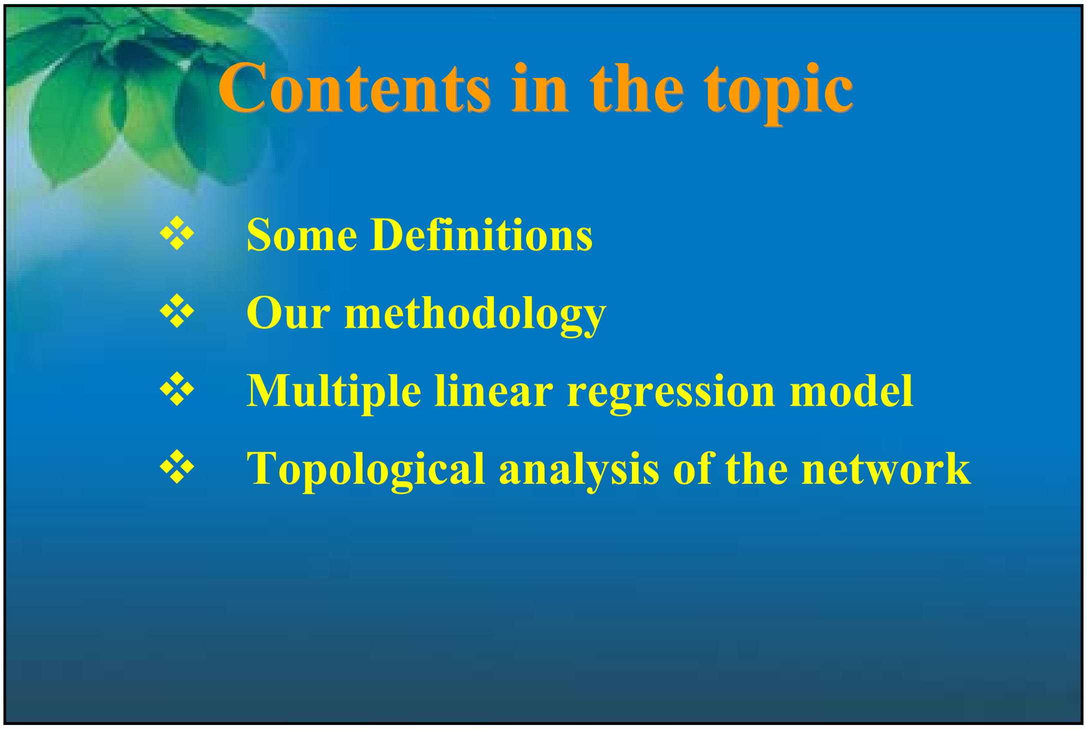
GIW2008, Gold Coast, Australia, 1-3 December 2008

Modeling genome-wide human regulatory network initiated by TFs and miRNAs through forward and reverse engineering

Yixue Li

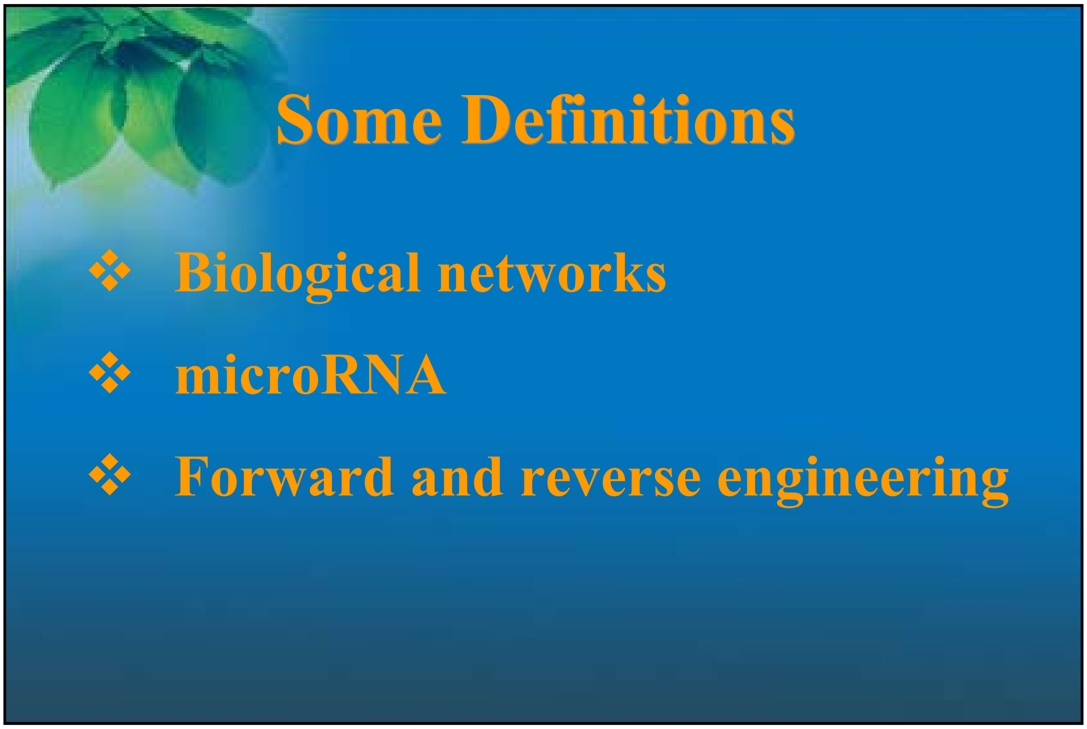
yxli@sibs.ac.cn

Shanghai Center for Bioinformation Technology
Key Laboratory of Systems Biology
Chinese Academy of Sciences



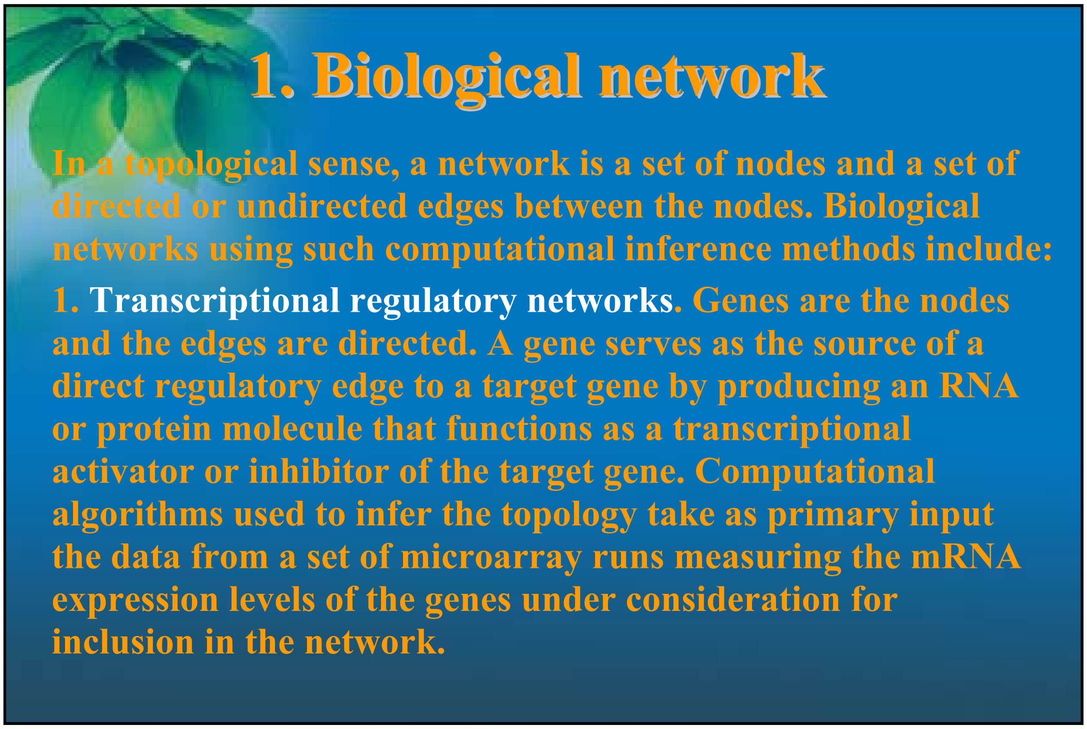
Contents in the topic

- ❖ Some Definitions
- ❖ Our methodology
- ❖ Multiple linear regression model
- ❖ Topological analysis of the network



Some Definitions

- Biological networks
- microRNA
- Forward and reverse engineering

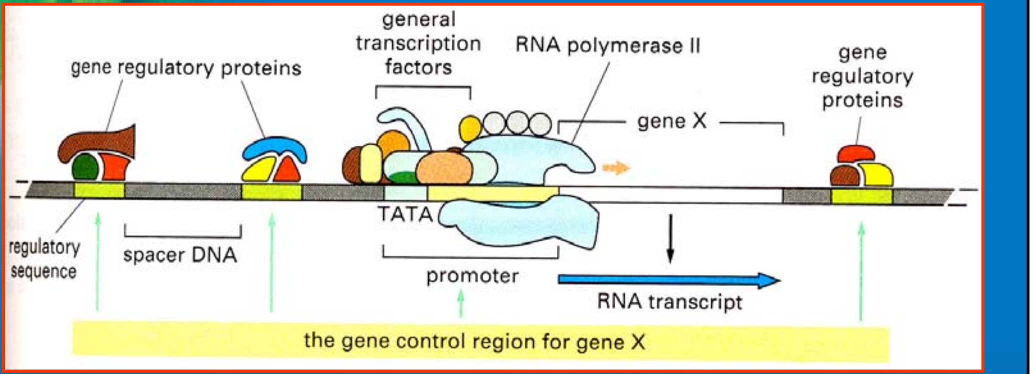


1. Biological network

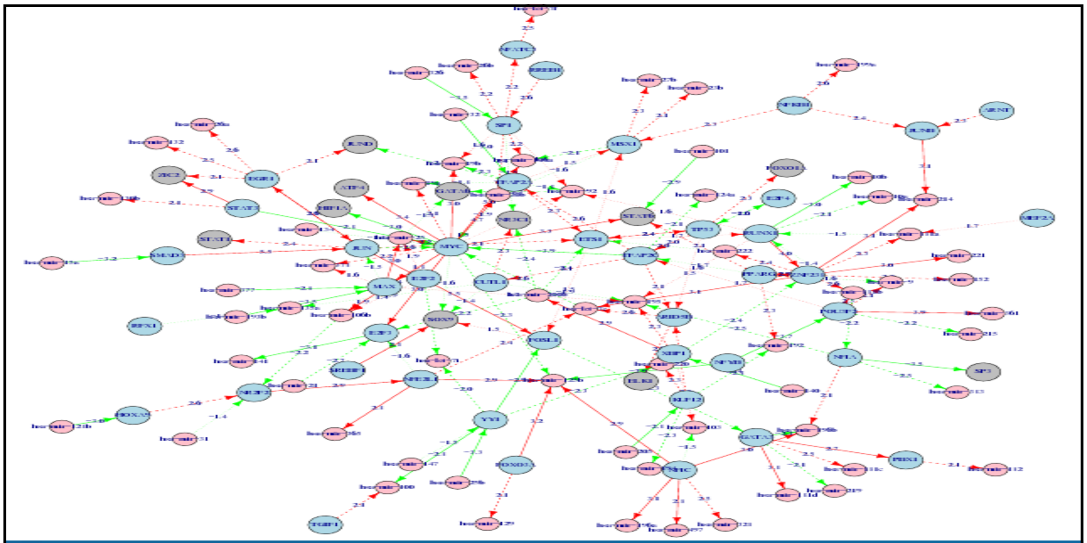
In a topological sense, a network is a set of nodes and a set of directed or undirected edges between the nodes. Biological networks using such computational inference methods include:

1. Transcriptional regulatory networks. Genes are the nodes and the edges are directed. A gene serves as the source of a direct regulatory edge to a target gene by producing an RNA or protein molecule that functions as a transcriptional activator or inhibitor of the target gene. Computational algorithms used to infer the topology take as primary input the data from a set of microarray runs measuring the mRNA expression levels of the genes under consideration for inclusion in the network.

Transcriptional regulatory networks



In the field of molecular biology, a transcription factor (sometimes called a sequence-specific DNA binding factor) is a protein that binds to specific parts of DNA using DNA binding domains and is part of the system that controls the transfer (or transcription) of genetic information from DNA to RNA. Transcription factors perform this function alone, or by using other proteins in a complex, by increasing (as an activator), or preventing (as a repressor) the presence of RNA polymerase, the enzyme which activates the transcription of genetic information from DNA to RNA.

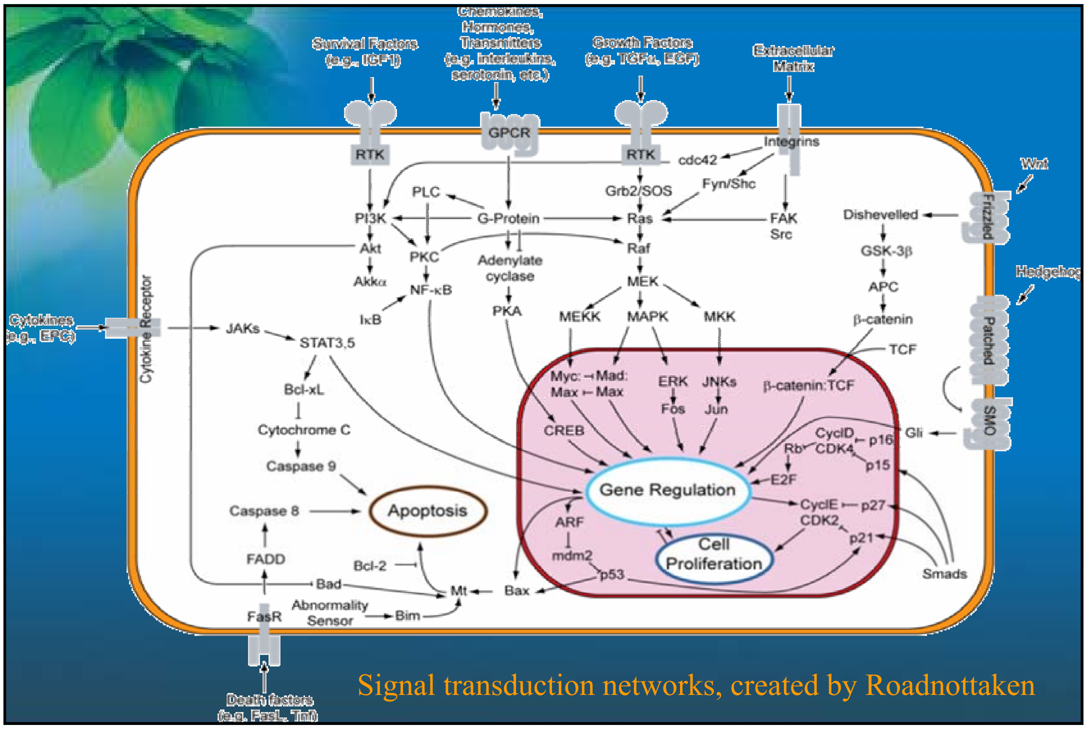


Transcriptional regulatory networks. Genes are the nodes and the edges are directed. A gene serves as the source of a direct regulatory edge to a target gene by producing an RNA or protein molecule that functions as a transcriptional activator or inhibitor of the target gene. If the gene is an activator, then it is the source of a positive regulatory connection; if an inhibitor, then it is the source of a negative regulatory connection.

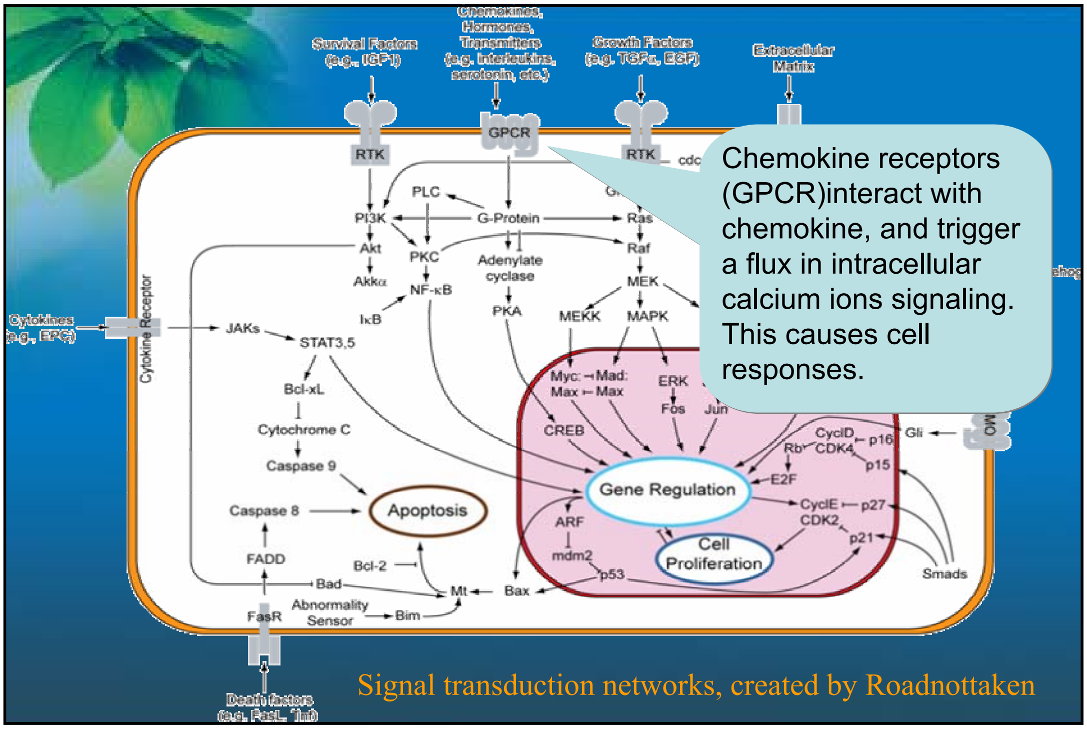
Biological network

2. **Signal transduction networks**(very important in the biology of cancer). Proteins are the nodes and the edges are directed. Primary input into the inference algorithm would be data from a set of experiments measuring protein activation / inactivation (e.g., **phosphorylation/ dephosphorylation**) across a set of proteins.

3. **Metabolic networks.** Metabolites are the nodes and the edges are directed. Primary input into an algorithm would be data from a set of experiments measuring metabolite levels.

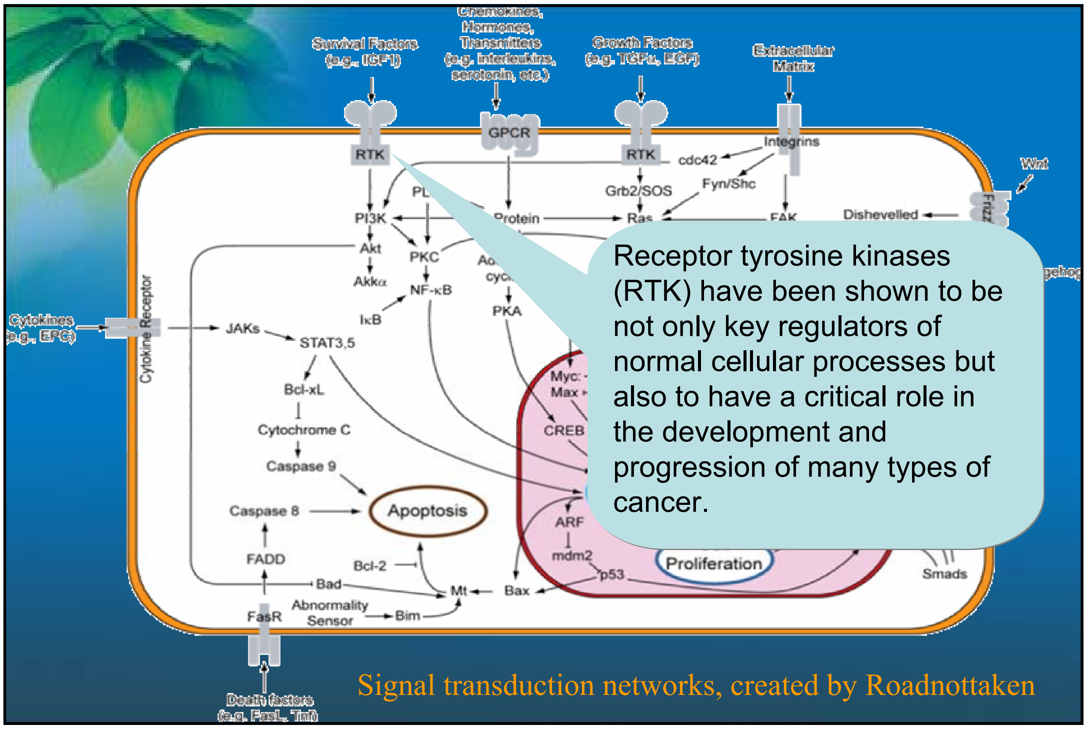


Signal transduction networks, created by Roadnottaken



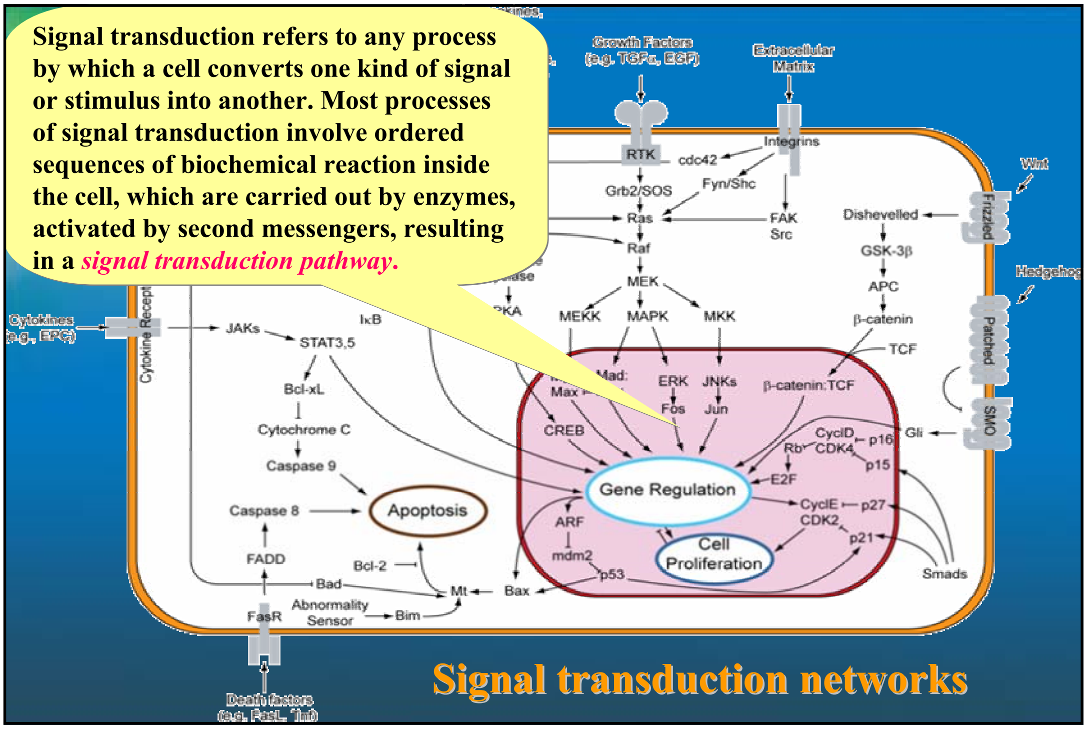
Chemokine receptors (GPCR)interact with chemokine, and trigger a flux in intracellular calcium ions signaling. This causes cell responses.

Signal transduction networks, created by Roadnottaken



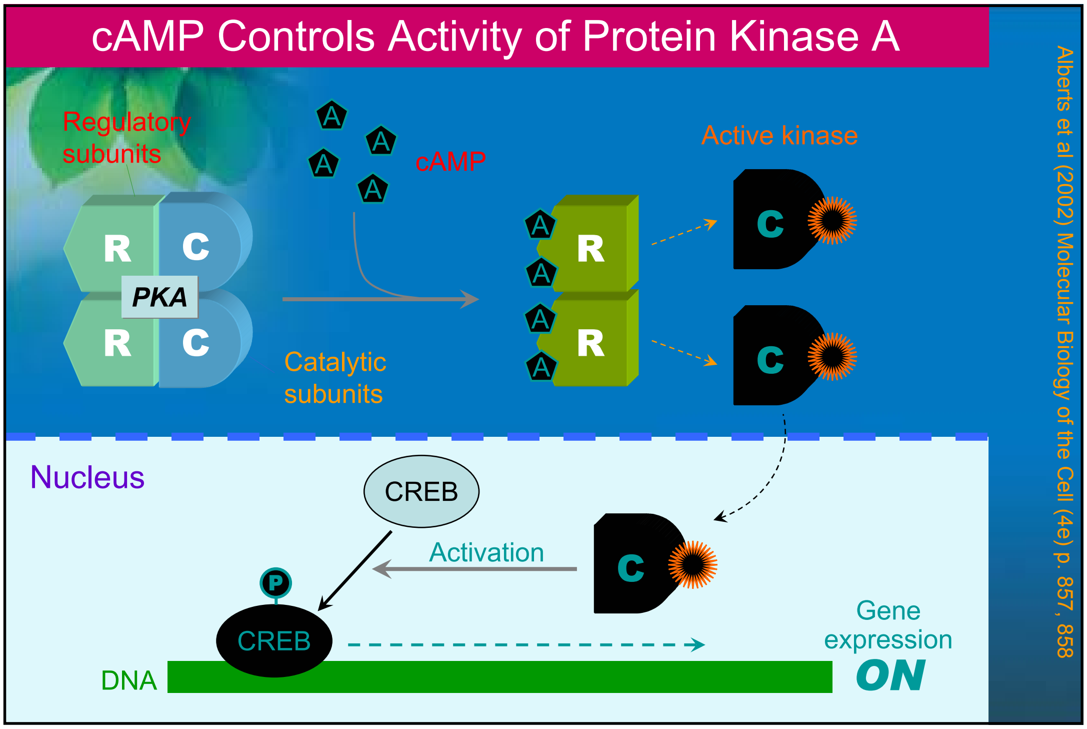
Receptor tyrosine kinases (RTK) have been shown to be not only key regulators of normal cellular processes but also to have a critical role in the development and progression of many types of cancer.

Signal transduction networks, created by Roadnottaken



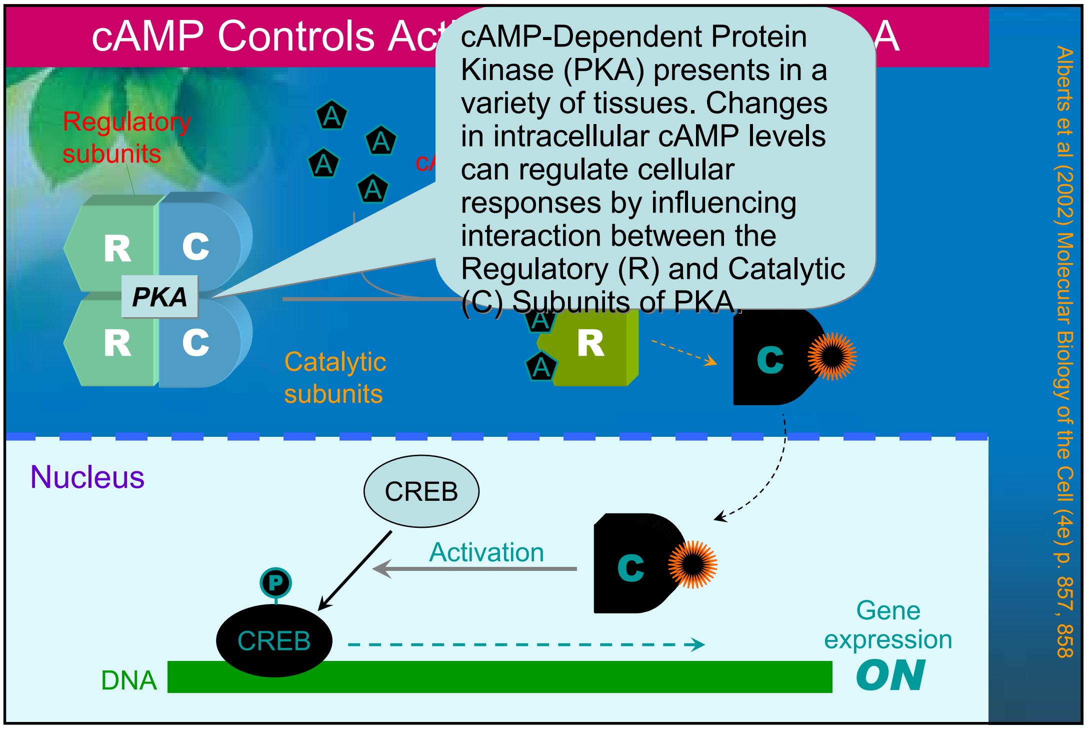
Signal transduction refers to any process by which a cell converts one kind of signal or stimulus into another. Most processes of signal transduction involve ordered sequences of biochemical reaction inside the cell, which are carried out by enzymes, activated by second messengers, resulting in a ***signal transduction pathway.***

Signal transduction networks



cAMP Controls Activity of Protein Kinase A

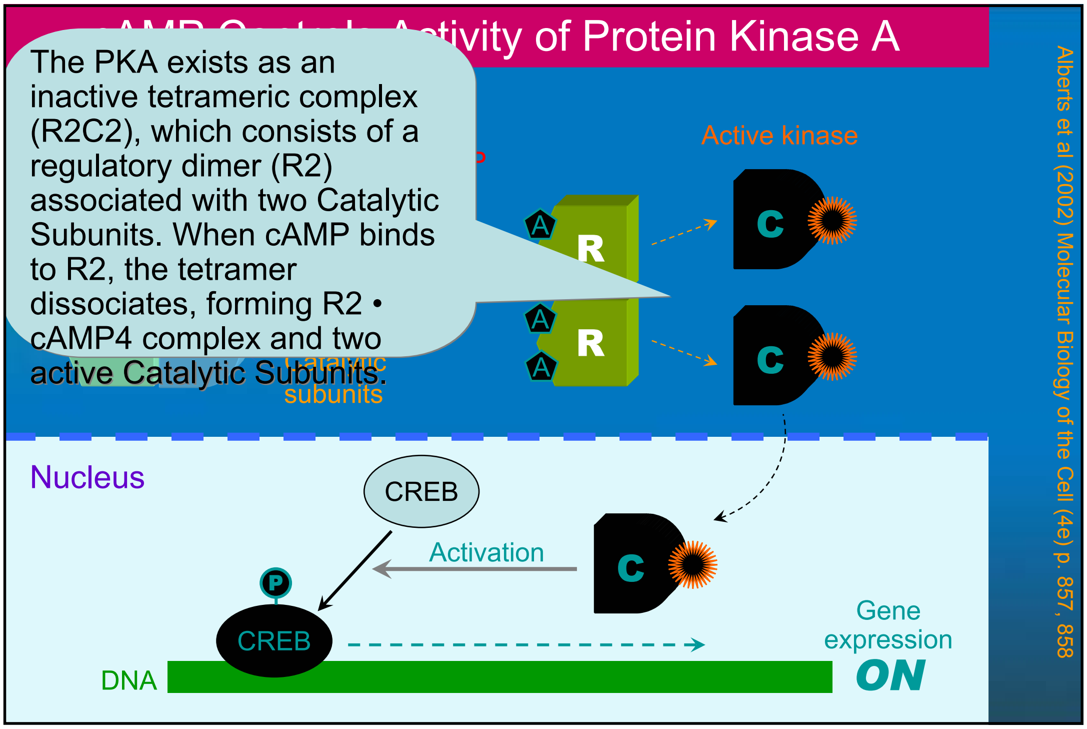
Alberts et al (2002) Molecular Biology of the Cell (4e) p. 857, 858



cAMP Controls Act A

cAMP-Dependent Protein Kinase (PKA) presents in a variety of tissues. Changes in intracellular cAMP levels can regulate cellular responses by influencing interaction between the Regulatory (R) and Catalytic (C) Subunits of PKA.

Alberts et al (2002) Molecular Biology of the Cell (4e) p. 857, 858

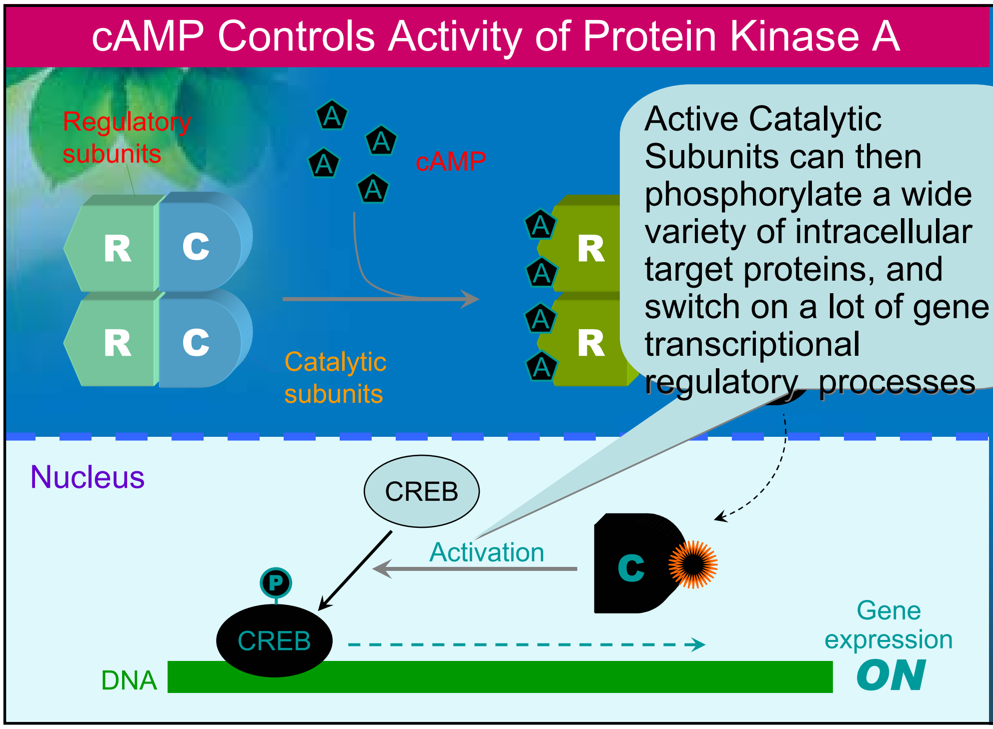


cAMP Controls Activity of Protein Kinase A

The PKA exists as an inactive tetrameric complex (R2C2), which consists of a regulatory dimer (R2) associated with two Catalytic Subunits. When cAMP binds to R2, the tetramer dissociates, forming R2 • cAMP4 complex and two active Catalytic Subunits.

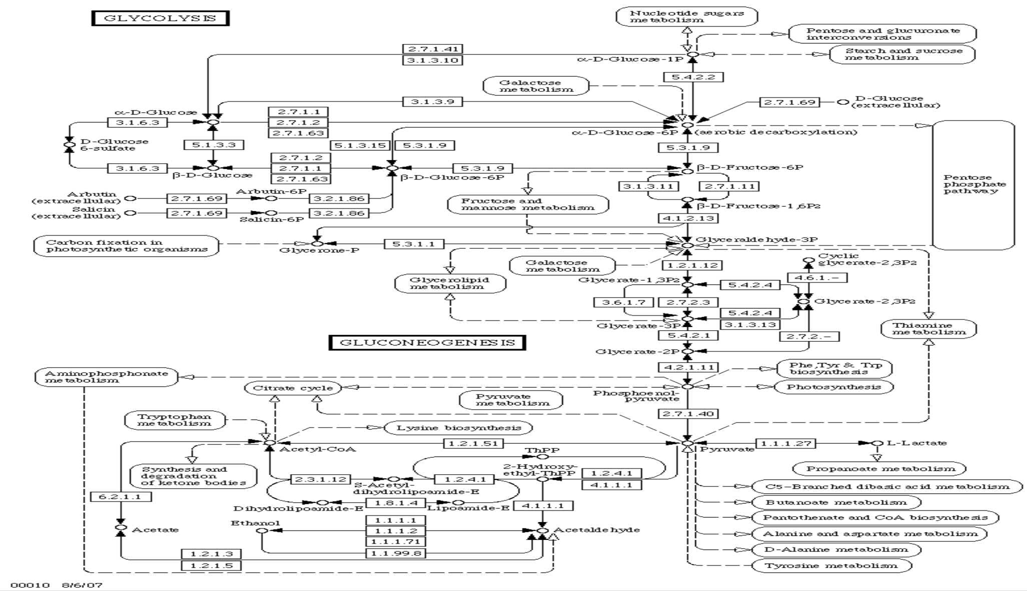
Alberts et al (2002) Molecular Biology of the Cell (4e) p. 857, 858

cAMP Controls Activity of Protein Kinase A

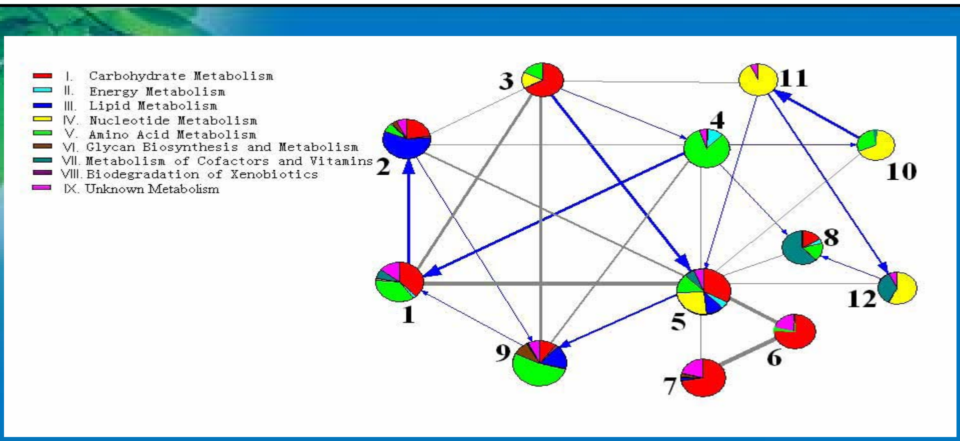


Active Catalytic Subunits can then phosphorylate a wide variety of intracellular target proteins, and switch on a lot of gene transcriptional regulatory processes

Alberts et al (2002) Molecular Biology of the Cell (4e) p. 857, 858



Metabolic networks, Glycolysis / Gluconeogenesis reference metabolic pathway from KEGG.



I. Carbohydrate Metabolism
II. Energy Metabolism
III. Lipid Metabolism
IV. Nucleotide Metabolism
V. Amino Acid Metabolism
VI. Glycan Biosynthesis and Metabolism
VII. Metabolism of Cofactors and Vitamins
VIII. Biodegradation of Xenobiotics
IX. Unknown Metabolism
1
2
3
4
5
6
7
8
9
10
11
12

Cartographic representation of the metabolic network for *E.coli*. Each circle represents a module and is coloured according to the KEGG pathway classification of the reactions belonging to it, while the arcs reflect the connection between clusters. The area of each colour in one circle is proportional to the number of reactions that belong to the corresponding metabolism. The width of an arc is proportional to the number of reactions between the two corresponding modules. For simplicity, bi-directed arcs are presented by grey edges. Jing et.al., ***BMC Bioinformatics*** 2006, 7:386.

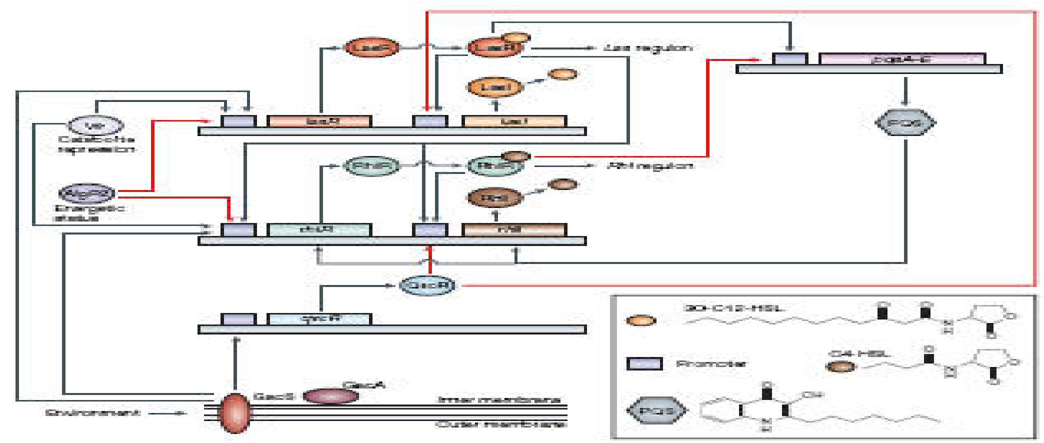
Biological network

5. Intraspecies or interspecies communication networks in microbial communities. Nodes are excreted organic compounds and the edges are directed. Input into an inference algorithm is data from a set of experiments measuring levels of excreted molecules.

6. Protein-protein interaction networks are also under very active study. However, reconstruction of these networks does not use correlation-based inference in the sense discussed for the networks already described (interaction does not necessarily imply a change in protein state)

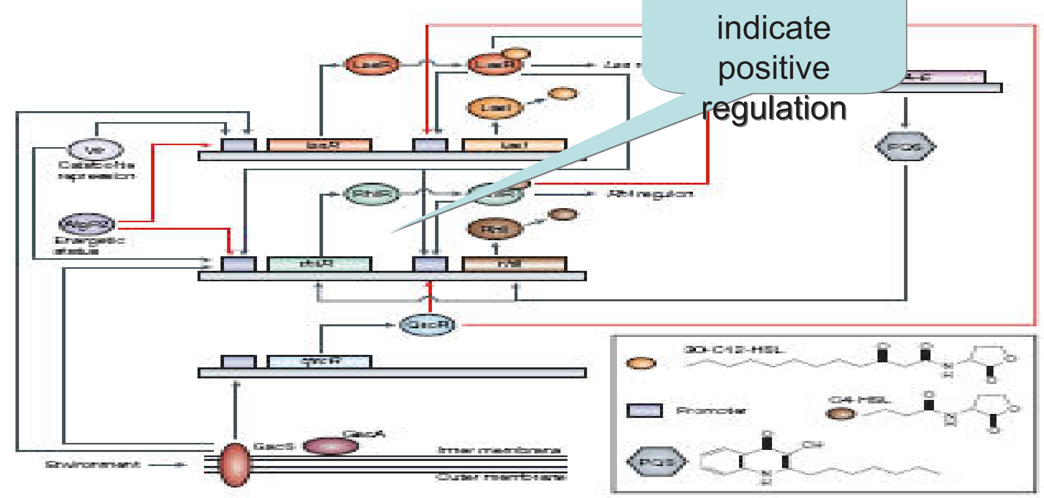
Intraspecies or interspecies communication networks

The quorum-sensing paradigm in Gram-negative and positive bacteria



The Las, Rhl and Qsc quorum-sensing systems in *Pseudomonas aeruginosa*: hierarchies and integration into cellular control circuits. For each circuit in the cell the interactions between the different QS systems are indicated by arrows. Black arrows indicate positive regulation and red arrows indicate negative regulation. Signals from the environment, the intracellular metabolic status of the cell and other regulators, such as RpoS, RsmA and

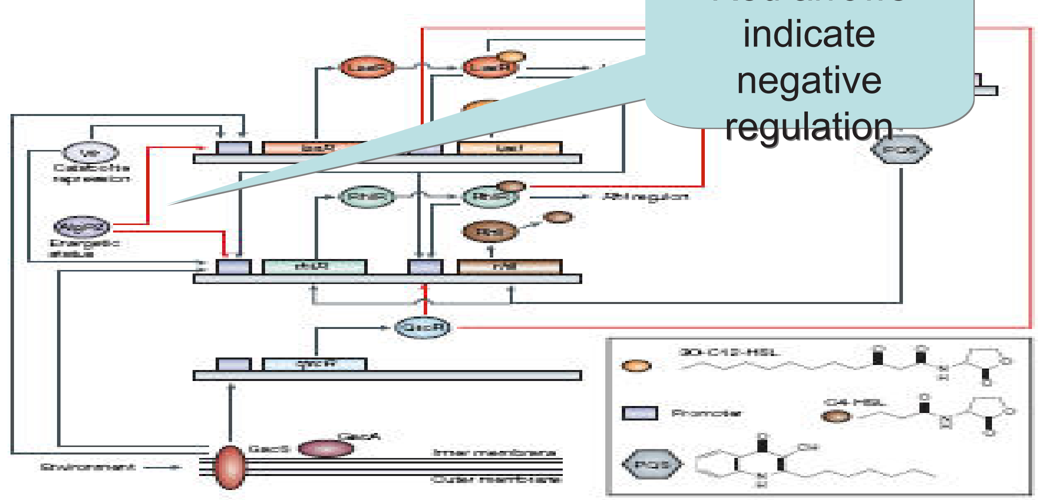
Intraspecies or interspecies communication networks



Black arrows indicate positive regulation

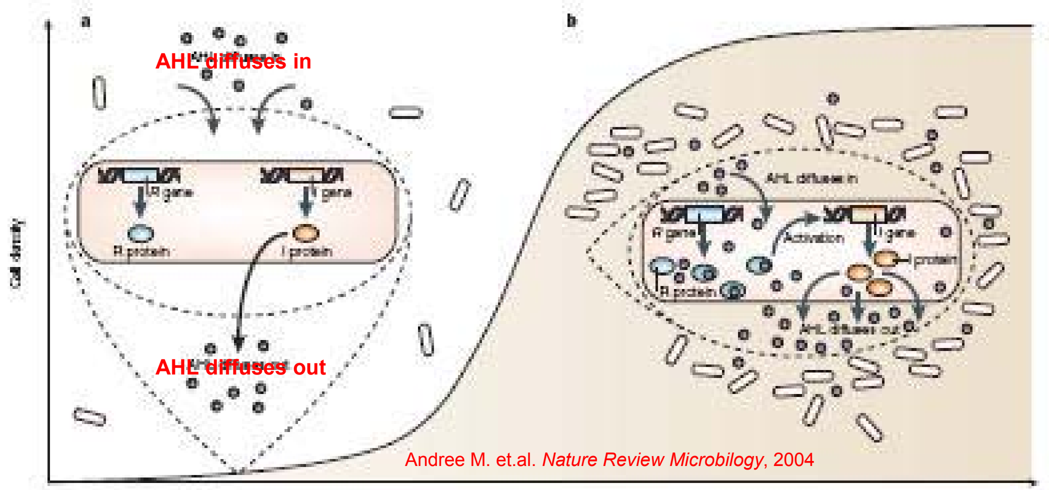
The Las, Rhl and Qsc quorum-sensing systems in *Pseudomonas aeruginosa*: hierarchies and integration into cellular control circuits. For each circuit in the cell the interactions between the different QS systems are indicated by arrows. Black arrows indicate positive regulation and red arrows indicate negative regulation. Signals from the environment, the intracellular metabolic status of the cell and other regulators, such as RpoS, RsmA and

Intraspecies or interspecies comm ks



Red arrows indicate negative regulation

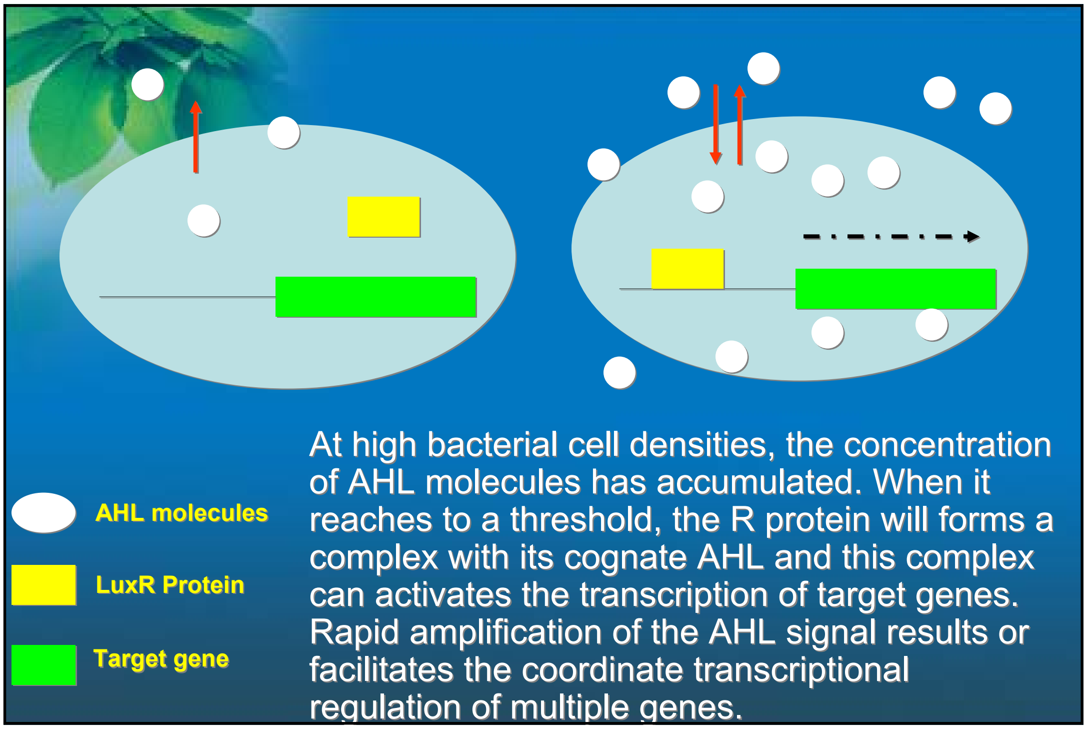
The Las, Rhl and Qsc quorum-sensing systems in *Pseudomonas aeruginosa*: hierarchies and integration into cellular control circuits. For each circuit in the cell the interactions between the different QS systems are indicated by arrows. Black arrows indicate positive regulation and red arrows indicate negative regulation. Signals from the environment, the intracellular metabolic status of the cell and other regulators, such as RpoS, RsmA and



Andree M. et.al. *Nature Review Microbilogy*, 2004

The quorum-sensing paradigm in Gram-negative bacteria

A. At low bacterial cell densities AHL molecules are synthesized and accumulate. Depending on the length of the acyl chain, AHLs either diffuse or are pumped out of the cell into the local environment, where the AHL molecules are available for diffusion into, or uptake by, bacterial cells.

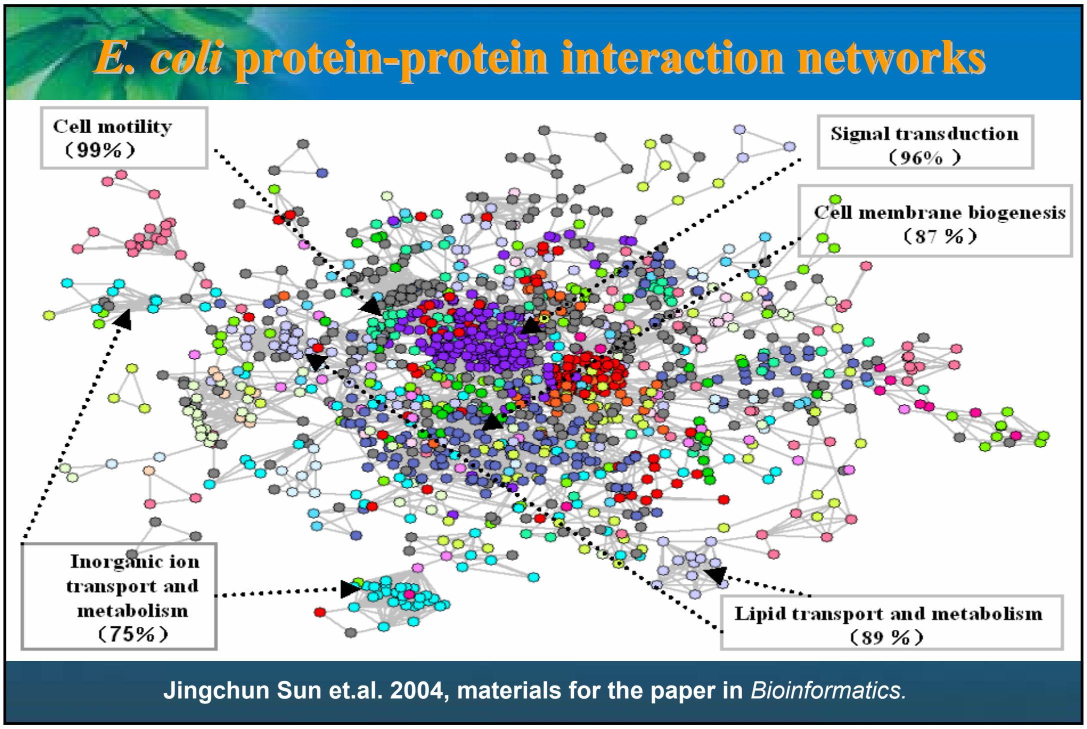


○ AHL molecules

🟨 LuxR Protein

🟩 Target gene

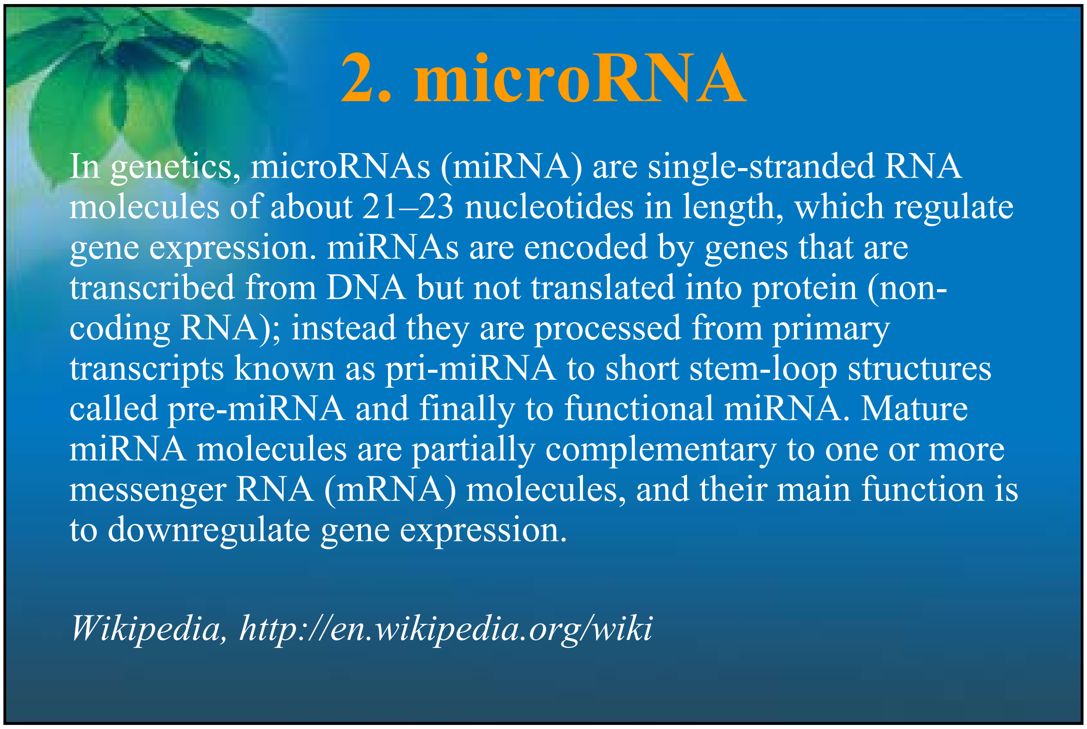
At high bacterial cell densities, the concentration of AHL molecules has accumulated. When it reaches to a threshold, the R protein will forms a complex with its cognate AHL and this complex can activates the transcription of target genes. Rapid amplification of the AHL signal results or facilitates the coordinate transcriptional regulation of multiple genes.



E. coli protein-protein interaction networks

Cell motility (99%)
Signal transduction (96%)
Cell membrane biogenesis (87 %)
Inorganic ion transport and metabolism (75%)
Lipid transport and metabolism (89 %)

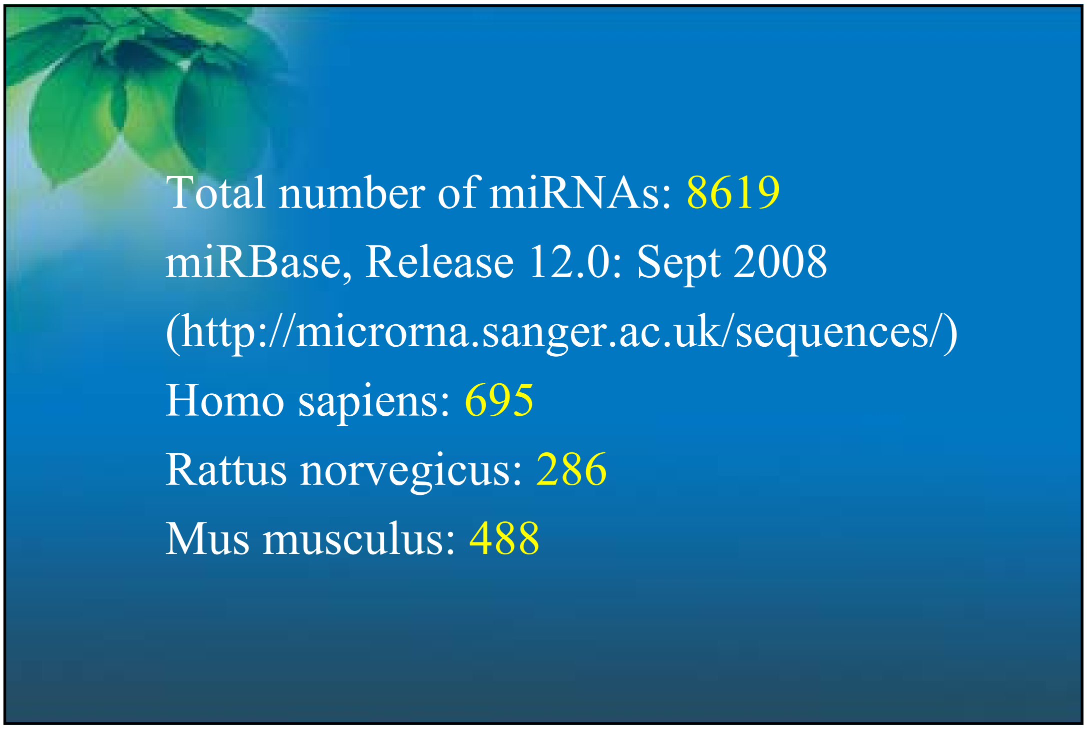
Jingchun Sun et.al. 2004, materials for the paper in *Bioinformatics.*



2. microRNA

In genetics, microRNAs (miRNA) are single-stranded RNA molecules of about 21–23 nucleotides in length, which regulate gene expression. miRNAs are encoded by genes that are transcribed from DNA but not translated into protein (non-coding RNA); instead they are processed from primary transcripts known as pri-miRNA to short stem-loop structures called pre-miRNA and finally to functional miRNA. Mature miRNA molecules are partially complementary to one or more messenger RNA (mRNA) molecules, and their main function is to downregulate gene expression.

Wikipedia, http://en.wikipedia.org/wiki



Total number of miRNAs: 8619

miRBase, Release 12.0: Sept 2008

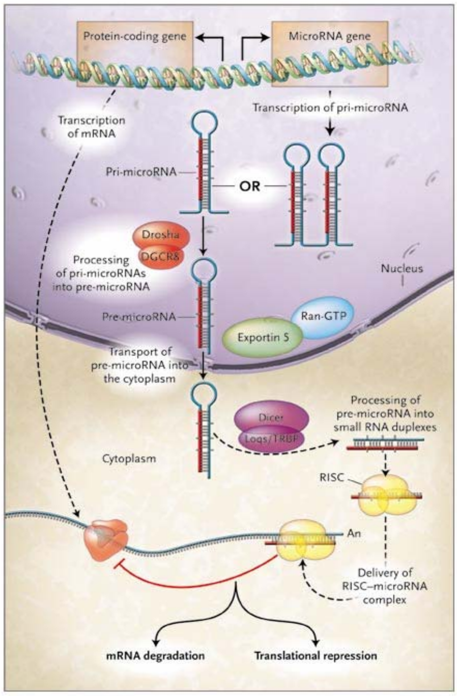
(http://microrna.sanger.ac.uk/sequences/)

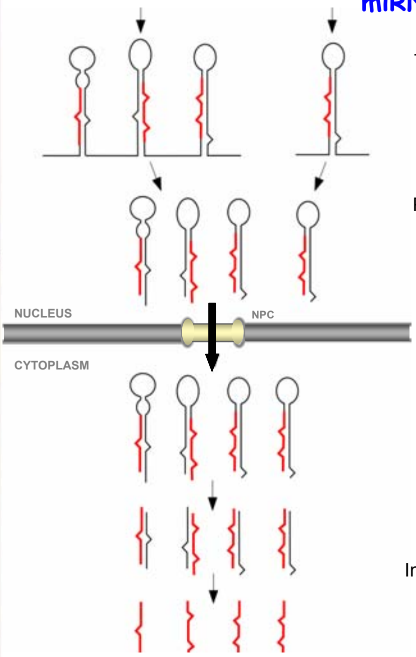
Homo sapiens: 695

Rattus norvegicus: 286

Mus musculus: 488

miRNA Biogenesis


Protein-coding gene
MicroRNA gene
Transcription of mRNA
Transcription of pri-microRNA
Pri-microRNA
OR
Drosha
DGCR8
Processing of pri-microRNAs into pre-microRNA
Nucleus
Pre-microRNA
Ran-GTP
Exportin 5
Transport of pre-microRNA into the cytoplasm
Processing of pre-microRNA into small RNA duplexes
Dicer
Loqs/TRBP
Cytoplasm
RISC
An
Delivery of RISC–microRNA complex
mRNA degradation
Translational repression



Transcription by Pol II
pri-miRNA
(100s-1000s nts.)

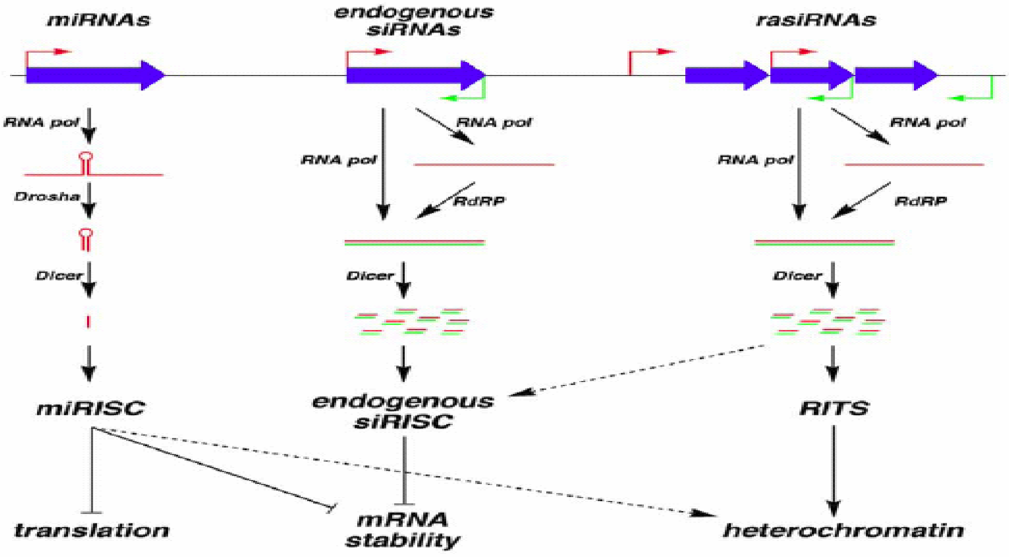
Processing by Drosha
pre-miRNAs
(~60 nts.)

Export by Exportin-5

Processing by Dicer
ds miRNAs
(~44 nts.)

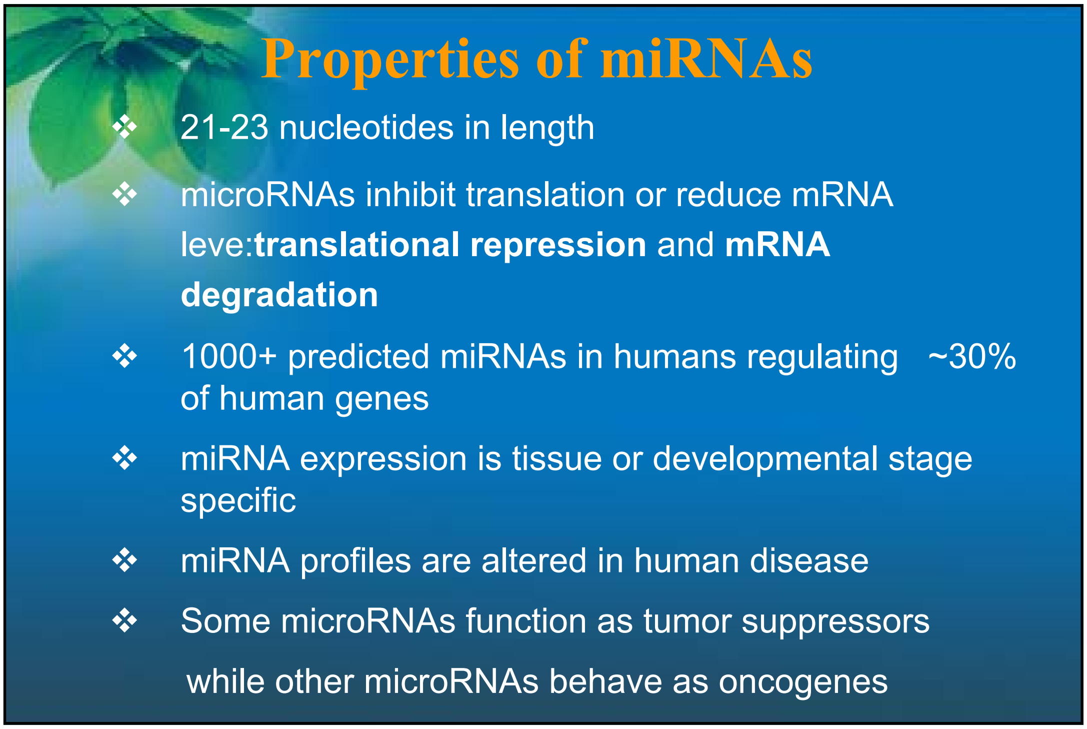
Incorporation into RISC
by Argonaute
miRNA
(~22 nts.)

(repeat-associated siRNAs)



Sontheimer and Carthew (2005) ***Cell*** 122, 9-12
Andre,Verdel (2004) ***Science*** 672-676

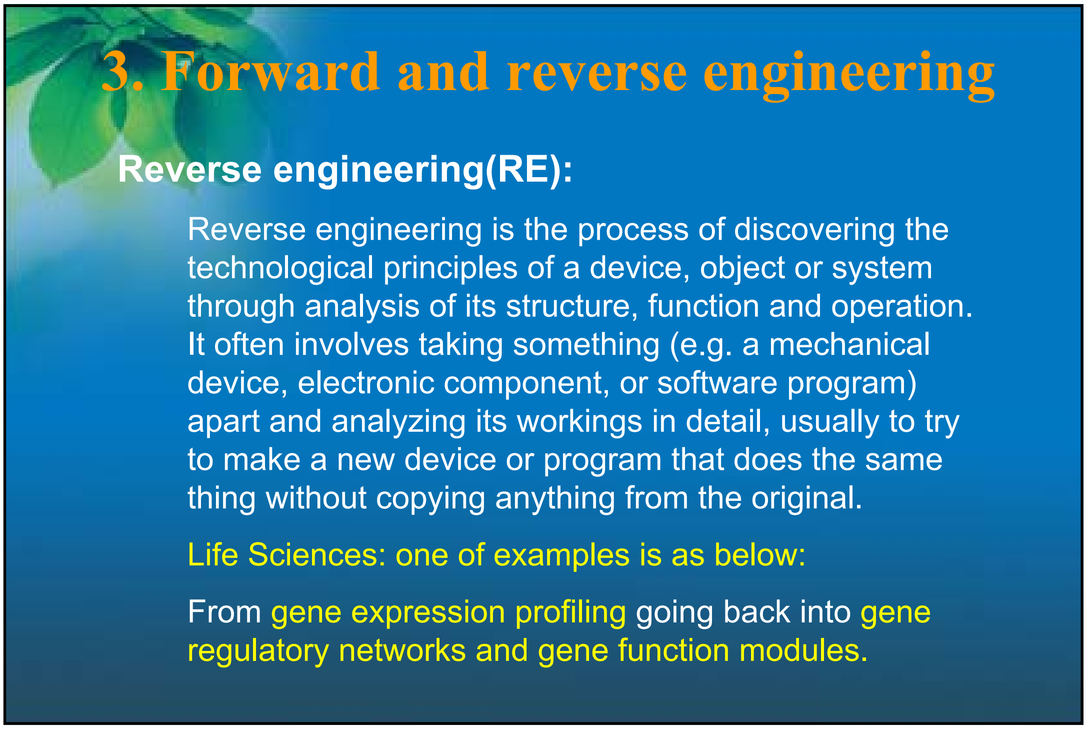
Guide heterochromatin
effectors to DNA



Properties of miRNAs

- 21-23 nucleotides in length
- microRNAs inhibit translation or reduce mRNA leve:**translational repression** and **mRNA degradation**
- 1000+ predicted miRNAs in humans regulating ~30% of human genes
- miRNA expression is tissue or developmental stage specific
- miRNA profiles are altered in human disease
- Some microRNAs function as tumor suppressors

  while other microRNAs behave as oncogenes



3. Forward and reverse engineering

Reverse engineering(RE):

Reverse engineering is the process of discovering the technological principles of a device, object or system through analysis of its structure, function and operation. It often involves taking something (e.g. a mechanical device, electronic component, or software program) apart and analyzing its workings in detail, usually to try to make a new device or program that does the same thing without copying anything from the original.

Life Sciences: one of examples is as below:

From gene expression profiling going back into gene regulatory networks and gene function modules.

Forward engineering(FE):

Forward engineering is the process of moving from a high-level abstraction and design to a low- level implementation. In the most time, the forward engineering based on the insights obtained from reverse engineering then systematically improve the protocols of implementations.

Life Sciences: one of examples is as below:

From mRNAs, miRNSs, proteins going to gene regulatory networks, PPI networks and metabolic pathways.

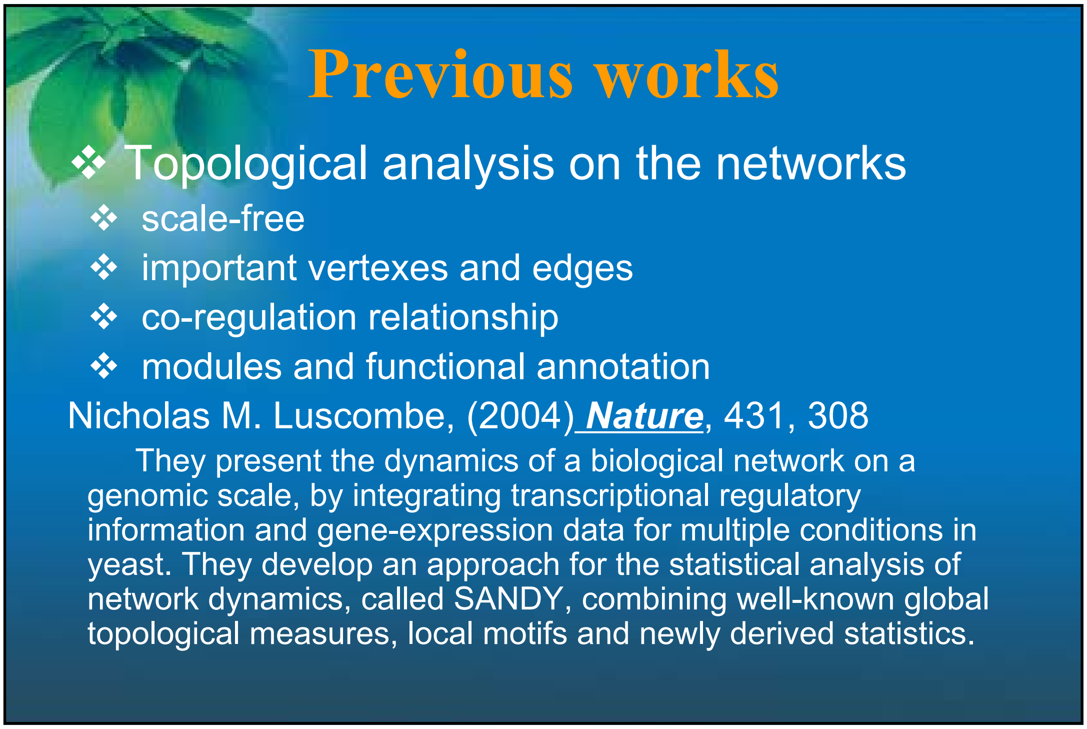
Previous works

- Constructing three different types of regulations
  - TF -> gene
  - microRNA -> gene → **Combinatory regulation networks**
  - TF -> microRNA
- M. Levine, R. Tjian, (2003) ***Nature*** 424, 147.

  Emerging evidence suggests that organismal complexity arises from progressively more elaborate regulation of gene expression.
- Hobert, (2004) Trends ***Biochem Sci***. 29, 462.

  TFs and miRNAs act in a largely combinatorial manner - that is, many different TFs or miRNAs control one gene - and they act cooperatively on their targets - that is, there are several cis-regulatory elements for a single TF or miRNA species in a target gene
- George A. Calin , (2006) ***Nature Reviews, CANCER***, 6

  MiRNA-expression profiling of human tumours has identified signatures associated with diagnosis, staging, progression, prognosis and response to treatment. Sometimes miRNA genes might represent downstream targets of activated oncogenic pathways, or they target protein-coding genes involved in cancer.

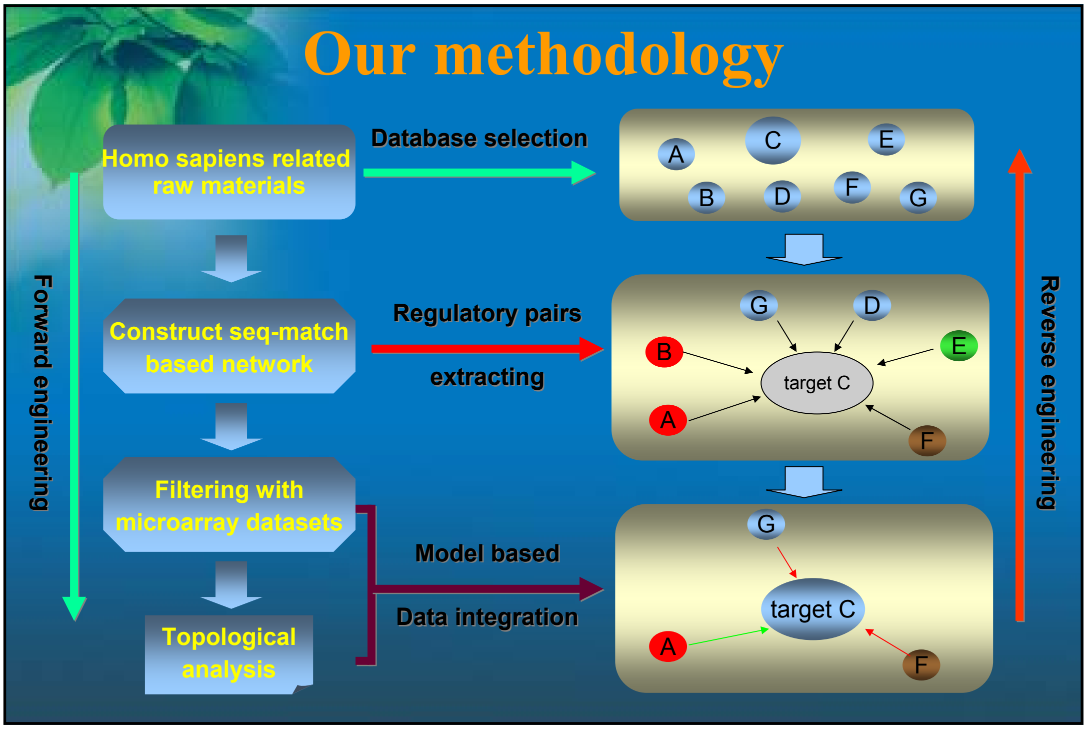


Previous works

- ❖ Topological analysis on the networks
  - ❖ scale-free
  - ❖ important vertexes and edges
  - ❖ co-regulation relationship
  - ❖ modules and functional annotation

Nicholas M. Luscombe, (2004) ***Nature***, 431, 308

They present the dynamics of a biological network on a genomic scale, by integrating transcriptional regulatory information and gene-expression data for multiple conditions in yeast. They develop an approach for the statistical analysis of network dynamics, called SANDY, combining well-known global topological measures, local motifs and newly derived statistics.



Our methodology

Homo sapiens related raw materials

Construct seq-match based network

Filtering with microarray datasets

Topological analysis

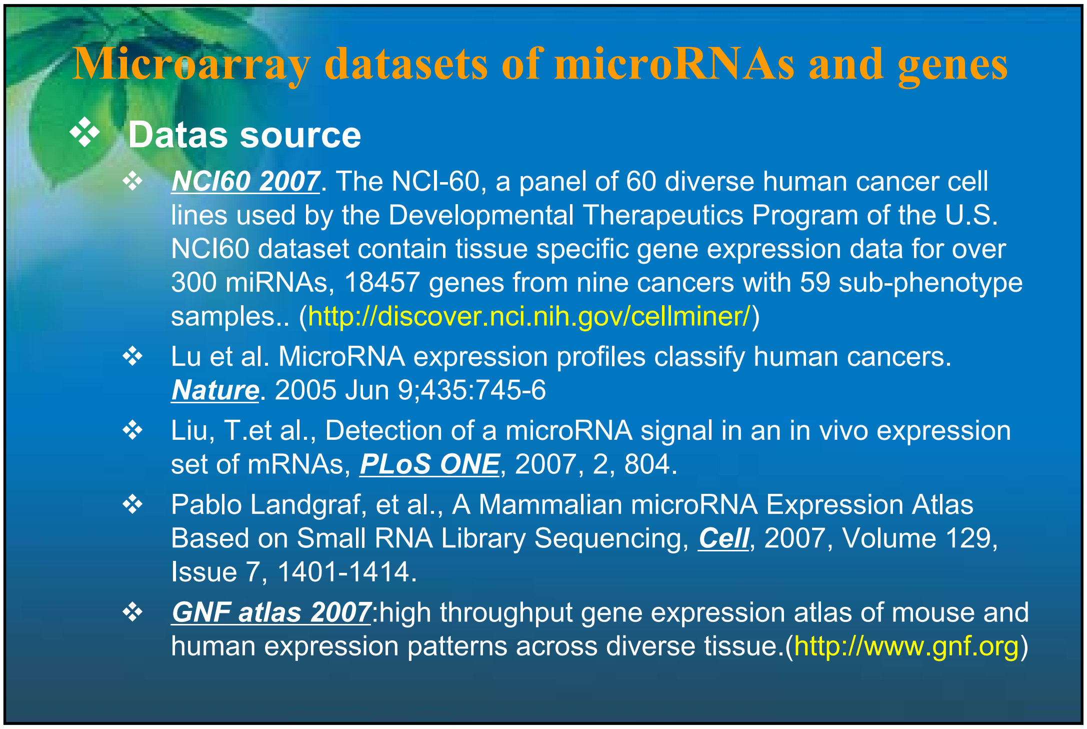
Database selection

Regulatory pairs extracting

Model based Data integration

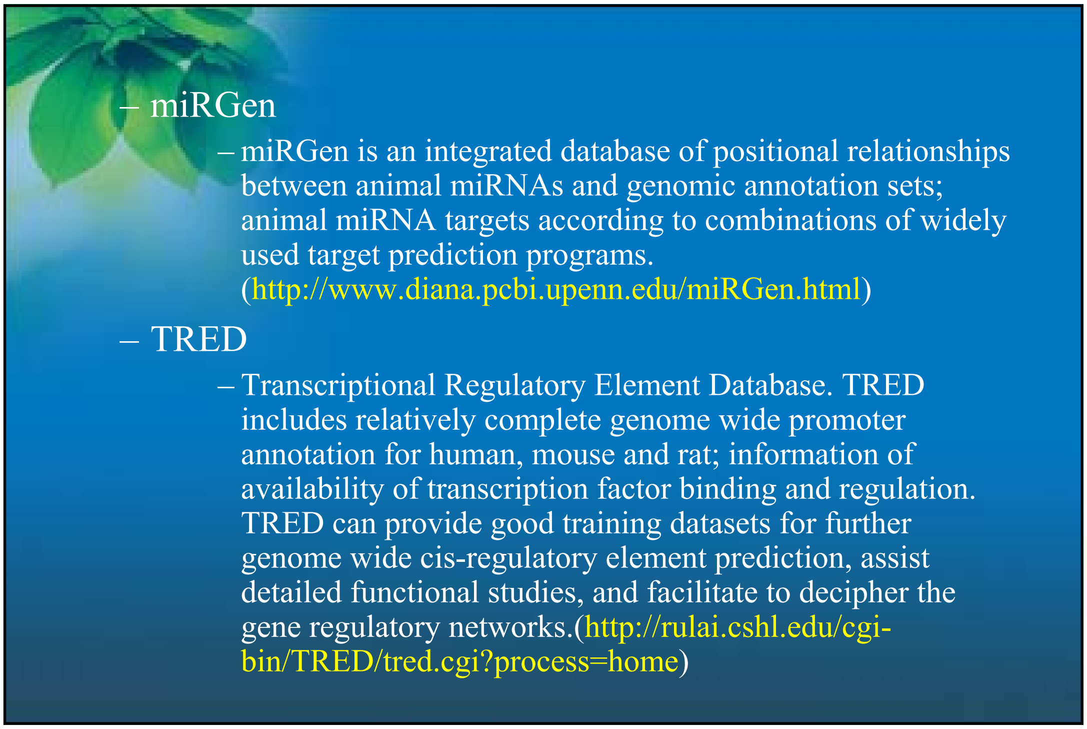
Forward engineering

Reverse engineering



Microarray datasets of microRNAs and genes

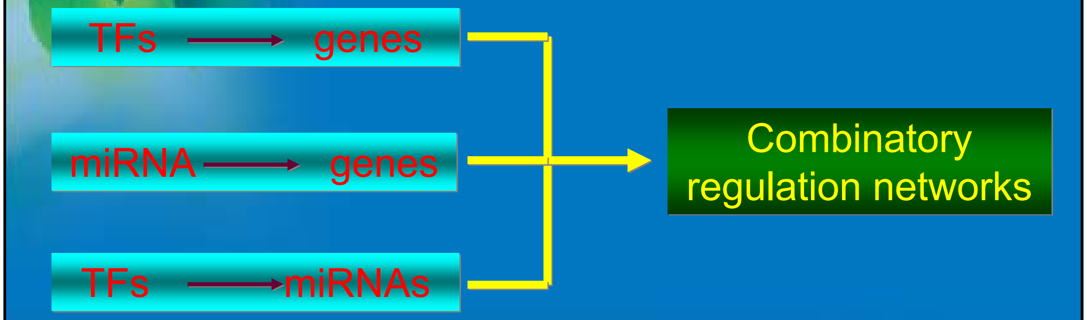
- **Datas source**
  - ***NCI60 2007***. The NCI-60, a panel of 60 diverse human cancer cell lines used by the Developmental Therapeutics Program of the U.S. NCI60 dataset contain tissue specific gene expression data for over 300 miRNAs, 18457 genes from nine cancers with 59 sub-phenotype samples.. (http://discover.nci.nih.gov/cellminer/)
  - Lu et al. MicroRNA expression profiles classify human cancers. ***Nature***. 2005 Jun 9;435:745-6
  - Liu, T.et al., Detection of a microRNA signal in an in vivo expression set of mRNAs, ***PLoS ONE***, 2007, 2, 804.
  - Pablo Landgraf, et al., A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing, ***Cell***, 2007, Volume 129, Issue 7, 1401-1414.
  - ***GNF atlas 2007***:high throughput gene expression atlas of mouse and human expression patterns across diverse tissue.(http://www.gnf.org)



- miRGen
  - miRGen is an integrated database of positional relationships between animal miRNAs and genomic annotation sets; animal miRNA targets according to combinations of widely used target prediction programs. (http://www.diana.pcbi.upenn.edu/miRGen.html)
- TRED
  - Transcriptional Regulatory Element Database. TRED includes relatively complete genome wide promoter annotation for human, mouse and rat; information of availability of transcription factor binding and regulation. TRED can provide good training datasets for further genome wide cis-regulatory element prediction, assist detailed functional studies, and facilitate to decipher the gene regulatory networks.(http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=home)

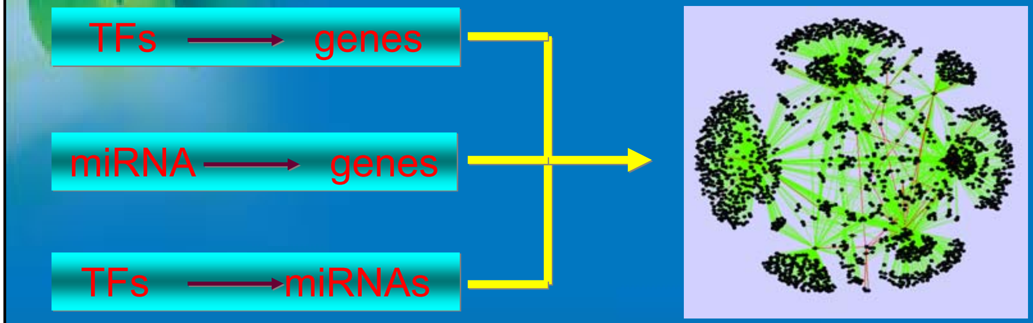
- UCSC hg18 databases
  - The hg18 database contains location information of miRNA on the human genome, and data report page contains links to sequence and annotation data for the genome assemblies featured in the UCSC Genome Browser. (http: // hgdownload.cse. ucsc. edu/downloads.html)
- sanger miRBase
  - miRBase is the new home for microRNA data, containing 3 main sections: all published miRNA sequences, genomic locations and associated annotation; a newly developed database of predicted miRNA target genes; confidential service assigning official names for novel miRNA genes prior to publication of their discovery. (http://microrna.sanger.ac.uk/)

Get regulatory information from data resources



Extract TF-gene pairs, miRNA-gene pairs and TF-miRNA pairs from our selected data resources, then put them as input into our reverse engineering model to get combinatory regulation networks made by TF-gene pairs, miRNA-gene pairs and TF-miRNA pairs. This is our strategy of building miRNA regulatory network based on **forward and reverse engineering**

Get regulatory information from data resources



Extract TF-gene pairs, miRNA-gene pairs and TF-miRNA pairs from our selected data resources, then put them as input into our reverse engineering model to get combinatory regulation networks made by TF-gene pairs, miRNA-gene pairs and TF-miRNA pairs. This is our strategy of building miRNA regulatory network based on **forward and reverse engineering**

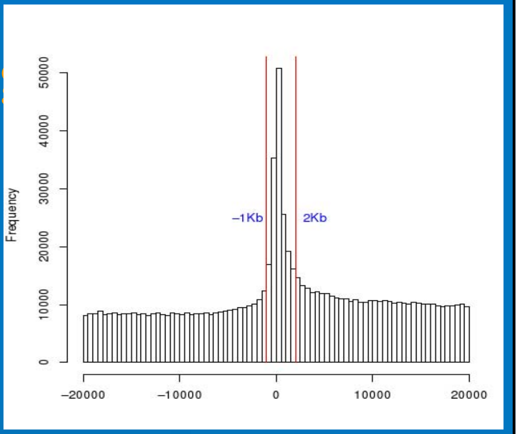
TF -> gene

- Union from
  - UCSC: promoter defined as -1000 ~ +500 bps
  - TRED

Source	transcription factor	target	relationship
ucsc predicted	137	15805	127444
TRED	134	3032	7059
Union	214	16354	130338

TF ->

- ❖ Union from
  - ❖UCSC: promoter defined
  - ❖TRED



Source	transcription factor	target	relationship
ucsc predicted	137	15805	127444
TRED	134	3032	7059
Union	214	16354	130338

microRNA -> target genes

- ❖ Union of
  - PicTar (http://pictar.mdc-berlin.de/)
  - TargetScan (http://www.targetscan.org/)
  - miRanda (http://cbio.mskcc.org/research/sander/data/miRNA2003/miranda_new.html)

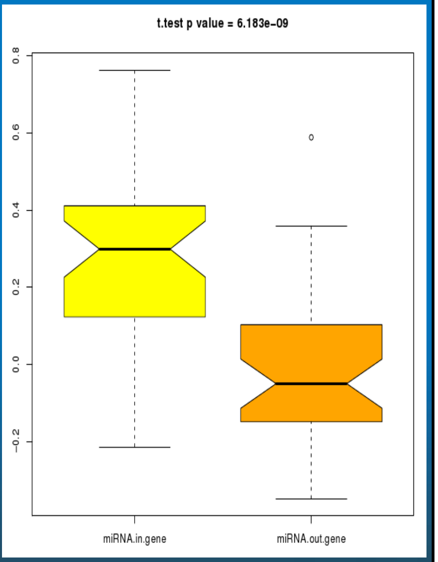
Prediction alg	microRNA	target	relationship
PicTar4way	178	6392	75968
TargetScan	238	7579	75613
miRandaXL	157	5448	41804
Union	276	10255	118408

TF->microRNA

- In order to obtain the information of how a TF regulate miRNAs, we searched UCSC hg18 database again and got total 421 human microRNAs, of which 123 located in genes. (most in intron).
- miRNA and gene Expression data are needed for calculate the correlation of miRNA-gene pairs and were downloaded from CellMiner (Blower, et al., 2007, Mol Cancer Ther ; Shankavaram, et al., 2007, Mol Cancer Ther. (http://discover.nci.nih.gov/cellminer/home.do).
- After carefully selection we got two miRNA and gene Expression datasets, 321 microRNA and 8388 gene are inclusive in these datasets.

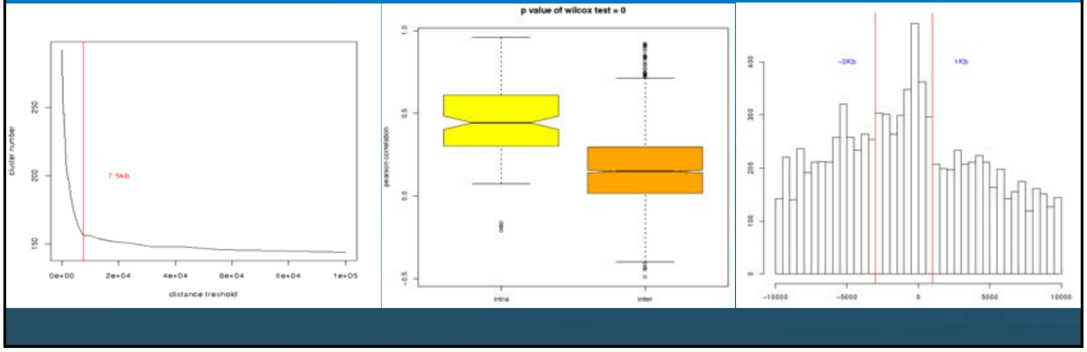
TF->microRNA

Map miRNA onto genome we found 177 of 421 miRNA were located at ORF region, most of them were located at intron region. Through calculation of gene expression correlation based on the related miRNA and gene expression data we found there exists significant positive correlation for the most of these kind of miRNA-gene pairs. A logic conclusion can be reached for this results that these miRNA-gene pairs may share the same regulatory region and mechanisms directly or indirectly. The right figure show the statistical significant of our conclusions.



TF->microRNA

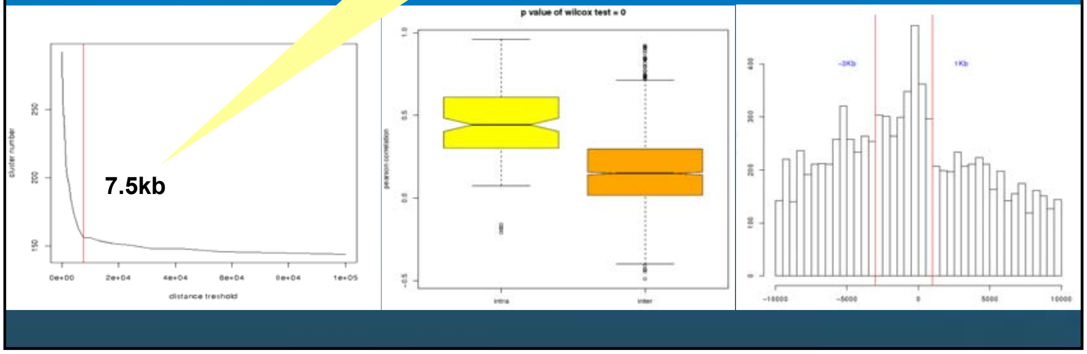
- the rest 298 intergenic microRNAs can be divided into 157 clusters based on a selected distance criteria(7.5kbp) on chromosome region, in which the largest one includes 43 microRNAs

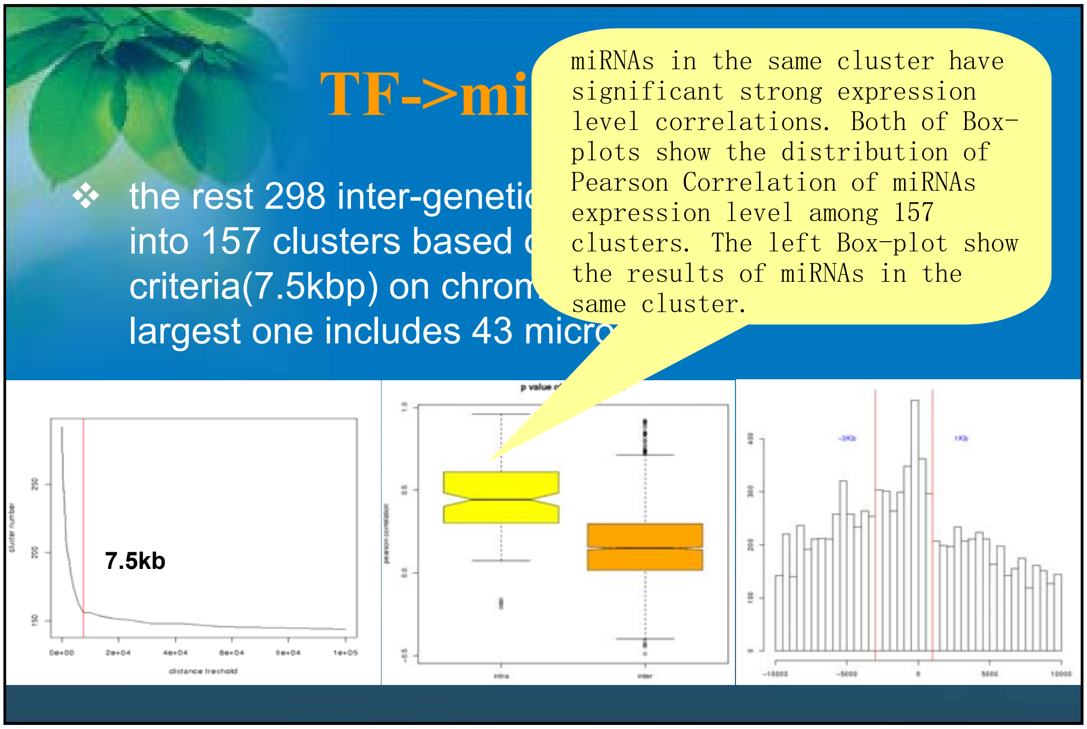


TF->m

- the rest 298 inter-gene ed into 157 clusters base criteria(7.5kbp) on ne largest one incl 43

On chromosome region when the distance of two miRNA is small then a certain criteria they can be regarded as belonging to the same cluster. From this figure we can see when distance is bigger than 7.5kb, the number of resulted clusters closes to a constant. We selected 7.5kb distance as our criteria for miRNA clustering.

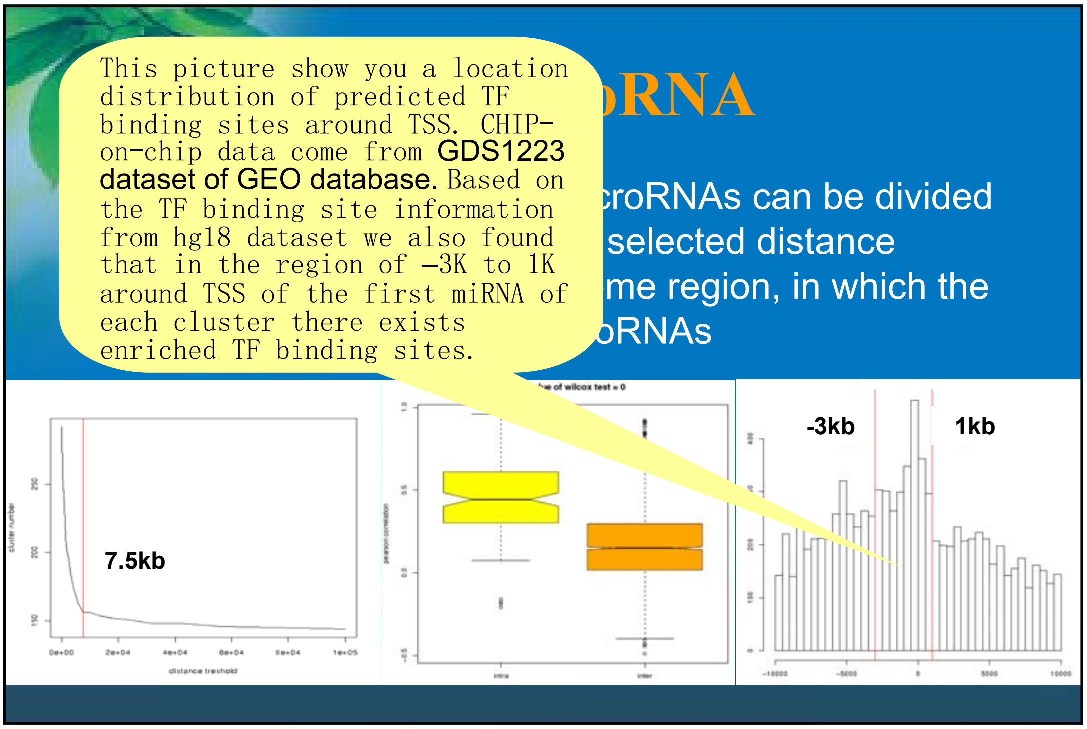




TF->mi

❖ the rest 298 inter-genetic into 157 clusters based criteria(7.5kbp) on chrom largest one includes 43 micro

miRNAs in the same cluster have significant strong expression level correlations. Both of Box-plots show the distribution of Pearson Correlation of miRNAs expression level among 157 clusters. The left Box-plot show the results of miRNAs in the same cluster.



oRNA

This picture show you a location distribution of predicted TF binding sites around TSS. CHIP-on-chip data come from GDS1223 dataset of GEO database. Based on the TF binding site information from hg18 dataset we also found that in the region of –3K to 1K around TSS of the first miRNA of each cluster there exists enriched TF binding sites.

croRNAs can be divided selected distance me region, in which the oRNAs

Linear approximation Hypothesis

Any gene's expression level is mainly controled by related TFs and miRNAs and can be expressed as a union of expression levels of TFs and miRNAs.

⬇

Forward and Reverse Engineering Model based Data Integration

Multiple linear regression model

-Reverse Engineering Model

- ❖**$E_{\_g} = A_{tf\_g} \times E_{tf\_g} + A_{m\_g} \times E_{m\_g}$ + interception + err**
- ❖**$E_{m\_g} = A_{tf\_g}(m) \times E_{tf\_g}(m)$ + interception' + err'**

$E_{\_g}$: Expression level of a gene

E_{m_g}: Expression level of a miRNA

A_{tf_g}: A vector of TFs multi-action to gene g

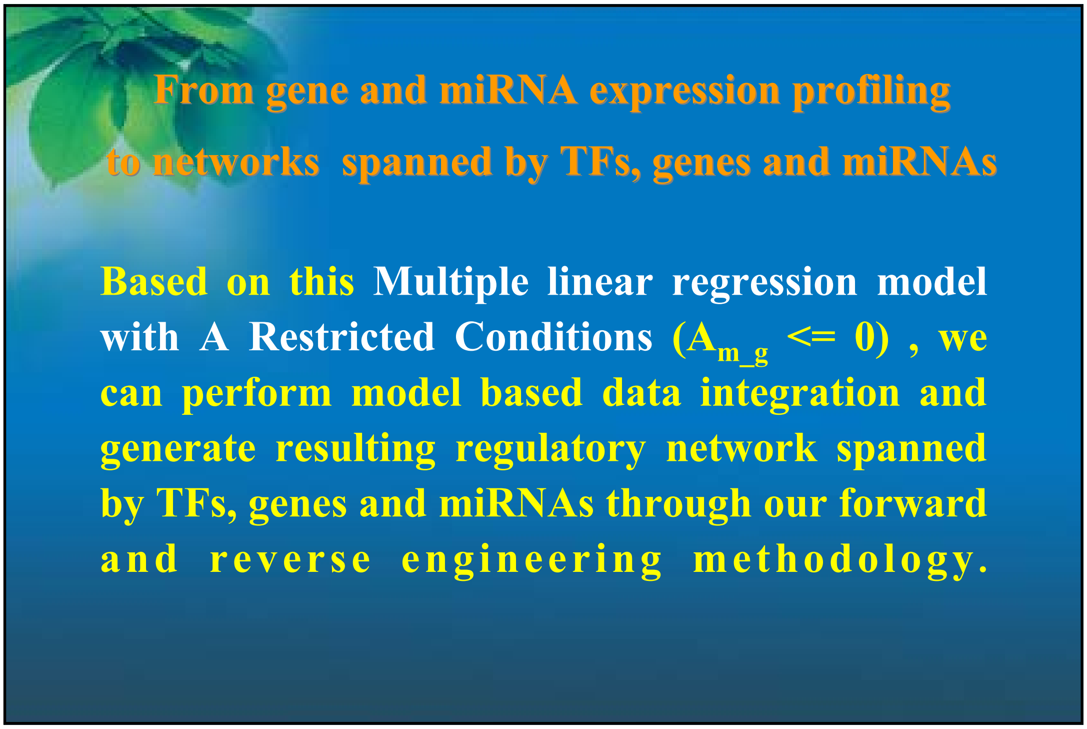
A_{m_g}: A vector of miRNAs multi-action to gene g

$A_{tf\_g}(m)$: A vector of TFs multi-action to miRNA

$E_{tf\_g}(m)$: Expression levels of miRNA related TFs

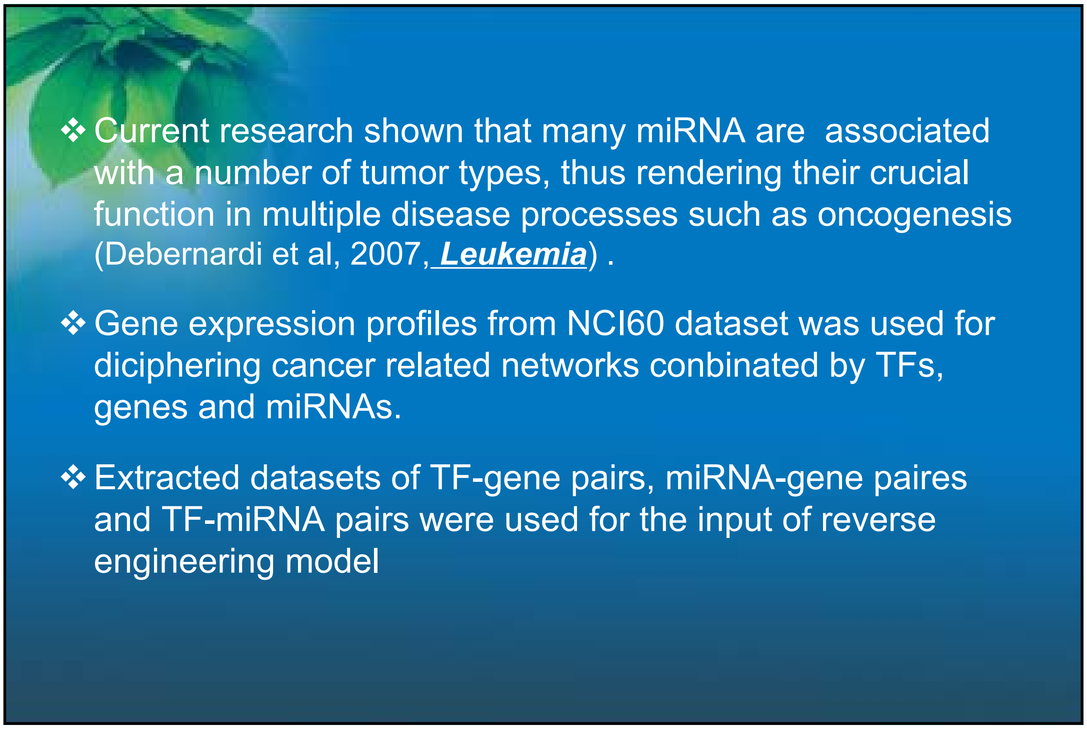
interception: a constant

err: random background from non-TFs factors

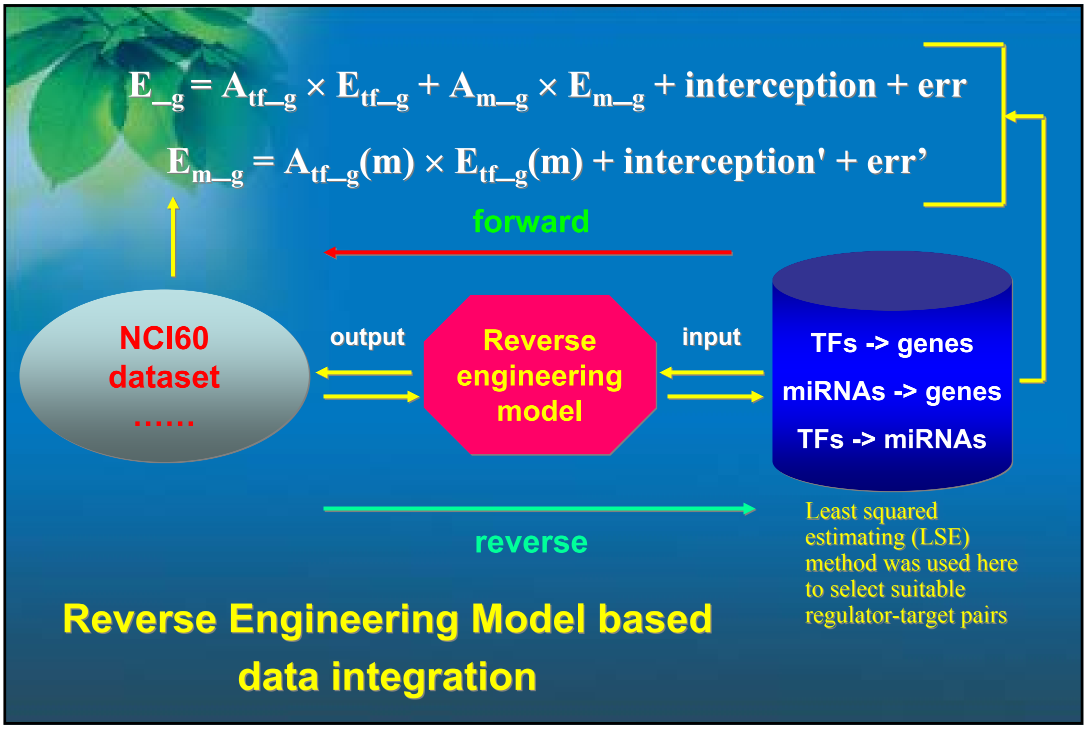


From gene and miRNA expression profiling to networks spanned by TFs, genes and miRNAs

Based on this Multiple linear regression model with A Restricted Conditions (A_{m_g} <= 0) , we can perform model based data integration and generate resulting regulatory network spanned by TFs, genes and miRNAs through our forward and reverse engineering methodology.



- ❖ Current research shown that many miRNA are associated with a number of tumor types, thus rendering their crucial function in multiple disease processes such as oncogenesis (Debernardi et al, 2007, ***Leukemia***) .
- ❖ Gene expression profiles from NCI60 dataset was used for diciphering cancer related networks conbinated by TFs, genes and miRNAs.
- ❖ Extracted datasets of TF-gene pairs, miRNA-gene paires and TF-miRNA pairs were used for the input of reverse engineering model

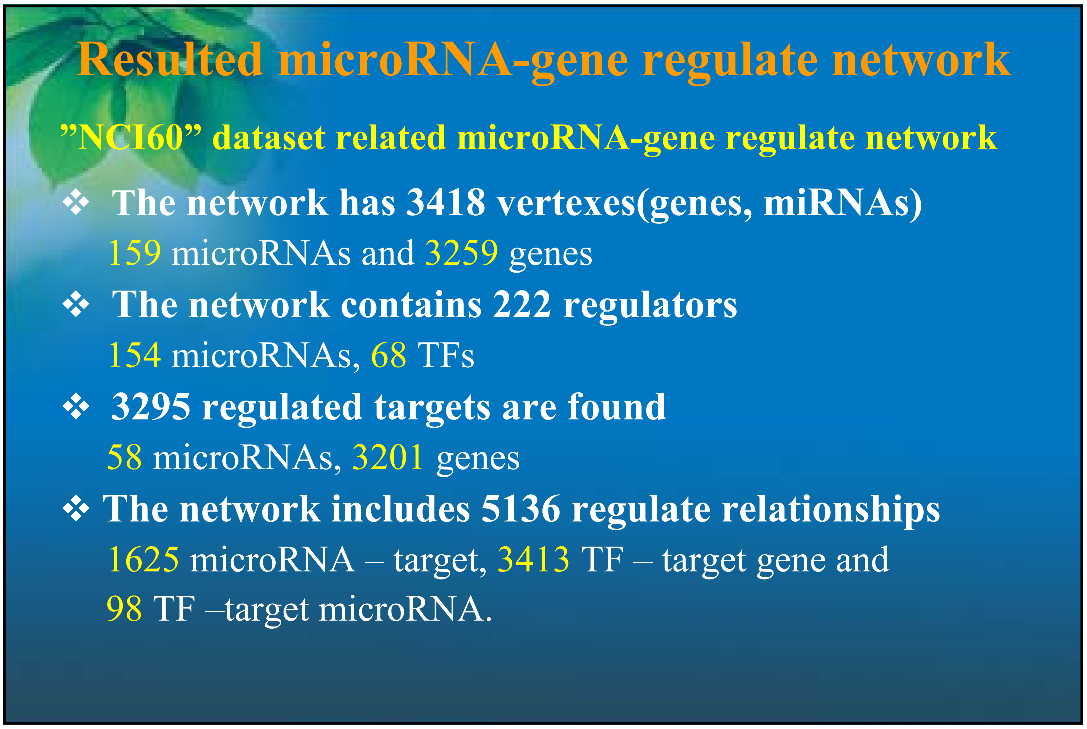


$$E_{\_g} = A_{tf\_g} \times E_{tf\_g} + A_{m\_g} \times E_{m\_g} + \text{interception} + \text{err}$$

$$E_{m\_g} = A_{tf\_g}(m) \times E_{tf\_g}(m) + \text{interception'} + \text{err'}$$

Least squared estimating (LSE) method was used here to select suitable regulator-target pairs

Reverse Engineering Model based data integration



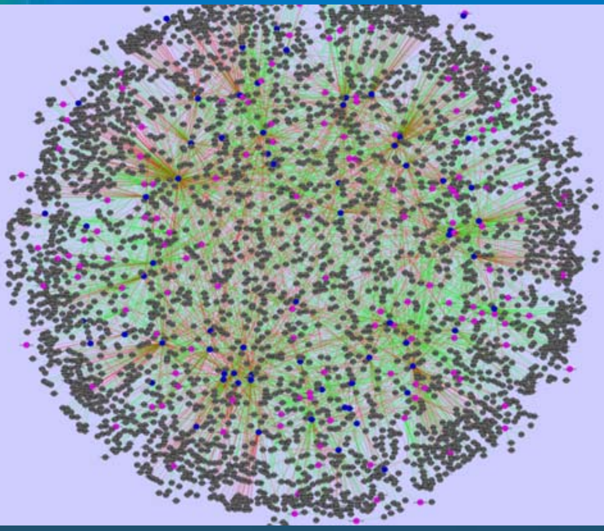
Resulted microRNA-gene regulate network

”NCI60” dataset related microRNA-gene regulate network

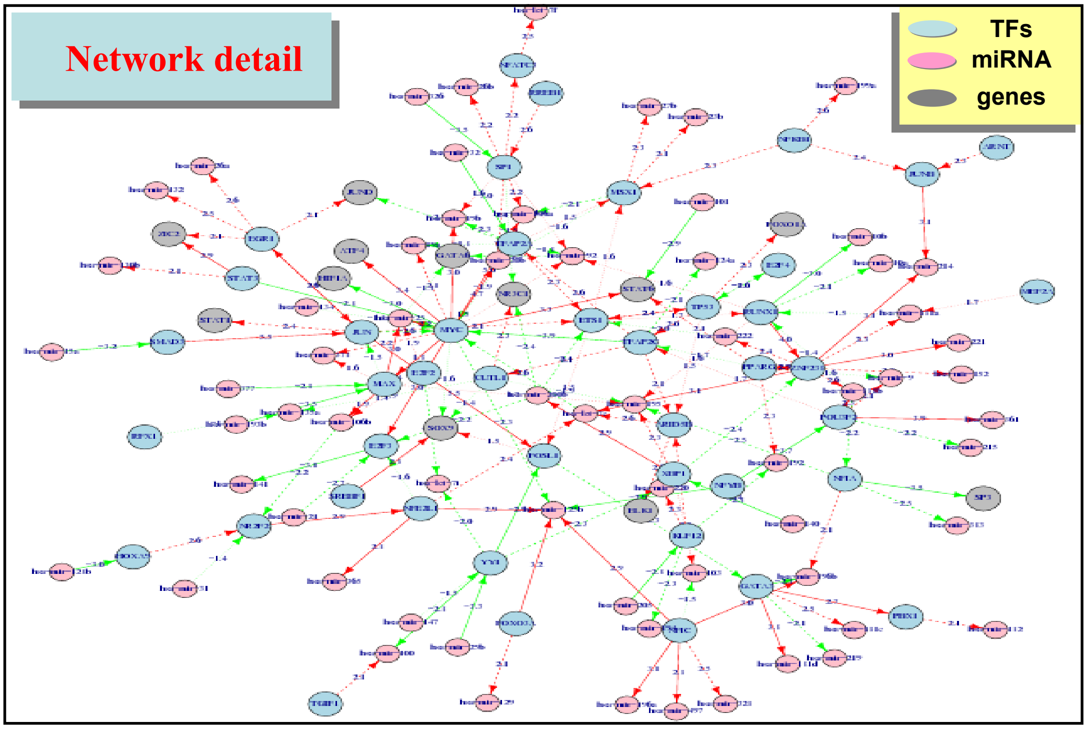
- **The network has 3418 vertexes(genes, miRNAs)**
  159 microRNAs and 3259 genes
- **The network contains 222 regulators**
  154 microRNAs, 68 TFs
- **3295 regulated targets are found**
  58 microRNAs, 3201 genes
- **The network includes 5136 regulate relationships**
  1625 microRNA – target, 3413 TF – target gene and
  98 TF –target microRNA.

- In generated regulatory network 32.3% are microRNA mediated regulations, 67.7% are TF mediated regulations;
- In this case, microRNAs initiated gene regulation is not only a fine spinner to a normal TF mediated regulation but alone a good supplementary to it;
- For cancer related biological processes, it looks like miRNA may play some very crucial role in the formation and development of tumors or oncogenesis.

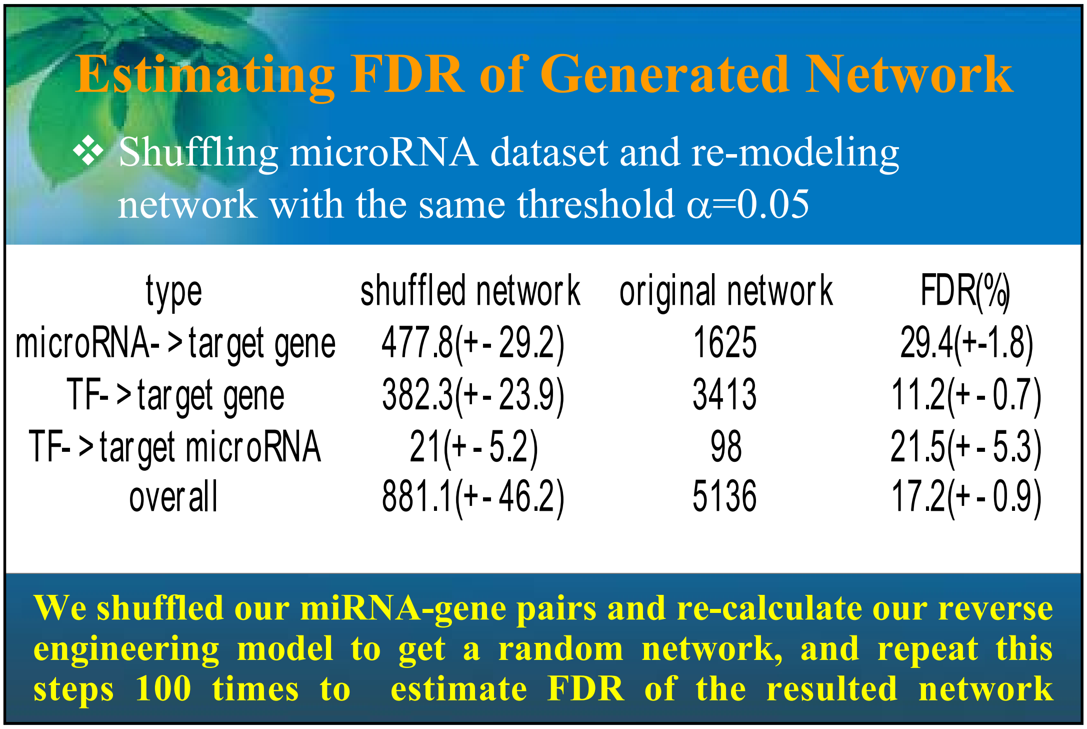
Global Network Spanned by Regulators



Network detail



Validation of generated network

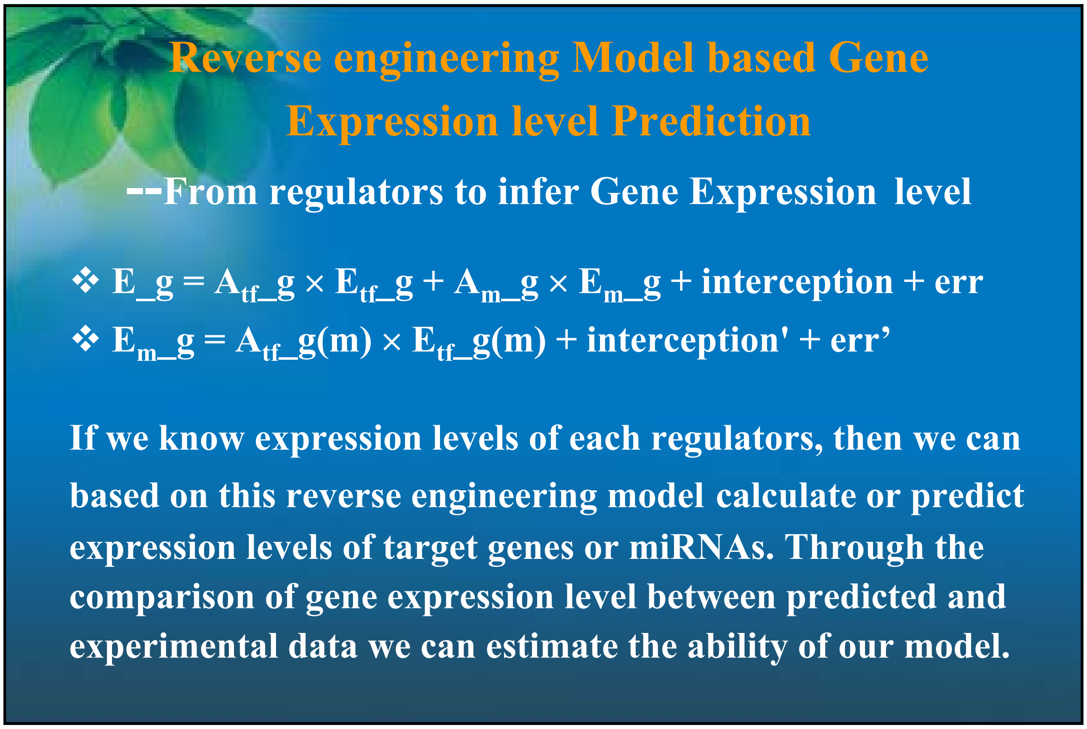


Estimating FDR of Generated Network

- Shuffling microRNA dataset and re-modeling network with the same threshold α=0.05

type	shuffled network	original network	FDR(%)
microRNA->target gene	477.8(+-29.2)	1625	29.4(+-1.8)
TF->target gene	382.3(+-23.9)	3413	11.2(+-0.7)
TF->target microRNA	21(+-5.2)	98	21.5(+-5.3)
overall	881.1(+-46.2)	5136	17.2(+-0.9)

We shuffled our miRNA-gene pairs and re-calculate our reverse engineering model to get a random network, and repeat this steps 100 times to estimate FDR of the resulted network



Reverse engineering Model based Gene Expression level Prediction

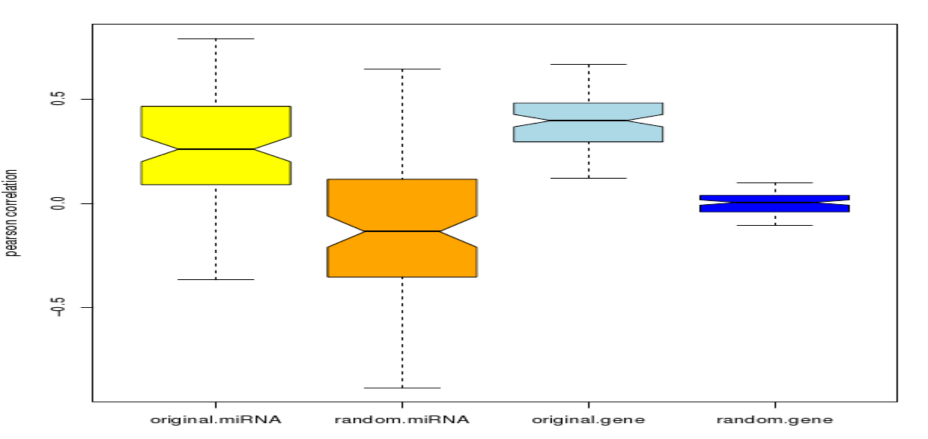
--From regulators to infer Gene Expression level

- $E\_g = A_{tf}\_g \times E_{tf}\_g + A_{m}\_g \times E_{m}\_g$ + interception + err
- $E_{m}\_g = A_{tf}\_g(m) \times E_{tf}\_g(m)$ + interception' + err'

If we know expression levels of each regulators, then we can based on this reverse engineering model calculate or predict expression levels of target genes or miRNAs. Through the comparison of gene expression level between predicted and experimental data we can estimate the ability of our model.

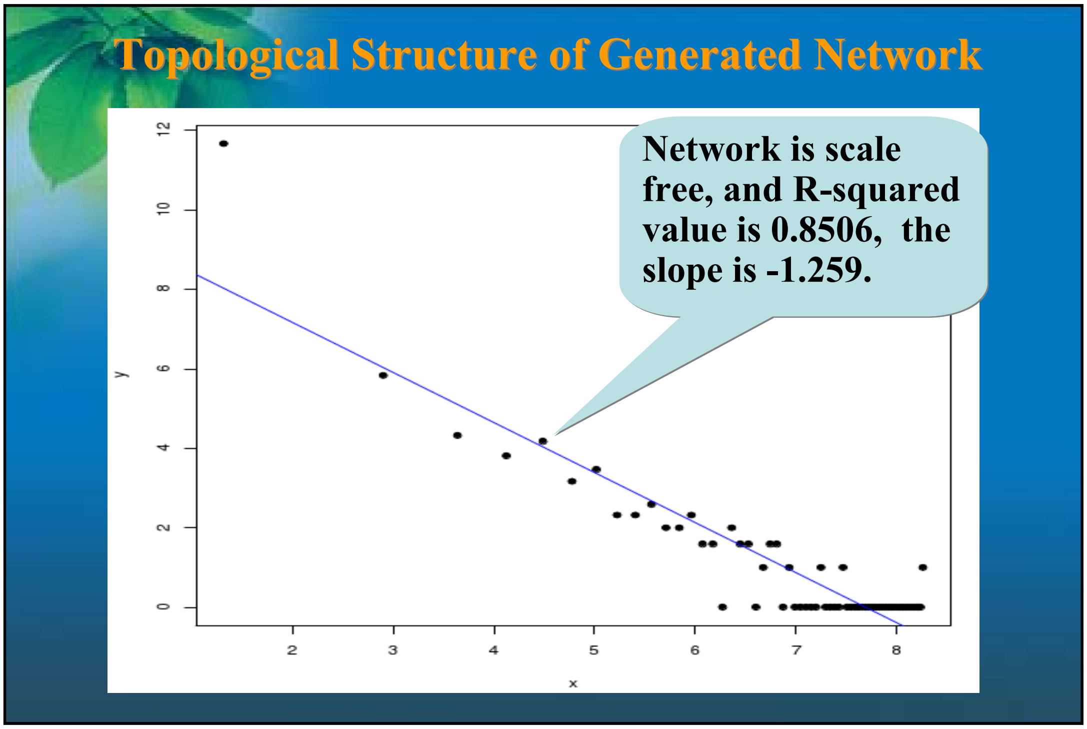
Reverse engineering Model based Gene Expression level Prediction

--From regulators to infer Gene Expression level



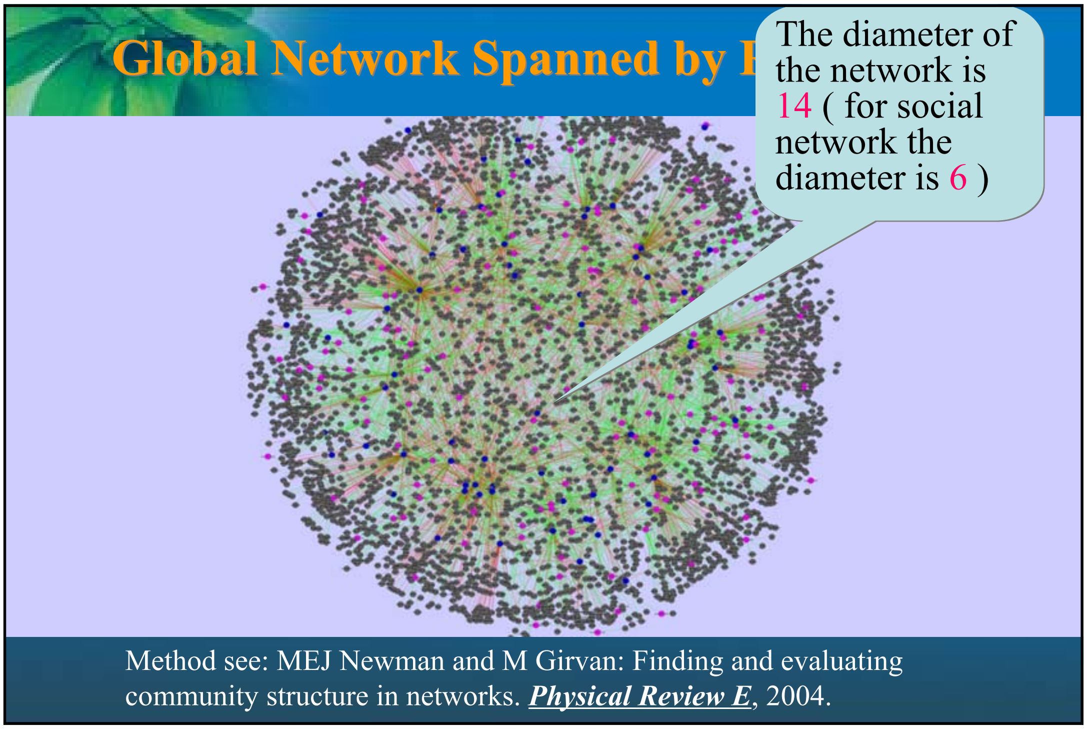
We calculated Pearson correlation coefficients among predicted and experimental expression level data. Results shown there exist high correlations among predicted results and experimental results.

Topological analysis of the network



Topological Structure of Generated Network

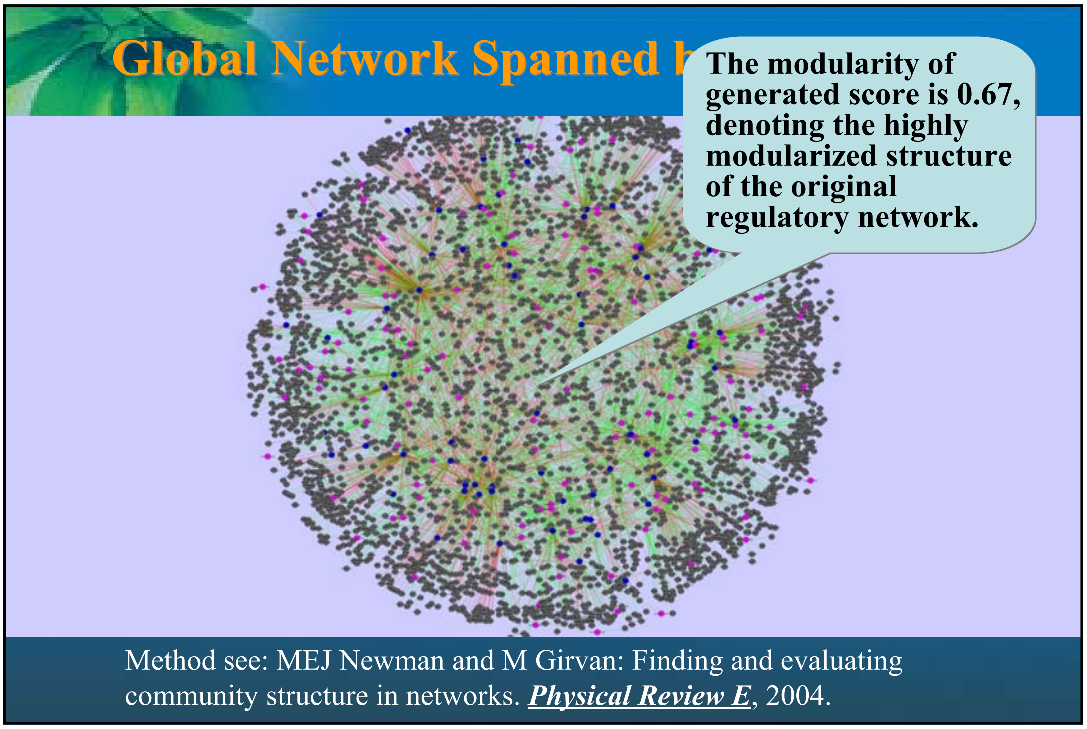
Network is scale free, and R-squared value is 0.8506, the slope is -1.259.



Global Network Spanned by F

The diameter of the network is 14 (for social network the diameter is 6)

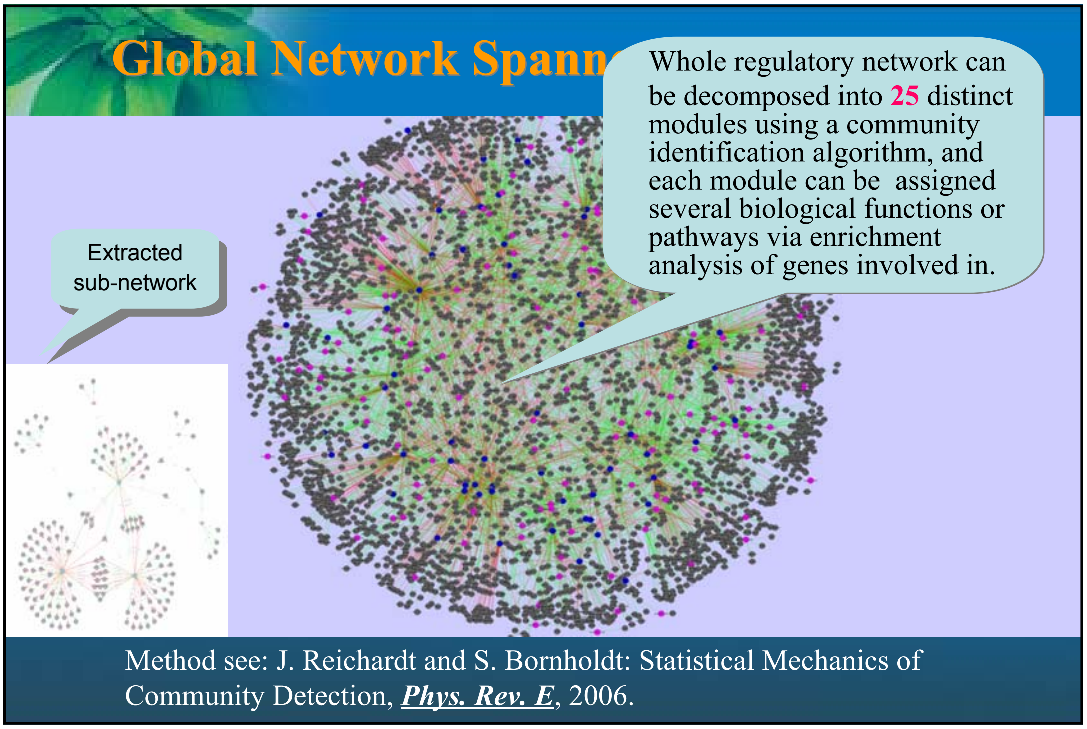
Method see: MEJ Newman and M Girvan: Finding and evaluating community structure in networks. ***Physical Review E***, 2004.



Global Network Spanned b

The modularity of generated score is 0.67, denoting the highly modularized structure of the original regulatory network.

Method see: MEJ Newman and M Girvan: Finding and evaluating community structure in networks. ***Physical Review E***, 2004.

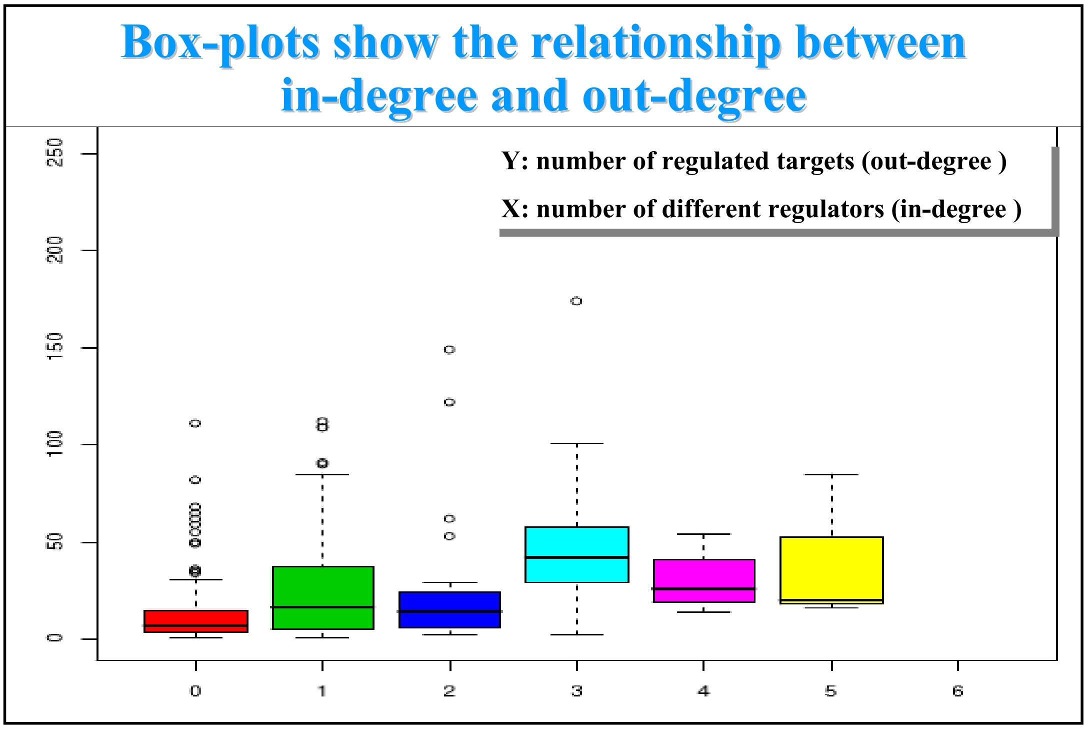


Global Network Spann

Whole regulatory network can be decomposed into **25** distinct modules using a community identification algorithm, and each module can be assigned several biological functions or pathways via enrichment analysis of genes involved in.

Extracted sub-network

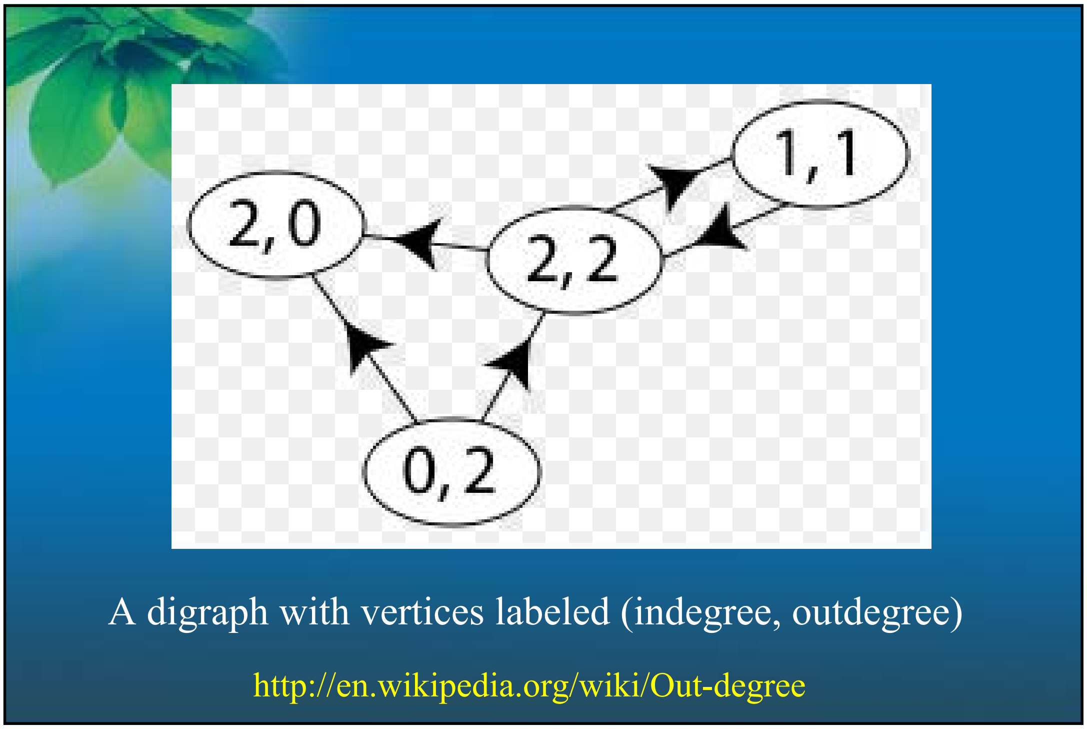
Method see: J. Reichardt and S. Bornholdt: Statistical Mechanics of Community Detection, ***Phys. Rev. E***, 2006.



Box-plots show the relationship between in-degree and out-degree

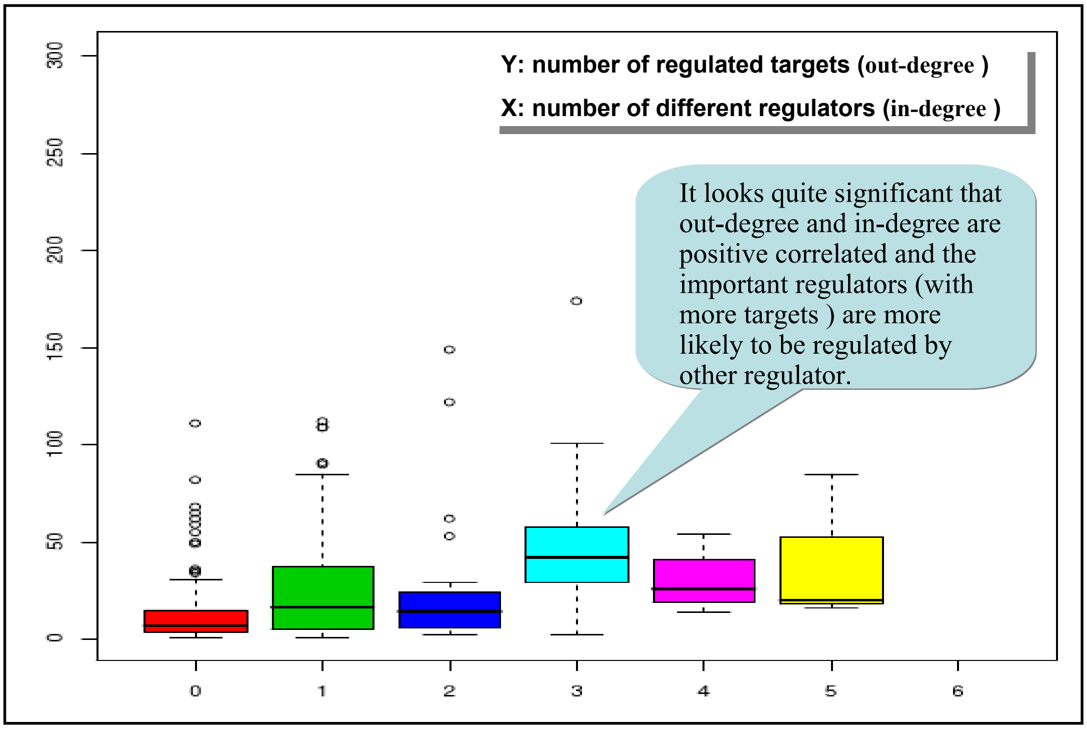
Y: number of regulated targets (out-degree)

X: number of different regulators (in-degree)



A digraph with vertices labeled (indegree, outdegree)

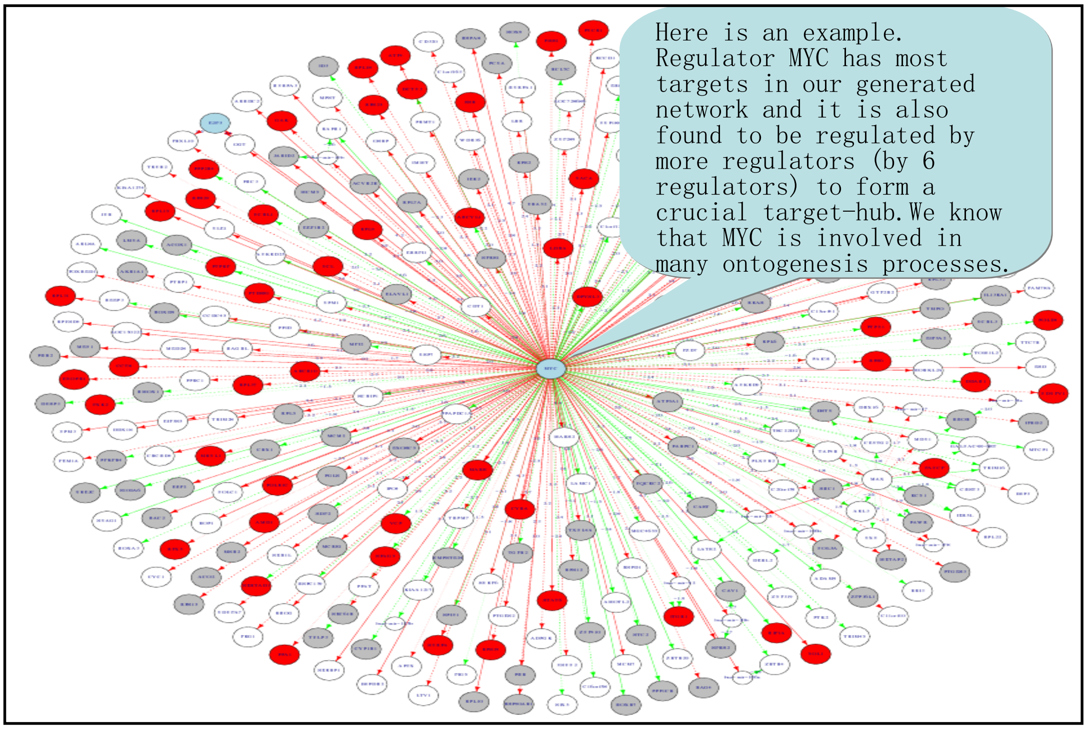
http://en.wikipedia.org/wiki/Out-degree



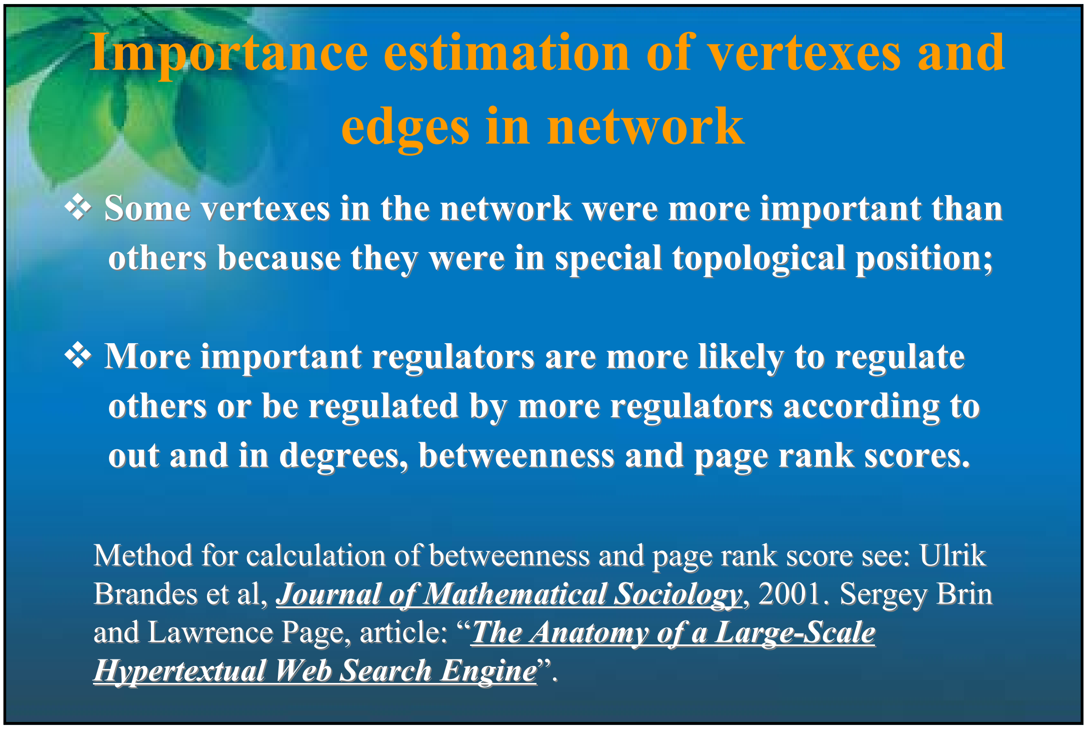
Y: number of regulated targets **(out-degree)**

X: number of different regulators **(in-degree)**

It looks quite significant that out-degree and in-degree are positive correlated and the important regulators (with more targets) are more likely to be regulated by other regulator.



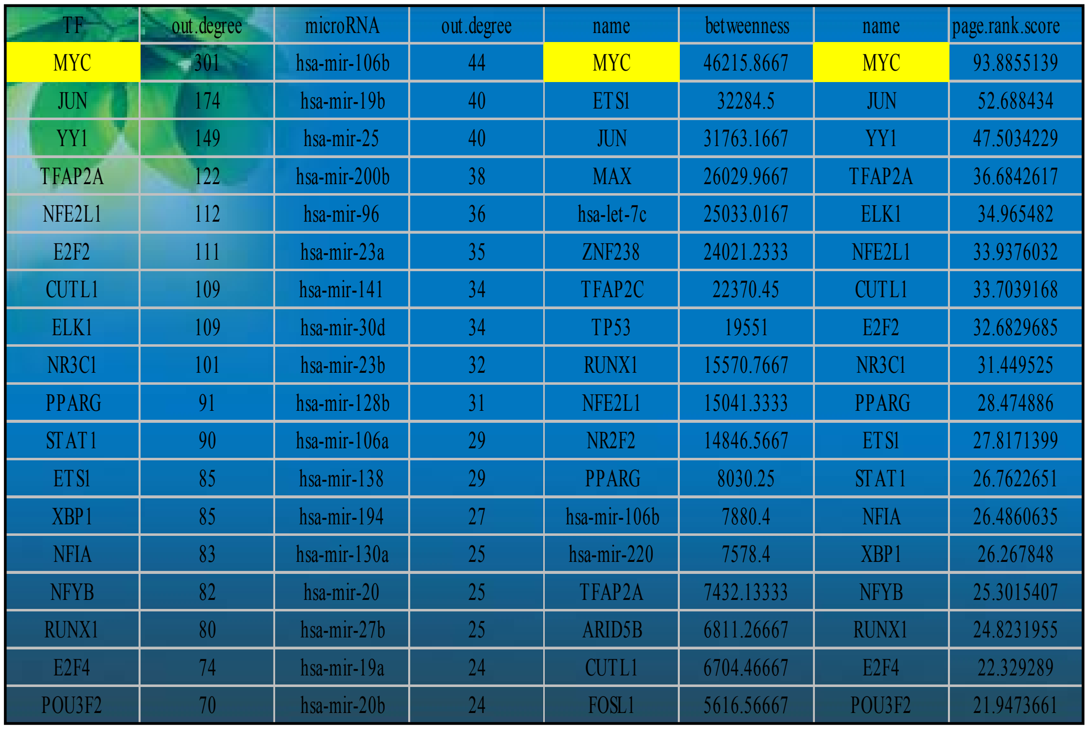
Here is an example. Regulator MYC has most targets in our generated network and it is also found to be regulated by more regulators (by 6 regulators) to form a crucial target-hub. We know that MYC is involved in many ontogenesis processes.



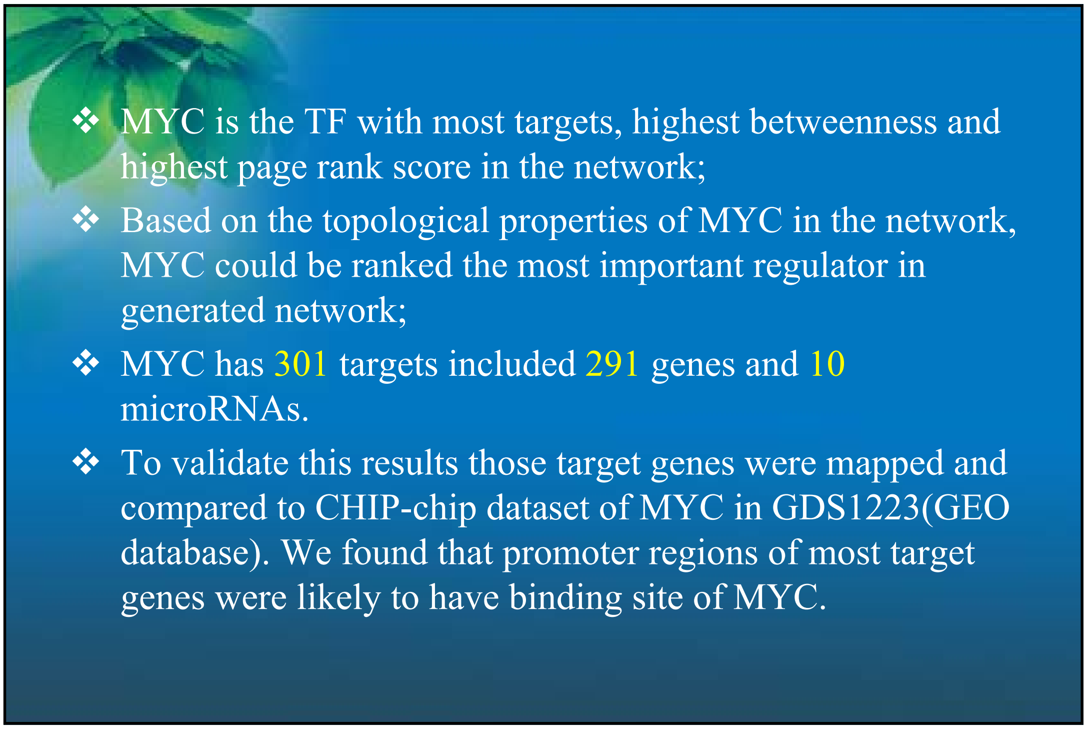
Importance estimation of vertexes and edges in network

- ❖ Some vertexes in the network were more important than others because they were in special topological position;
- ❖ More important regulators are more likely to regulate others or be regulated by more regulators according to out and in degrees, betweenness and page rank scores.

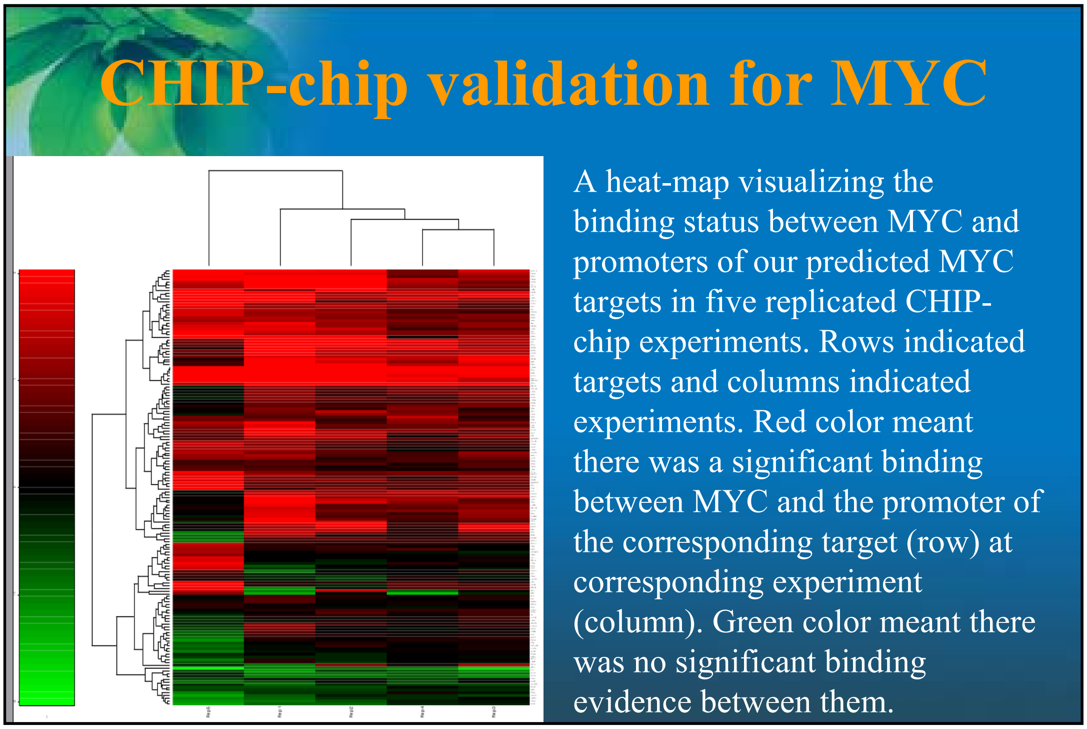
Method for calculation of betweenness and page rank score see: Ulrik Brandes et al, ***Journal of Mathematical Sociology***, 2001. Sergey Brin and Lawrence Page, article: "***The Anatomy of a Large-Scale Hypertextual Web Search Engine***".



TF	out.degree	microRNA	out.degree	name	betweenness	name	page.rank.score
MYC	301	hsa-mir-106b	44	MYC	46215.8667	MYC	93.8855139
JUN	174	hsa-mir-19b	40	ETS1	32284.5	JUN	52.688434
YY1	149	hsa-mir-25	40	JUN	31763.1667	YY1	47.5034229
TFAP2A	122	hsa-mir-200b	38	MAX	26029.9667	TFAP2A	36.6842617
NFE2L1	112	hsa-mir-96	36	hsa-let-7c	25033.0167	ELK1	34.965482
E2F2	111	hsa-mir-23a	35	ZNF238	24021.2333	NFE2L1	33.9376032
CUTL1	109	hsa-mir-141	34	TFAP2C	22370.45	CUTL1	33.7039168
ELK1	109	hsa-mir-30d	34	TP53	19551	E2F2	32.6829685
NR3C1	101	hsa-mir-23b	32	RUNX1	15570.7667	NR3C1	31.449525
PPARG	91	hsa-mir-128b	31	NFE2L1	15041.3333	PPARG	28.474886
STAT1	90	hsa-mir-106a	29	NR2F2	14846.5667	ETS1	27.8171399
ETS1	85	hsa-mir-138	29	PPARG	8030.25	STAT1	26.7622651
XBP1	85	hsa-mir-194	27	hsa-mir-106b	7880.4	NFIA	26.4860635
NFIA	83	hsa-mir-130a	25	hsa-mir-220	7578.4	XBP1	26.267848
NFYB	82	hsa-mir-20	25	TFAP2A	7432.13333	NFYB	25.3015407
RUNX1	80	hsa-mir-27b	25	ARID5B	6811.26667	RUNX1	24.8231955
E2F4	74	hsa-mir-19a	24	CUTL1	6704.46667	E2F4	22.329289
POU3F2	70	hsa-mir-20b	24	FOSL1	5616.56667	POU3F2	21.9473661



- ❖ MYC is the TF with most targets, highest betweenness and highest page rank score in the network;
- ❖ Based on the topological properties of MYC in the network, MYC could be ranked the most important regulator in generated network;
- ❖ MYC has 301 targets included 291 genes and 10 microRNAs.
- ❖ To validate this results those target genes were mapped and compared to CHIP-chip dataset of MYC in GDS1223(GEO database). We found that promoter regions of most target genes were likely to have binding site of MYC.



CHIP-chip validation for MYC

A heat-map visualizing the binding status between MYC and promoters of our predicted MYC targets in five replicated CHIP-chip experiments. Rows indicated targets and columns indicated experiments. Red color meant there was a significant binding between MYC and the promoter of the corresponding target (row) at corresponding experiment (column). Green color meant there was no significant binding evidence between them.

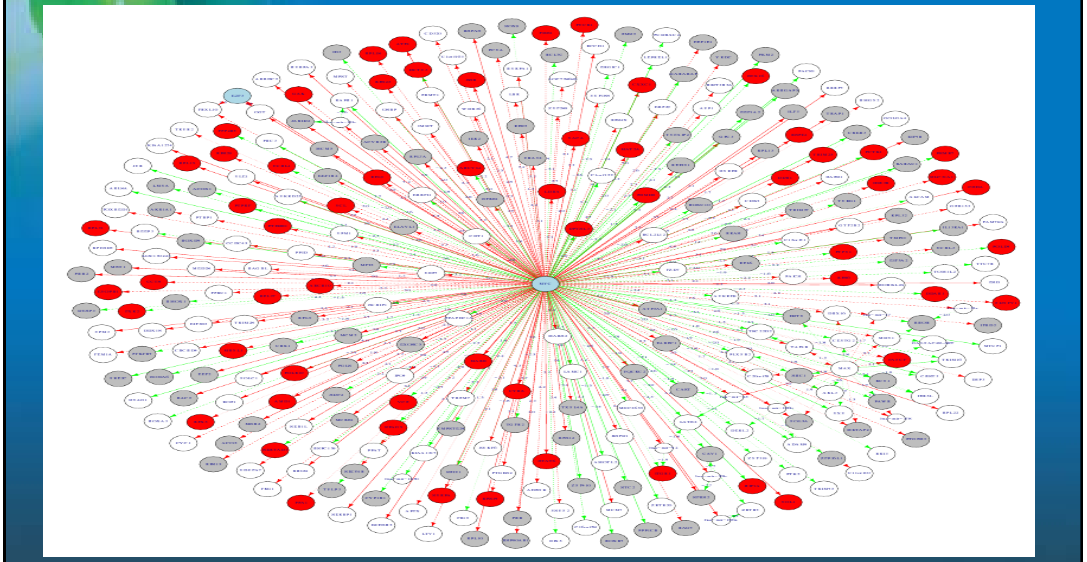
CHIP-chip validation for MYC

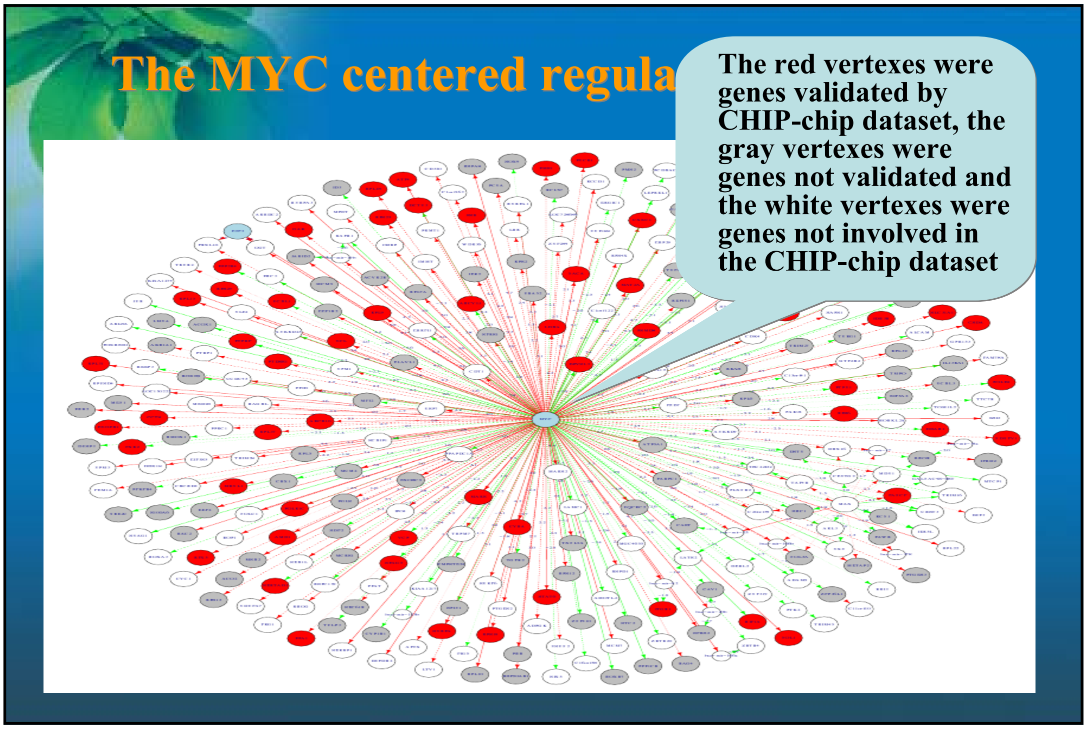
Prediction	Binding	Non-binding
target	60	231
non-target	386	7711

RTDR = 87%

RTDR: Relative True Discovery Rate

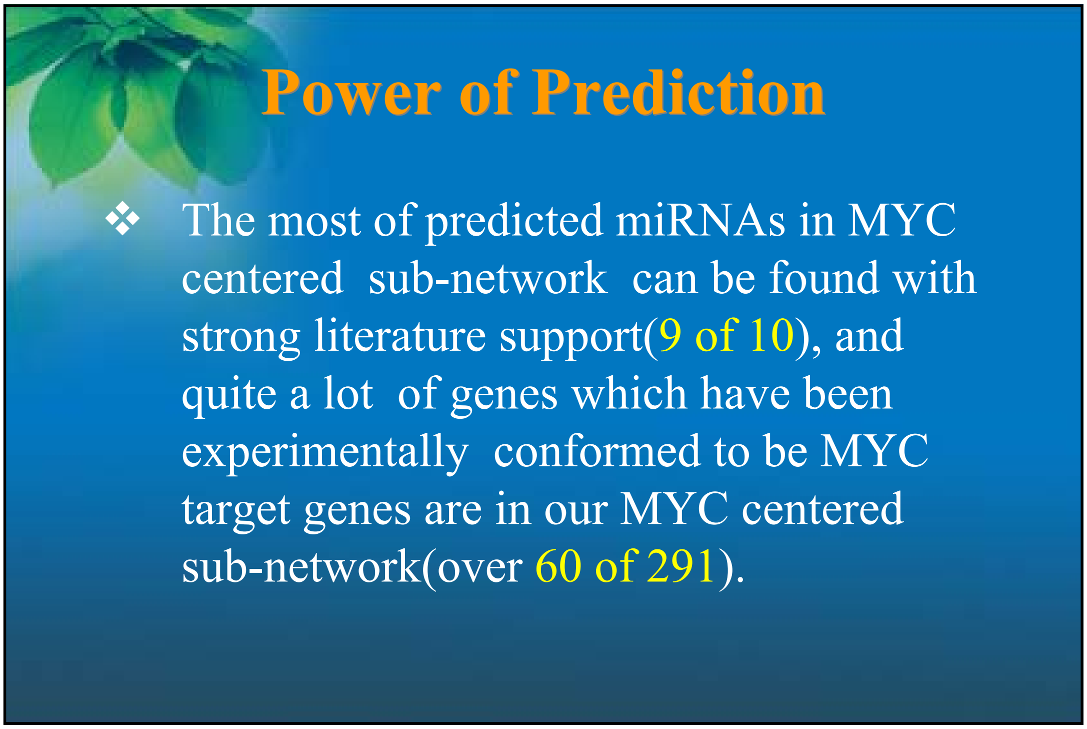
The MYC centered regulatory network





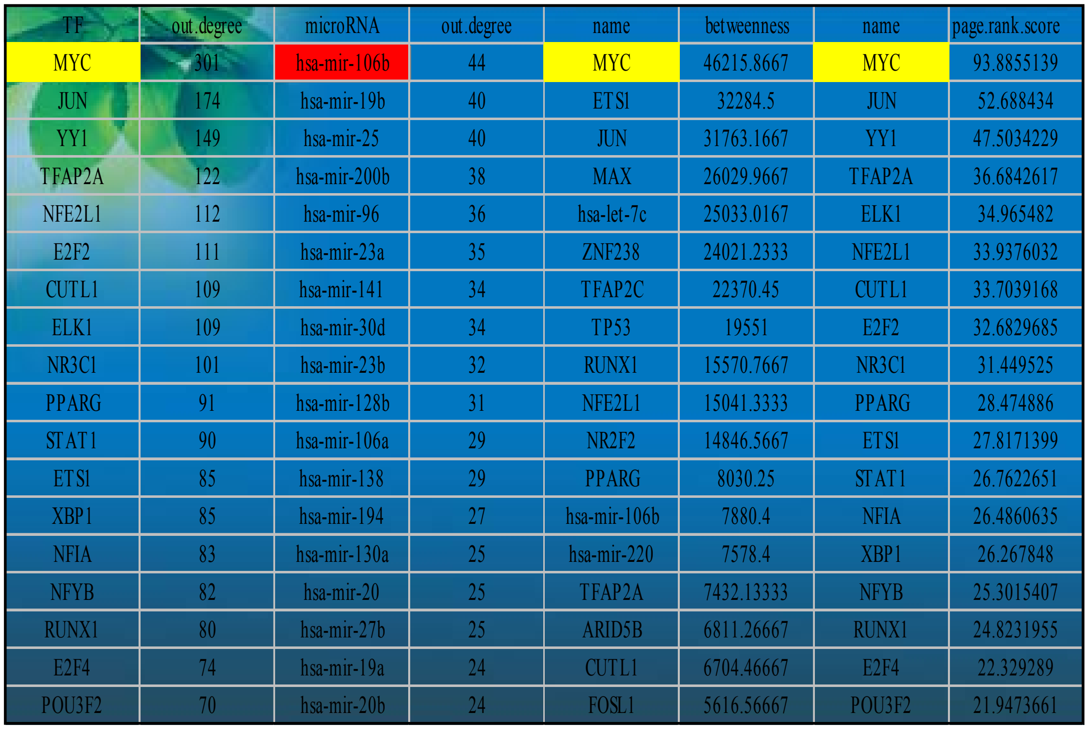
The MYC centered regula

The red vertexes were genes validated by CHIP-chip dataset, the gray vertexes were genes not validated and the white vertexes were genes not involved in the CHIP-chip dataset

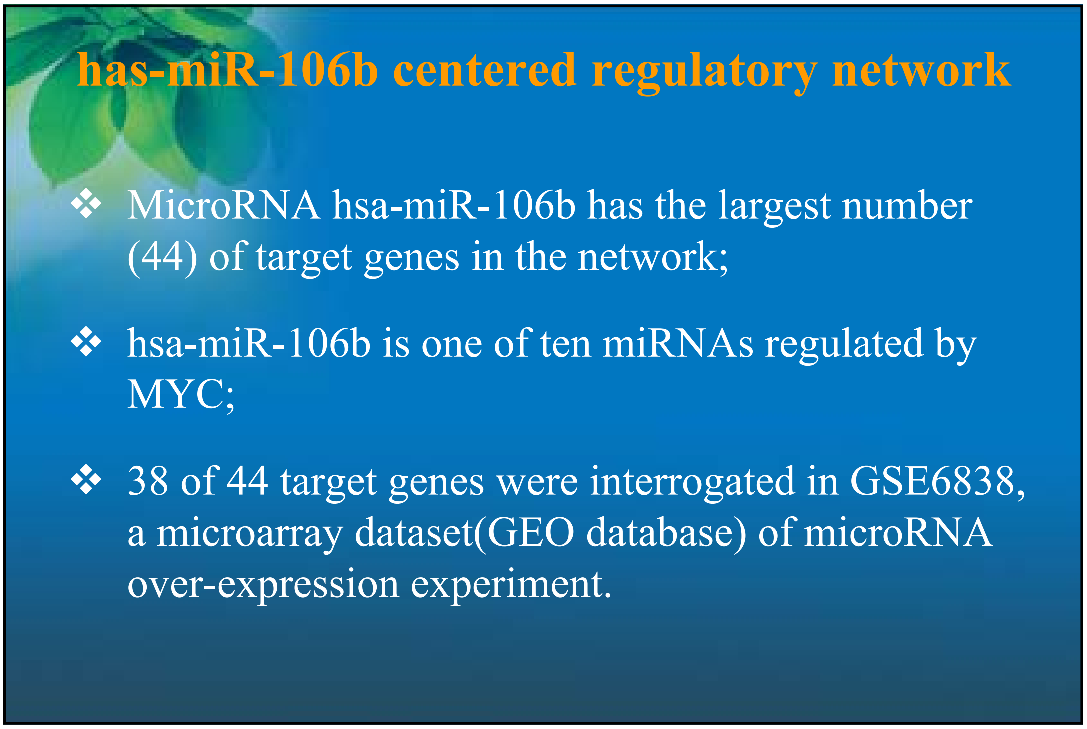


Power of Prediction

- The most of predicted miRNAs in MYC centered sub-network can be found with strong literature support(9 of 10), and quite a lot of genes which have been experimentally conformed to be MYC target genes are in our MYC centered sub-network(over 60 of 291).

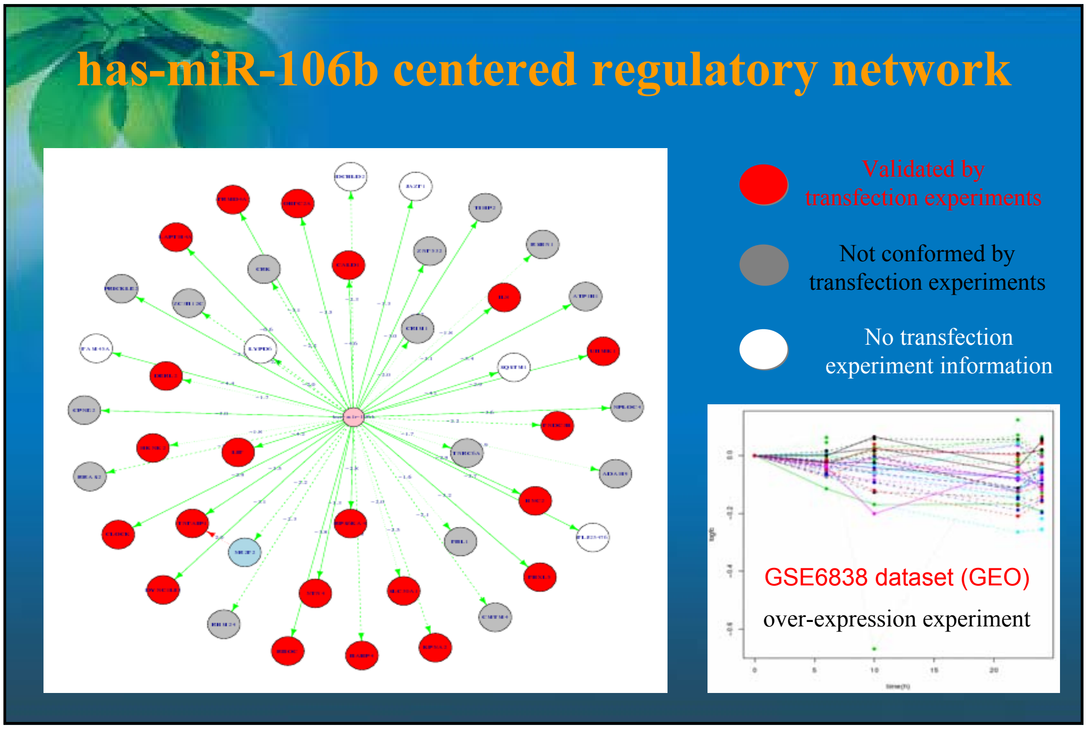


TF	out.degree	microRNA	out.degree	name	betweenness	name	page.rank.score
MYC	301	hsa-mir-106b	44	MYC	46215.8667	MYC	93.8855139
JUN	174	hsa-mir-19b	40	ETSl	32284.5	JUN	52.688434
YY1	149	hsa-mir-25	40	JUN	31763.1667	YY1	47.5034229
TFAP2A	122	hsa-mir-200b	38	MAX	26029.9667	TFAP2A	36.6842617
NFE2L1	112	hsa-mir-96	36	hsa-let-7c	25033.0167	ELK1	34.965482
E2F2	111	hsa-mir-23a	35	ZNF238	24021.2333	NFE2L1	33.9376032
CUTL1	109	hsa-mir-141	34	TFAP2C	22370.45	CUTL1	33.7039168
ELK1	109	hsa-mir-30d	34	TP53	19551	E2F2	32.6829685
NR3C1	101	hsa-mir-23b	32	RUNX1	15570.7667	NR3C1	31.449525
PPARG	91	hsa-mir-128b	31	NFE2L1	15041.3333	PPARG	28.474886
STAT1	90	hsa-mir-106a	29	NR2F2	14846.5667	ETSl	27.8171399
ETSl	85	hsa-mir-138	29	PPARG	8030.25	STAT1	26.7622651
XBP1	85	hsa-mir-194	27	hsa-mir-106b	7880.4	NFIA	26.4860635
NFIA	83	hsa-mir-130a	25	hsa-mir-220	7578.4	XBP1	26.267848
NFYB	82	hsa-mir-20	25	TFAP2A	7432.13333	NFYB	25.3015407
RUNX1	80	hsa-mir-27b	25	ARID5B	6811.26667	RUNX1	24.8231955
E2F4	74	hsa-mir-19a	24	CUTL1	6704.46667	E2F4	22.329289
POU3F2	70	hsa-mir-20b	24	FOSL1	5616.56667	POU3F2	21.9473661



has-miR-106b centered regulatory network

- MicroRNA hsa-miR-106b has the largest number (44) of target genes in the network;
- hsa-miR-106b is one of ten miRNAs regulated by MYC;
- 38 of 44 target genes were interrogated in GSE6838, a microarray dataset(GEO database) of microRNA over-expression experiment.



has-miR-106b centered regulatory network

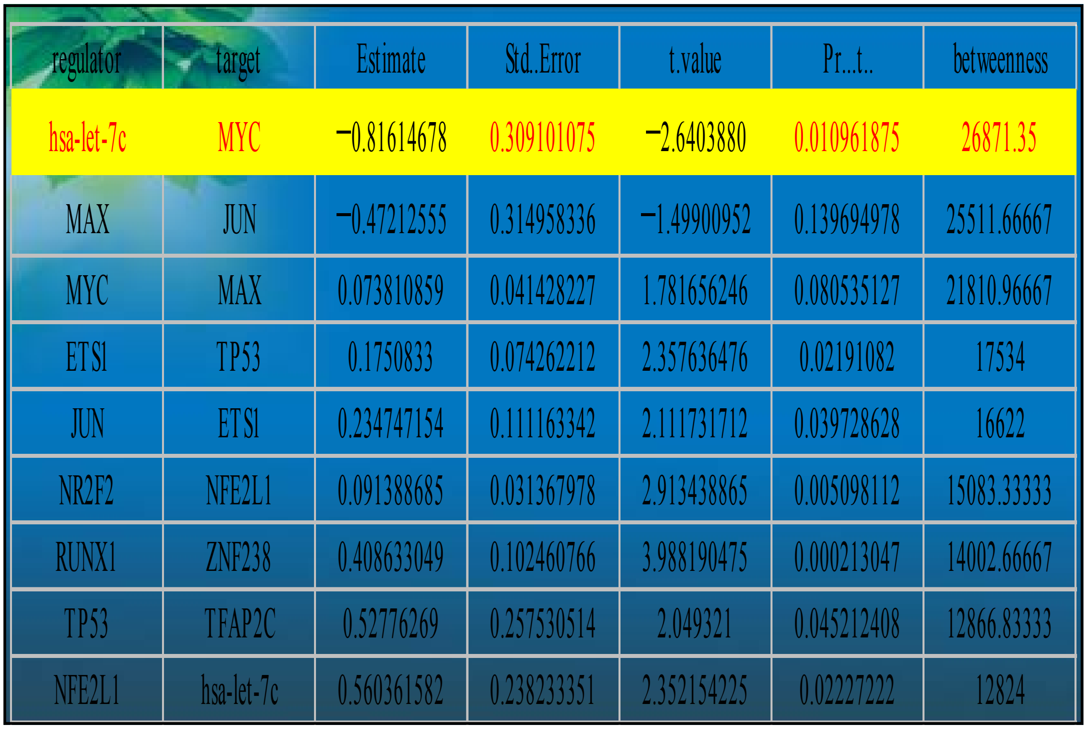
● Validated by transfection experiments

● Not conformed by transfection experiments

○ No transfection experiment information

GSE6838 dataset (GEO)

over-expression experiment



regulator	target	Estimate	Std..Error	t.value	Pr...t..	betweenness
hsa-let-7c	MYC	−0.81614678	0.309101075	−2.6403880	0.010961875	26871.35
MAX	JUN	−0.47212555	0.314958336	−1.49900952	0.139694978	25511.66667
MYC	MAX	0.073810859	0.041428227	1.781656246	0.080535127	21810.96667
ETS1	TP53	0.1750833	0.074262212	2.357636476	0.02191082	17534
JUN	ETS1	0.234747154	0.111163342	2.111731712	0.039728628	16622
NR2F2	NFE2L1	0.091388685	0.031367978	2.913438865	0.005098112	15083.33333
RUNX1	ZNF238	0.408633049	0.102460766	3.988190475	0.000213047	14002.66667
TP53	TFAP2C	0.52776269	0.257530514	2.049321	0.045212408	12866.83333
NFE2L1	hsa-let-7c	0.560361582	0.238233351	2.352154225	0.02227222	12824

miRNA with the highest betweenness as a regulator of MYC

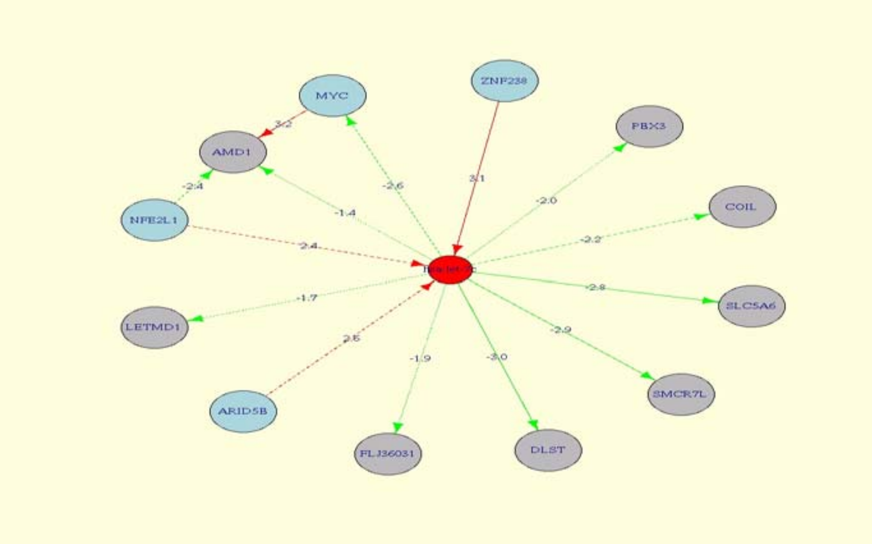
- ❖ Experiment shown that has-let-7c regulates MYC;

  (Yatrik M. Shah et al., ***Molecular and Cellular Biology***, 2007)

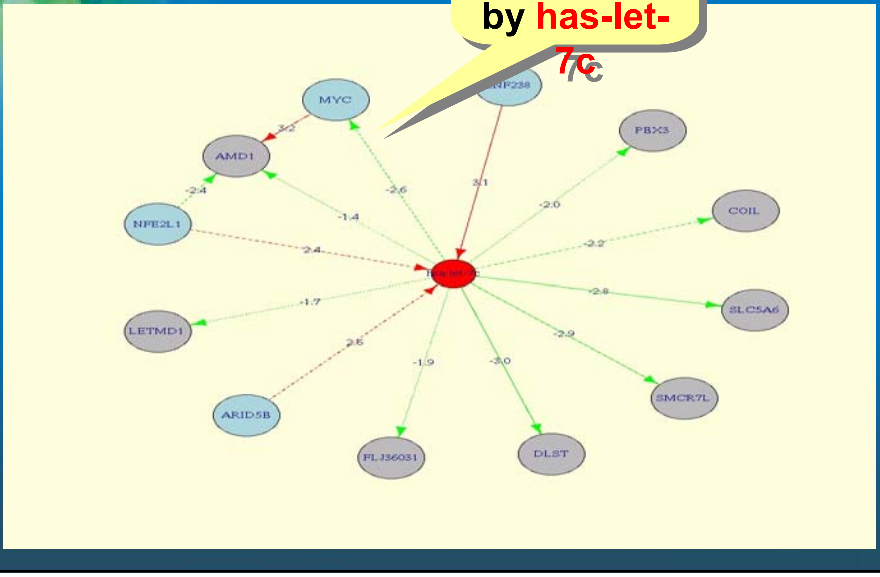
- ❖ Our predicted model is consistent with experiment results, has-let-7c regulates MYC in the network. Besides, has-let-7c has the highest betweenness in generated network, which means has-let-7c will possibly have more functional connections with other genes in the network than that of those genes with lower betweenness;

- ❖ Topological structure properties may also help to decipher the importance of biological functions.

Sub-network centered by has-let-7c



Sub-network cente s-let-7c

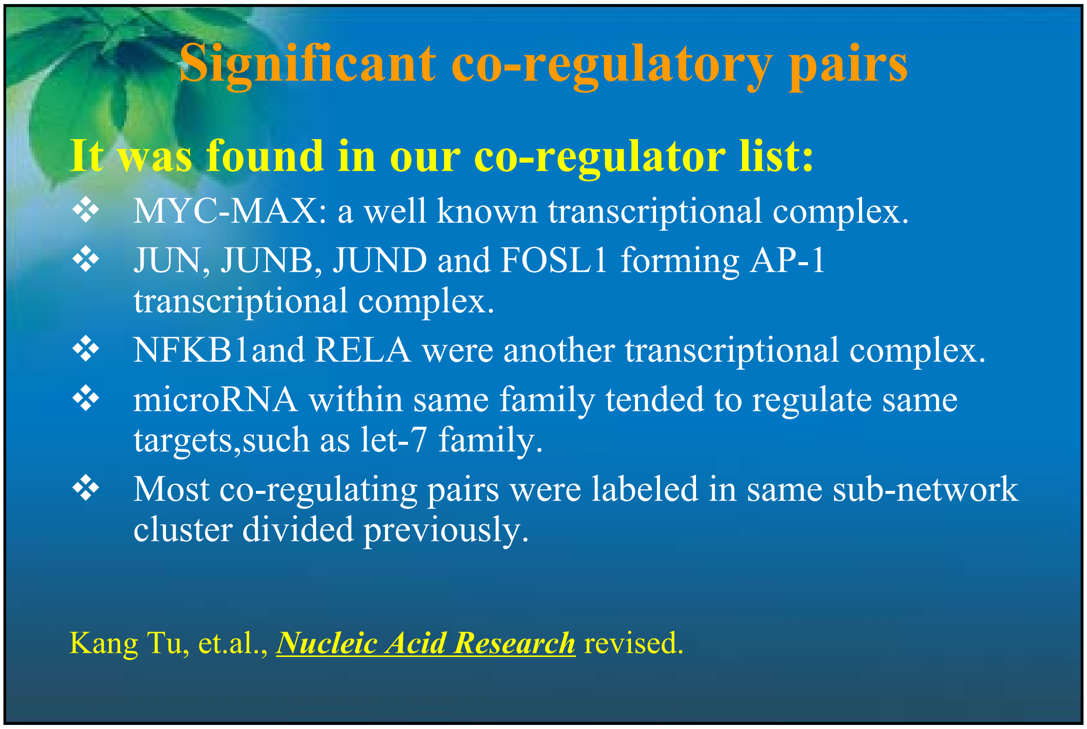


MYC regulated by has-let-7c7c

Significant co-regulating regulator pairs

If two regulators (TFs, miRNAs) significantly share more targets in the generated network, it can be considered a co-regulator pair.

- 17 TF-TF pairs
- 21 microRNA-microRNA pairs
- 7 TF-microRNA pairs

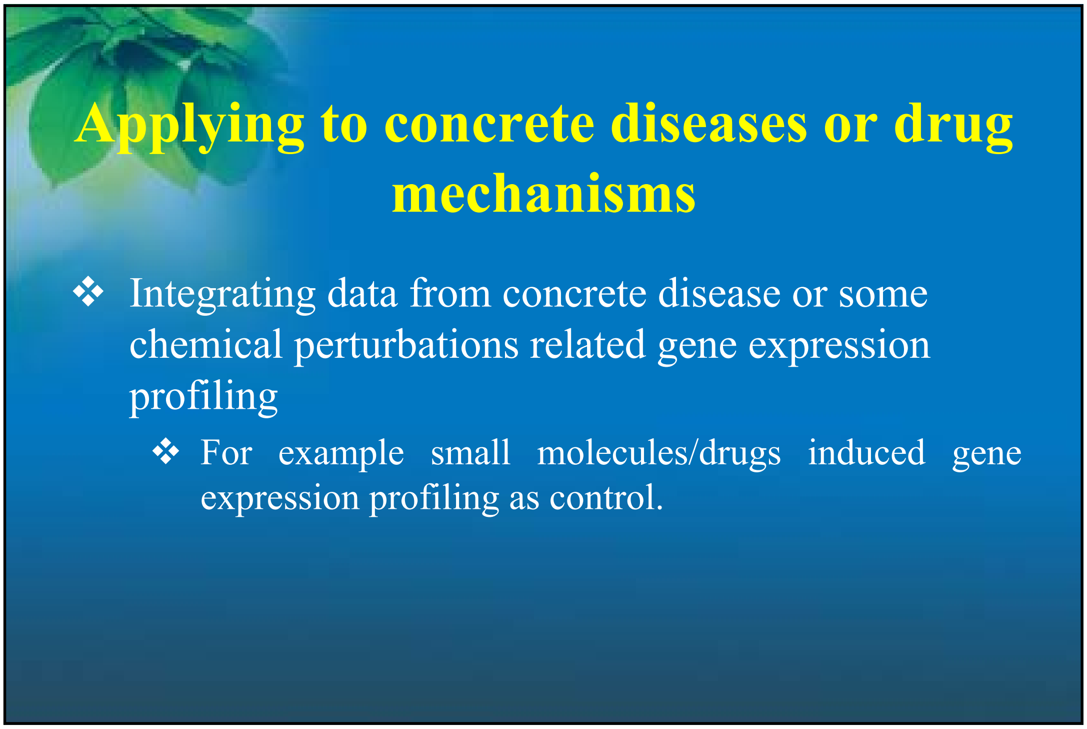


Significant co-regulatory pairs

It was found in our co-regulator list:

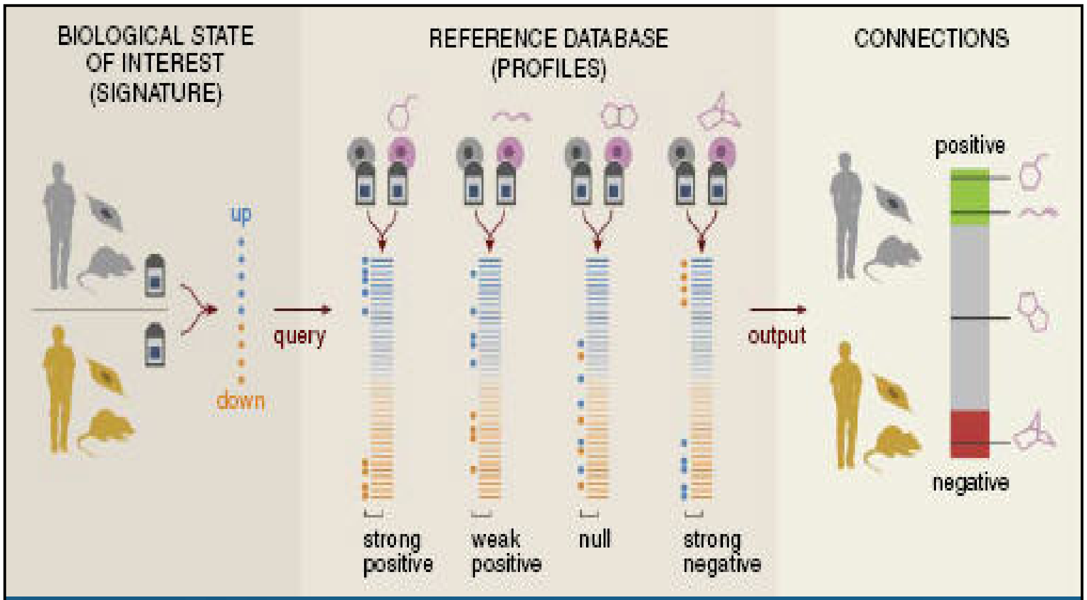
- MYC-MAX: a well known transcriptional complex.
- JUN, JUNB, JUND and FOSL1 forming AP-1 transcriptional complex.
- NFKB1and RELA were another transcriptional complex.
- microRNA within same family tended to regulate same targets,such as let-7 family.
- Most co-regulating pairs were labeled in same sub-network cluster divided previously.

Kang Tu, et.al., ***Nucleic Acid Research*** revised.

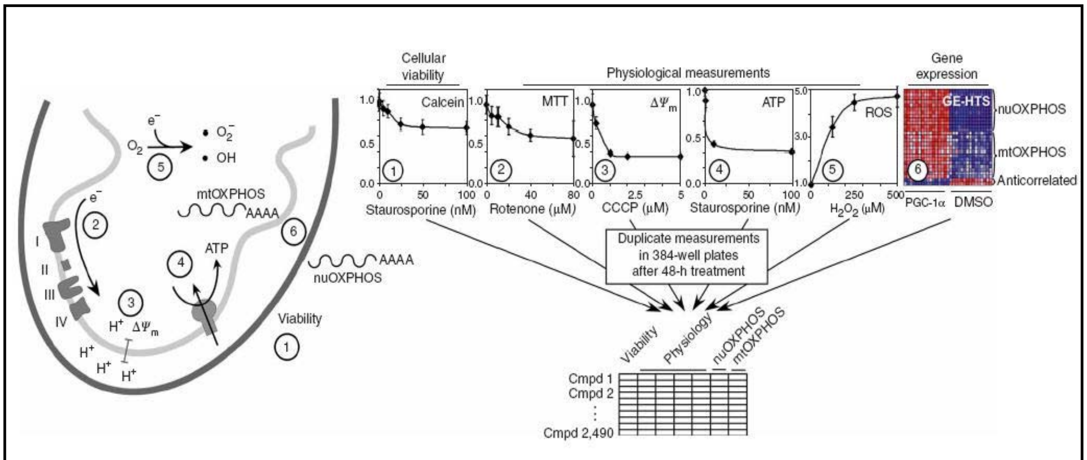


Applying to concrete diseases or drug mechanisms

- Integrating data from concrete disease or some chemical perturbations related gene expression profiling
  - For example small molecules/drugs induced gene expression profiling as control.

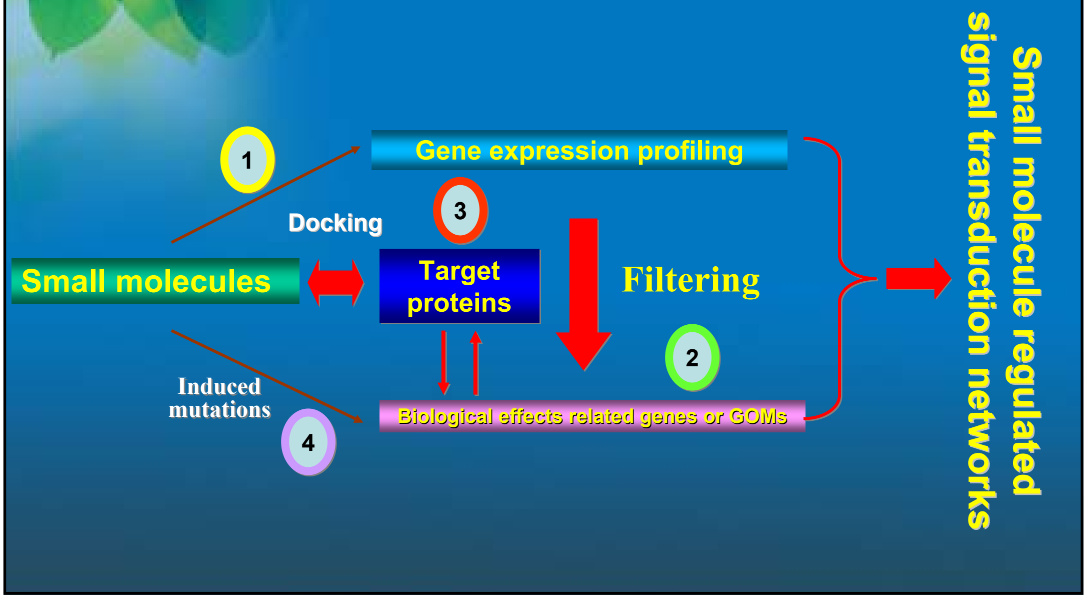


Lamb, ea al., 2006. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science, V. 313, 1929-1935.*

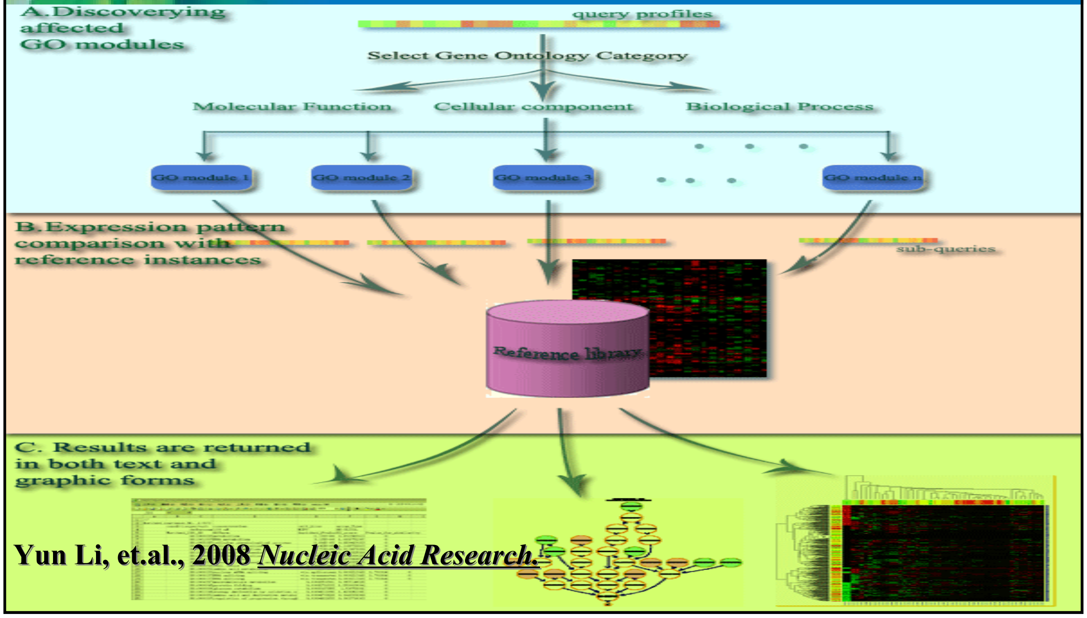


Mitochondrial physiology and gene expression across 2,490 chemical perturbations. The calcein assay (1)measures cell viability and filters out overtly toxic compounds. The MTT assay (2) measures cellular dehydrogenase activity, which is inhibited by the complex I inhibitor rotenone. The JC-1 assay (3) measures the mitochondrial membrane potential (DCm) and drops acutely after the addition of the mitochondrial uncoupler carbonyl cyanide m-chlorophenylhydrazone (CCCP). A luciferase-based assay measures ATP (4), which is reduced by staurosporine. CM-H2DCFDA is a fluorescent probe of cellular ROS (5), which can be stimulated by the addition of H2O2. The expression of both nuOXPHOS and mtOXPHOS transcripts is measured by a multiplex PCR technique, GE-HTS (6). Each column of the heat map represents one sample replicate; expression levels for each gene are row-normalized. Treatment with PGC-1a, an inducer of OXPHOS gene expression, is used as a positive control. All assays were performed in biological duplicate in 384-well format after 48 h of treatment in differentiated murine C2C12 myotubes. Data from 2,490 distinct compounds are incorporated into the screening compendium. ***Bridget, et al., 2008, Large-scale chemical***

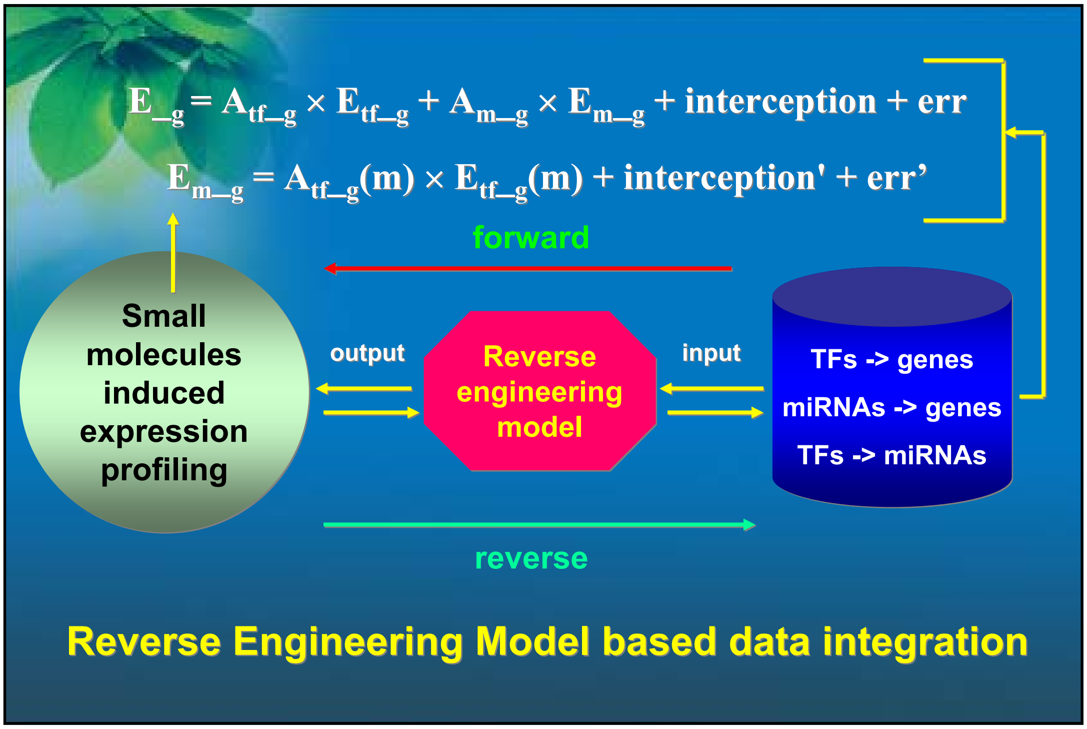
Small Molecule (drugs) Regulation


1
Gene expression profiling
3
Docking
Small molecules
Target proteins
Filtering
2
Induced mutations
4
Biological effects related genes or GOMs
Small molecule regulated signal transduction networks

Procedures for the gene expression profiling related small molecule alignment


A.Discoverying affected GO modules
query profiles
Select Gene Ontology Category
Molecular Function
Cellular component
Biological Process
GO module 1
GO module 2
GO module 3
GO module n
B.Expression pattern comparison with reference instances
sub-queries
Reference library
C. Results are returned in both text and graphic forms

Yun Li, et.al., 2008 *Nucleic Acid Research.*

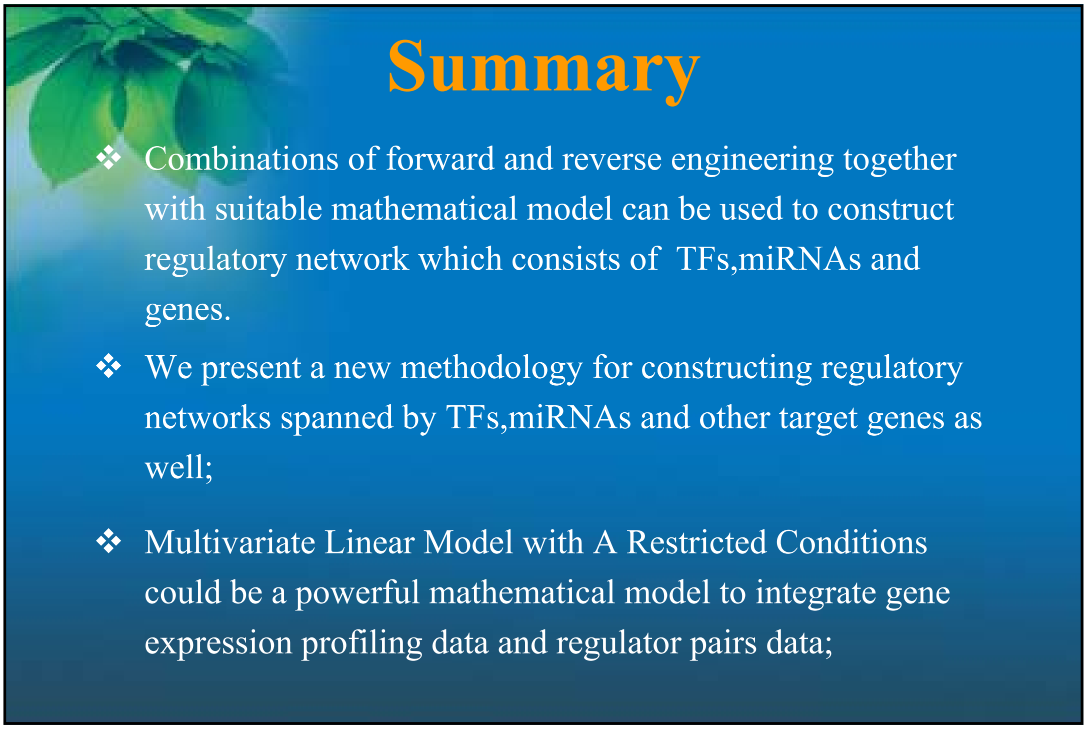


$$E_{\_g} = A_{tf\_g} \times E_{tf\_g} + A_{m\_g} \times E_{m\_g} + \text{interception} + \text{err}$$

$$E_{m\_g} = A_{tf\_g}(m) \times E_{tf\_g}(m) + \text{interception}' + \text{err}'$$

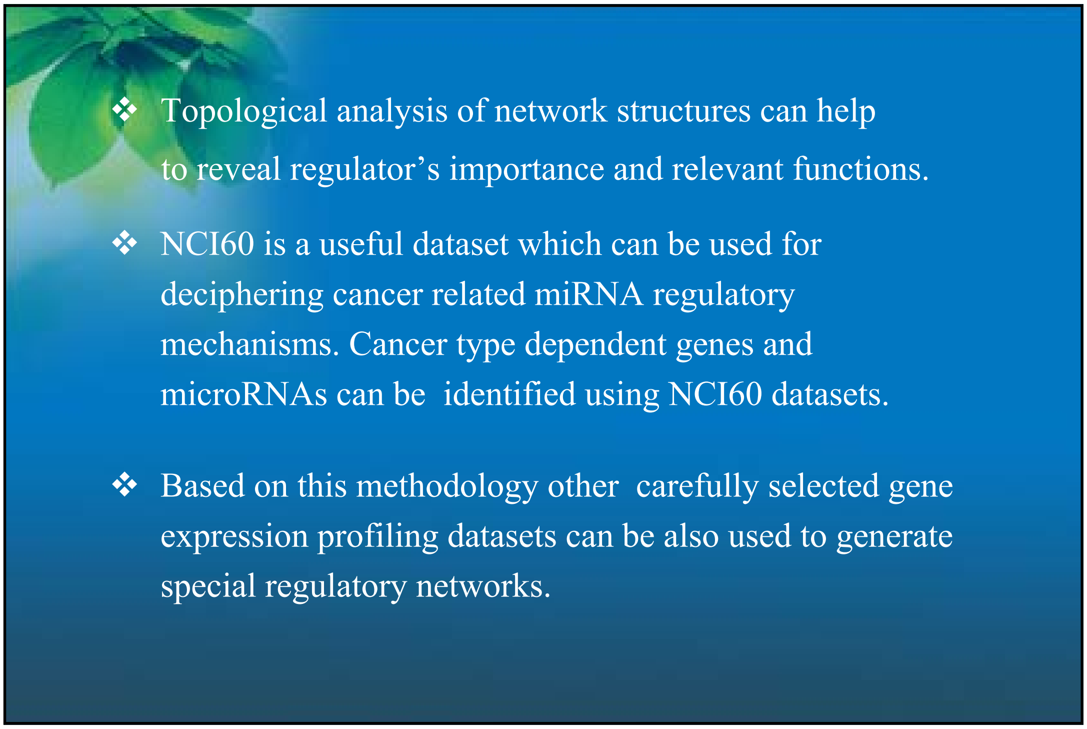
forward
Small molecules induced expression profiling
output
Reverse engineering model
input
TFs -> genes
miRNAs -> genes
TFs -> miRNAs
reverse

Reverse Engineering Model based data integration



Summary

- Combinations of forward and reverse engineering together with suitable mathematical model can be used to construct regulatory network which consists of TFs,miRNAs and genes.
- We present a new methodology for constructing regulatory networks spanned by TFs,miRNAs and other target genes as well;
- Multivariate Linear Model with A Restricted Conditions could be a powerful mathematical model to integrate gene expression profiling data and regulator pairs data;



- ❖ Topological analysis of network structures can help to reveal regulator's importance and relevant functions.
- ❖ NCI60 is a useful dataset which can be used for deciphering cancer related miRNA regulatory mechanisms. Cancer type dependent genes and microRNAs can be identified using NCI60 datasets.
- ❖ Based on this methodology other carefully selected gene expression profiling datasets can be also used to generate special regulatory networks.

Acknowledgements

Dr. Kang Tu, Dr. Xie Lu, Prof. Lei Liu

Thank you for your Attention!