Intra-relation Reconstruction from Inter-relation: miRNA to Gene Expression

Dokyoon Kim[1,2, †], Hyunjung Shin[3, †,*] ,Su-Yeon Lee[1,2], Ju Han Kim[1,2,*]

[1]Seoul National University Biomedical Informatics (SNUBI), Div. of Biomedical Informatics, Seoul National University College of Medicine, Seoul 110-799, Korea
[2]Systems Biomedical Informatics Research Center, Seoul National University, Seoul 110-799, Korea
[3]Department of Industrial & Information Systems Engineering, Ajou University, San 5, Wonchun-dong, Yeoungtong-gu, 443-749, Suwon, Korea

Abstract

Gene expression profiling has been used to molecularly characterize various tumors and tissues. However, regulation of gene expression by microRNAs (miRNAs) has attracted much attention recently. MicroRNAs are regulators of gene expression, mainly functioning by decreasing mRNA levels of their multiple targets. Normally, intra-relations from gene expression or miRNA data can be constructed for explaining cancer phenotype. However, intra-relations are not fully elucidating complex cancer mechanism because the information that miRNAs and target genes are strongly associated with different biological processes is missing. As the recent studies for target prediction of miRNAs are getting increased, the inter-relation between miRNA and gene expression can be constructed from biological experimental data and genomic knowledge. In this study, we propose an integrated framework that combines genomic dataset from gene expression and genomic knowledge from inter-relation between miRNA and gene expression for the molecular-based classification of clinical outcomes. According to our results, accuracy of prediction model increases because of incorporation of information fused over genomic dataset (gene expression) and genomic knowledge (target relation between miRNA and gene expression). This suggests that gene expression regulation through mechanisms that involve miRNAs has valid knowledge for elucidating the cancer phenotype.

Acknowledgement: This work was supported by Post Brain Korea 21 and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (2009-0065043/2011-0018257).

Abstract (Summary)

Normally, intra-relations from gene expression or miRNA data can be constructed for explaining cancer phenotype

However, intra-relations are not fully elucidating complex cancer mechanism because the information that miRNAs and target genes are strongly associated with different biological processes is missing

In this study, we propose an integrated framework that combines genomic dataset from gene expression and genomic knowledge from inter-relation between miRNA and gene expression for the molecular-based classification of clinical outcomes

Contents

1 Introduction

2 Data

3 Methods

4 Results

5 Conclusion

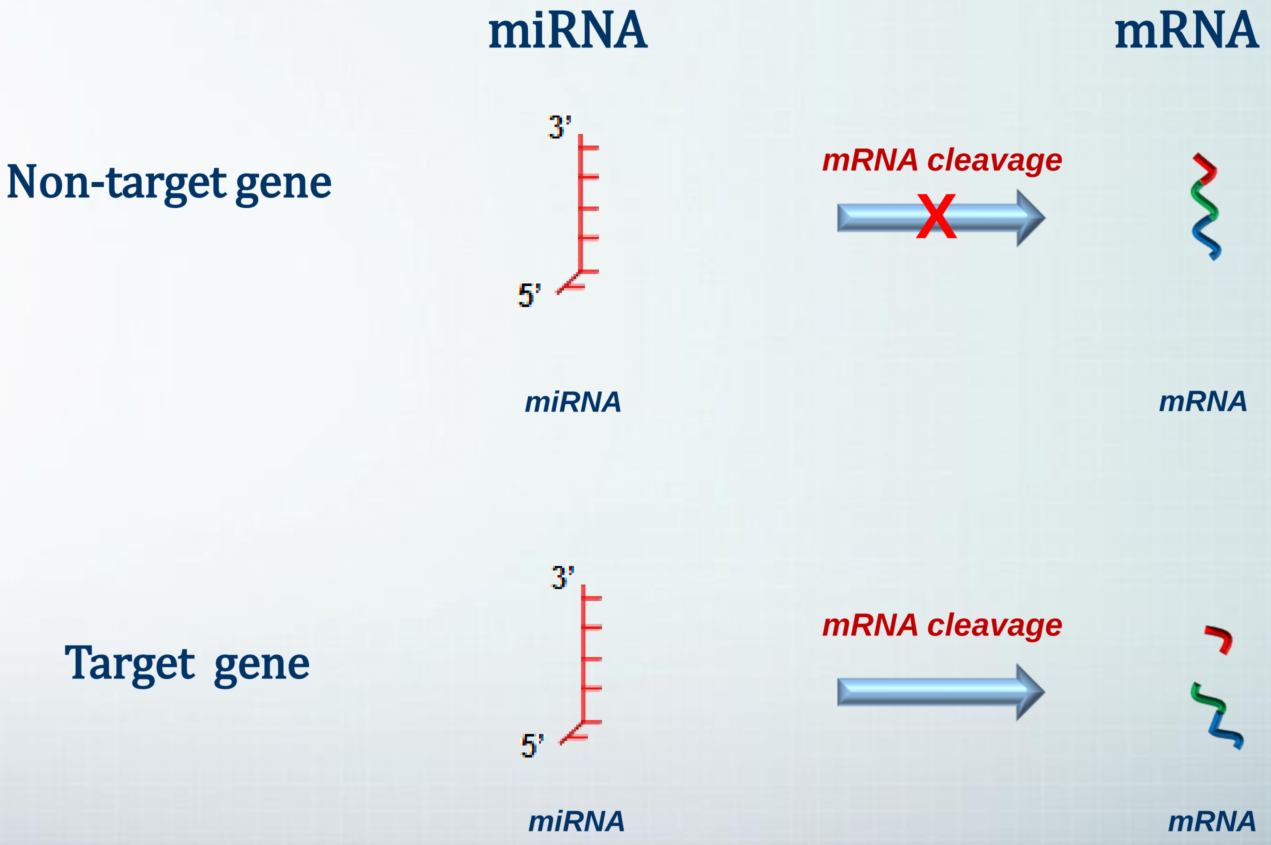
Introduction

Introduction

- Gene expression profiling has been used to molecularly characterize various tumors and tissues
- However, regulation of gene expression by microRNAs (miRNAs) has attracted much attention recently
- miRNAs regulate many genes associated with different biological processes such as development, stress response, apoptosis, proliferation, and tumourigenesis

Regulation mechanism of miRNA and target genes

- **miRNAs are involved in the post-transcriptional regulation of genes either by mRNA cleavage and degradation or by repressing the translation of mRNA into protein**



miRNA

mRNA

Non-target gene

mRNA cleavage

miRNA

mRNA

Target gene

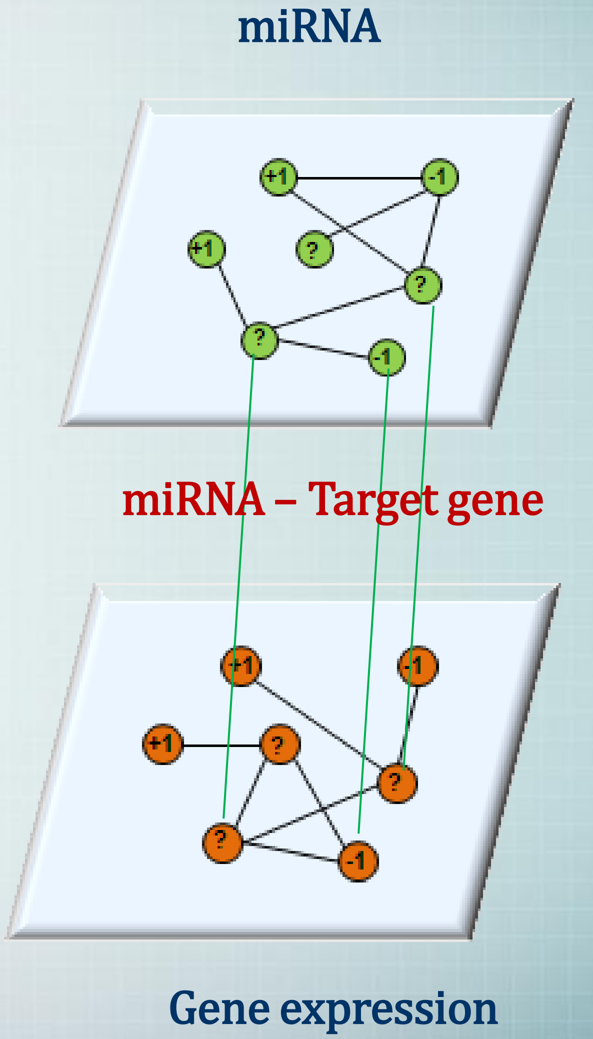
mRNA cleavage

miRNA

mRNA

Motivation

- **Intra-relation: the relation between entities on a specific biological level**
- **Inter-relation: the relation between different levels**
- **Normally, intra-relations from gene expression or miRNA data can be constructed for explaining cancer phenotype**
- **However, intra-relations are not fully elucidating complex cancer mechanism because the information that miRNAs and target genes are strongly associated with different biological processes is missing**



miRNA

miRNA – Target gene

Gene expression

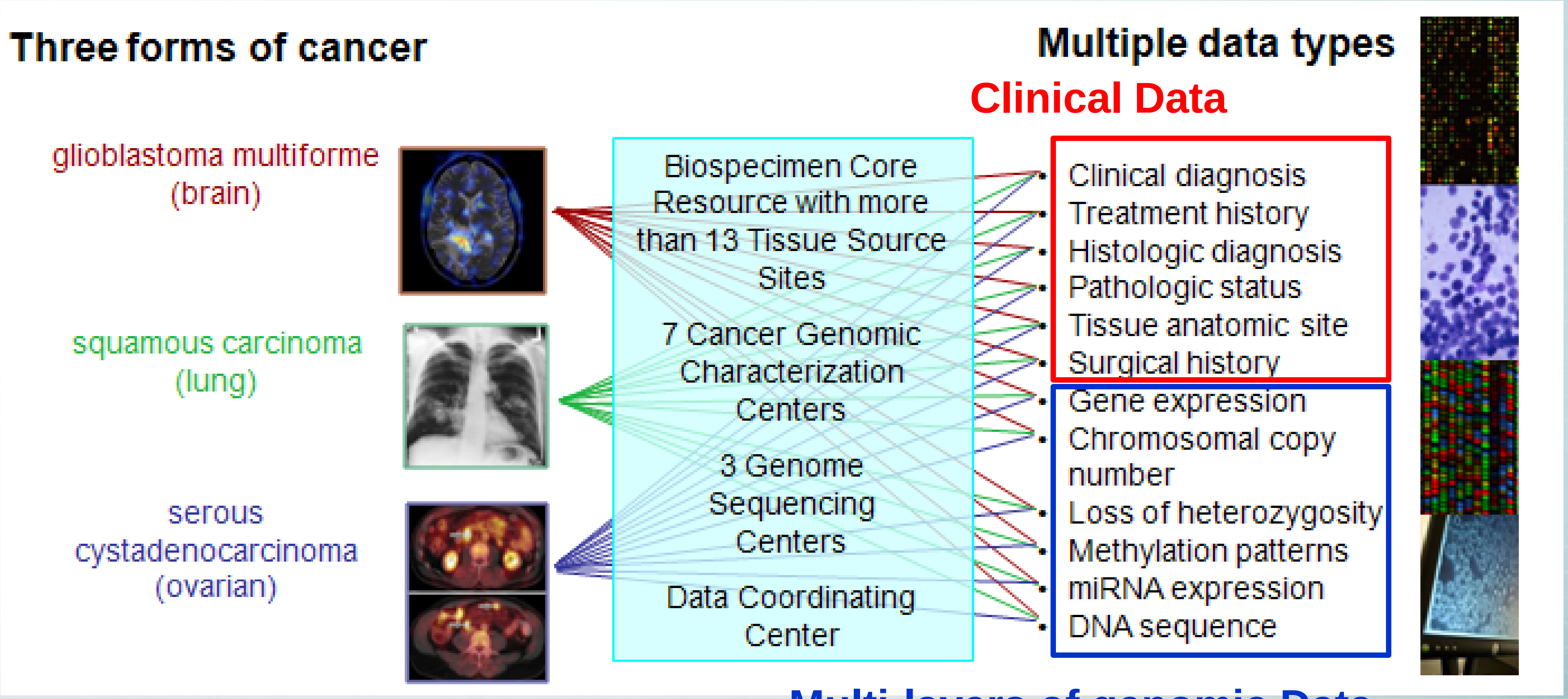
Purpose of the study

- **How informative is inter-relationship between miRNA and gene expression for cancer clinical outcome prediction?**

- **Propose an integrated framework that combines genomic dataset from gene expression and genomic knowledge from inter-relation between miRNA and gene expression for the molecular-based classification of clinical outcomes**

Data

TCGA: Connecting multiple sources, experiments, and data types



Three forms of cancer

glioblastoma multiforme (brain)

squamous carcinoma (lung)

serous cystadenocarcinoma (ovarian)

Biospecimen Core Resource with more than 13 Tissue Source Sites

7 Cancer Genomic Characterization Centers

3 Genome Sequencing Centers

Data Coordinating Center

Multiple data types

Clinical Data

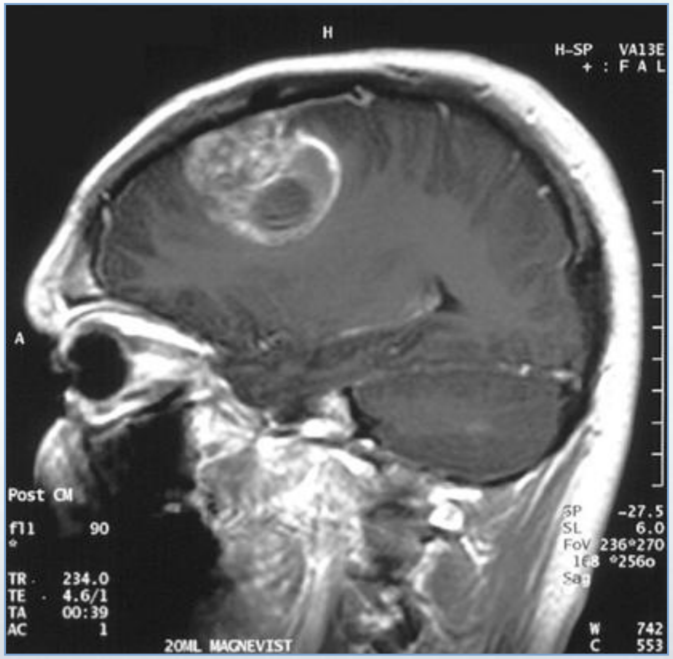
- Clinical diagnosis
- Treatment history
- Histologic diagnosis
- Pathologic status
- Tissue anatomic site
- Surgical history

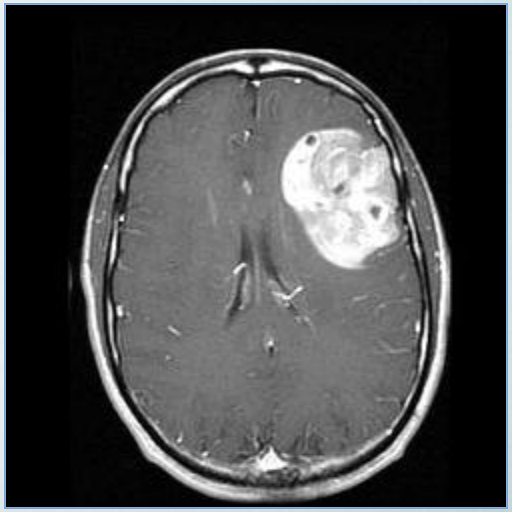
- Gene expression
- Chromosomal copy number
- Loss of heterozygosity
- Methylation patterns
- miRNA expression
- DNA sequence

Multi-layers of genomic Data

Glioblastoma Multiforme (GBM)

- **Most common and aggressive primary brain tumor in adults**
  - Median survival of one year
  - One of the hallmarks of GBM is its inherent tendency to recur





Data description

Data type	Platform	Num of Features
Gene Expression	Affymetrix HT Human Genome U133 Array Plate Set	12,043
miRNA	Agilent Human miRNA Microarray Rel12.0	799

Clinical outcome	Num of samples (Neg / Pos)
Survival status (short-term vs. long-term)	82 (54 / 28)

miRNA – target gene relation

- **In order to get target information between miRNA and mRNA**
  - **Used miRecords which is integrated resources of miRNA that store target interactions produced by 11 established miRNA target prediction program**

Xiao *et al*., 2009

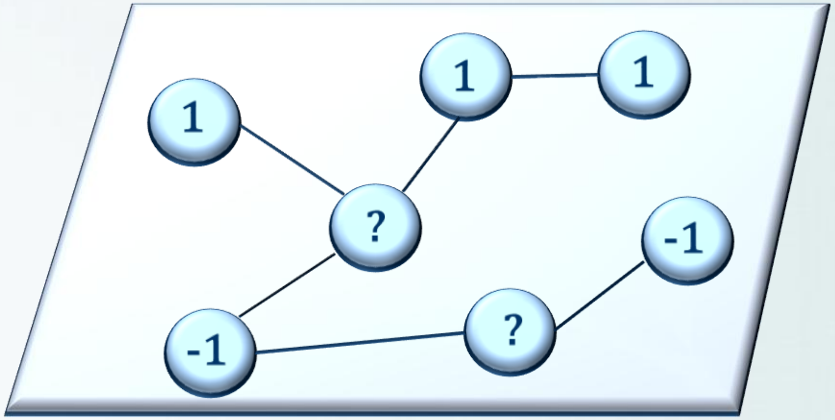
- **Among 11 algorithms, a binary relation between miRNA and mRNA was set when more than 3 algorithms provide the target relation**

Methods

Approaches

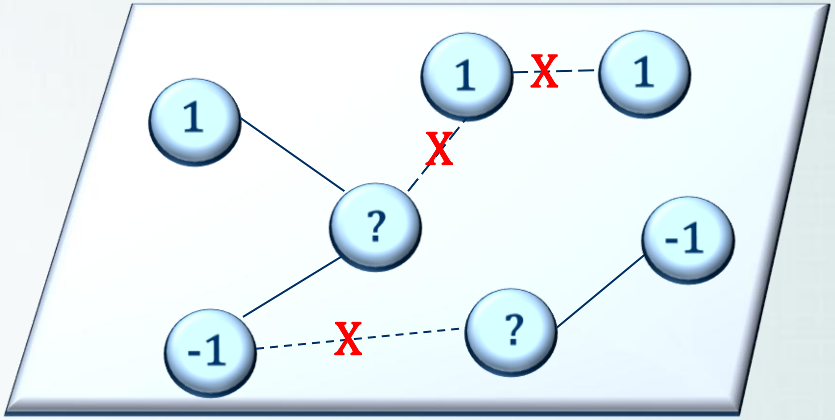
- G_O: Original graph from gene expression
- G_{D50}: Gene expression graph with 50% damages
- G_R: Reconstructed graph via inter-relationship between miRNA and gene expression
- G_A: Augmented graph by 50% damaged graph and reconstructed graph

G_O: Original graph from gene expression



Gene expression (G_O)

G_{D50}: Gene expression graph with 50% damages



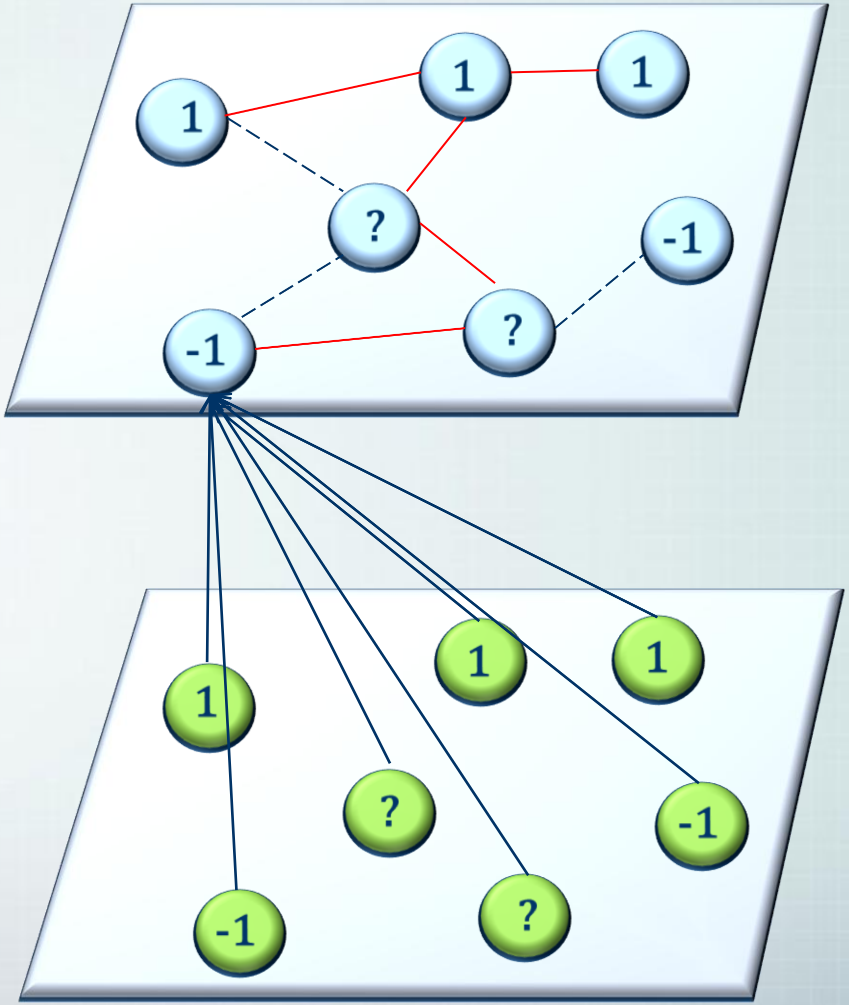
Gene expression (G_{D50})

G_R: Reconstructed graph via inter-relationship between miRNA and gene expression

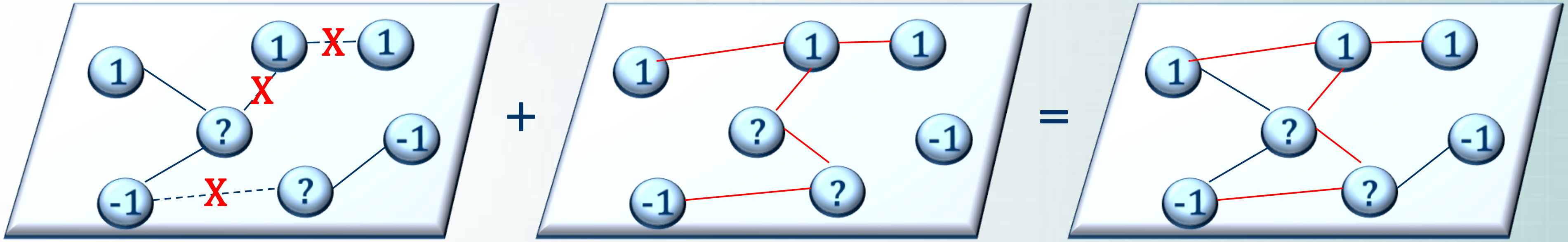
Gene expression (G_R)

Inter-relationship

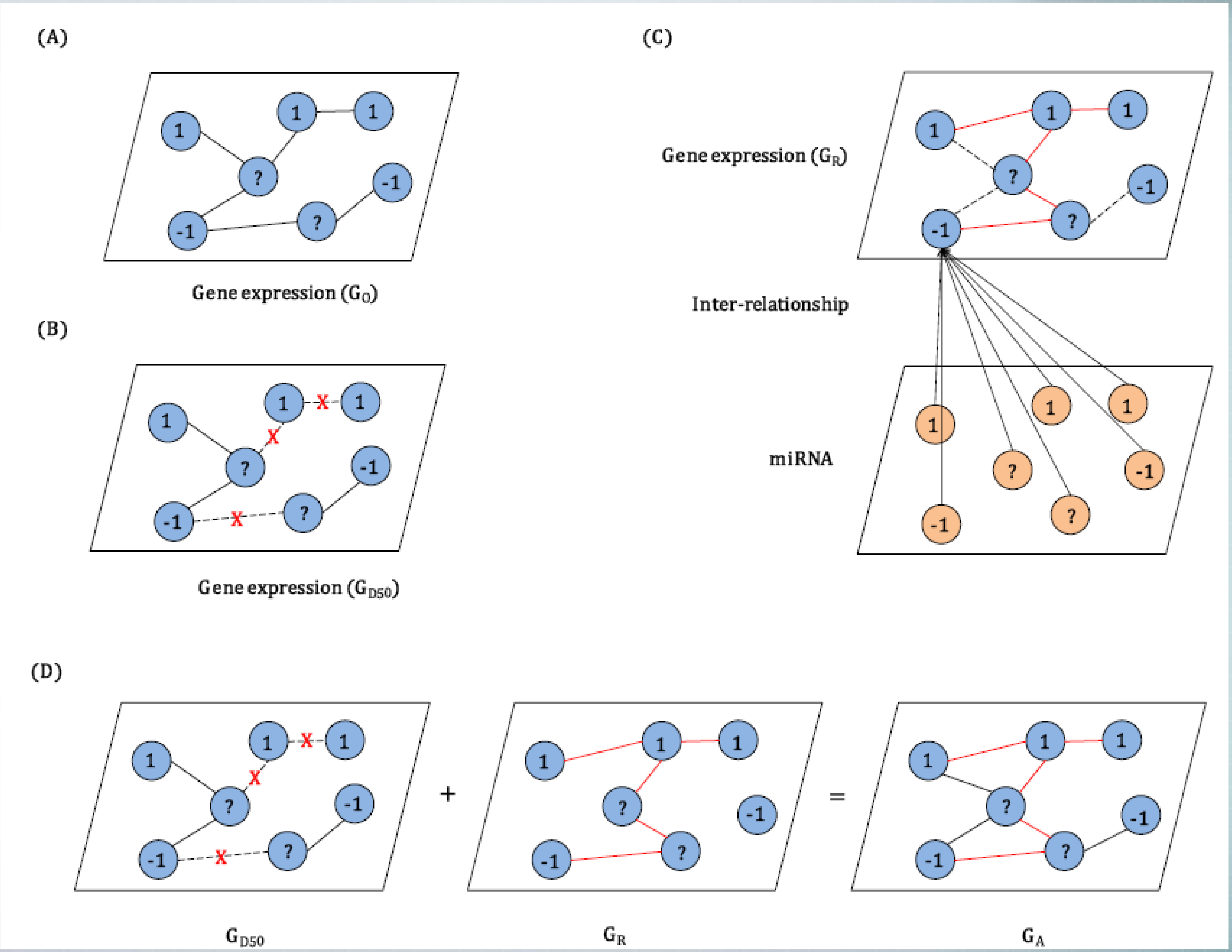
miRNA



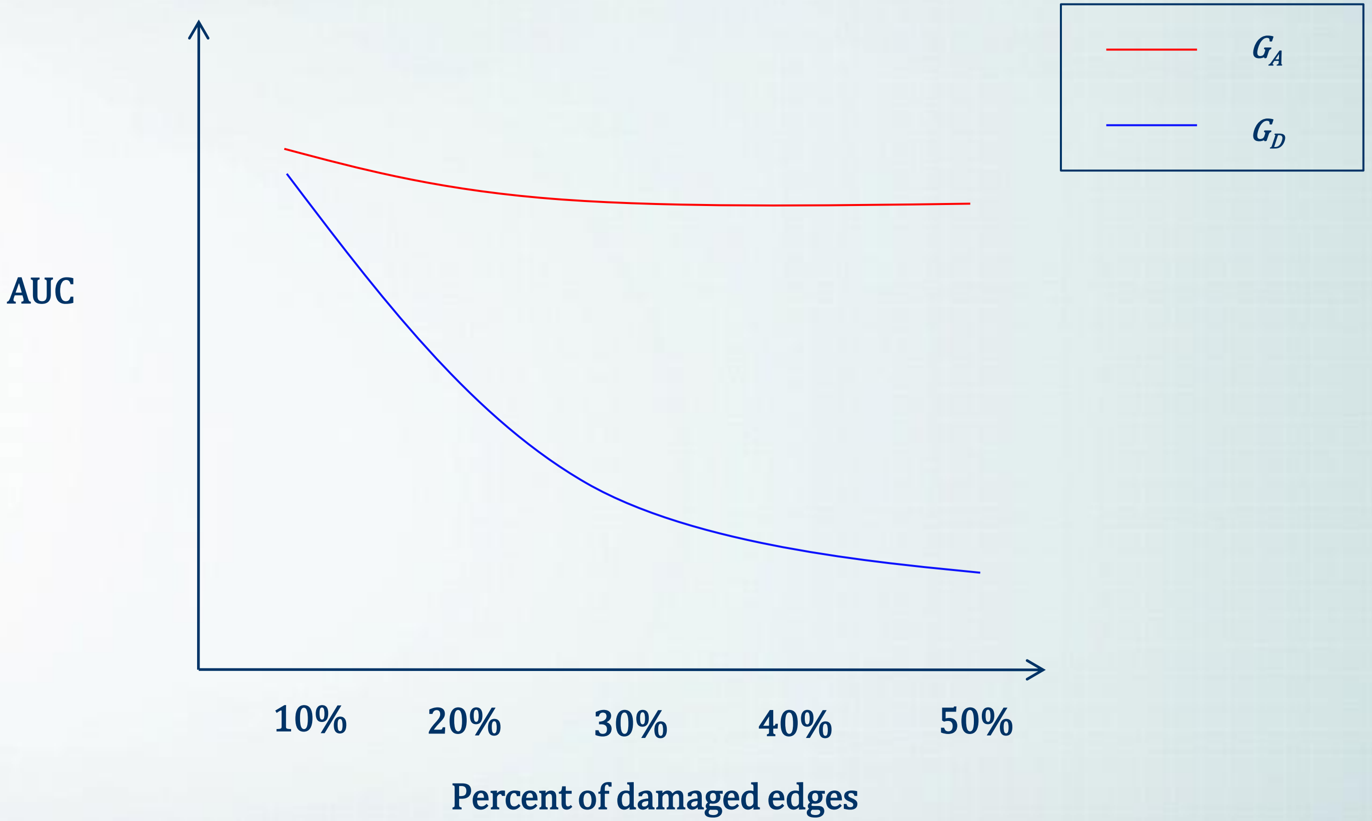
G_A: Augmented graph by 50% damaged graph and reconstructed graph



G_{D50} G_R G_A

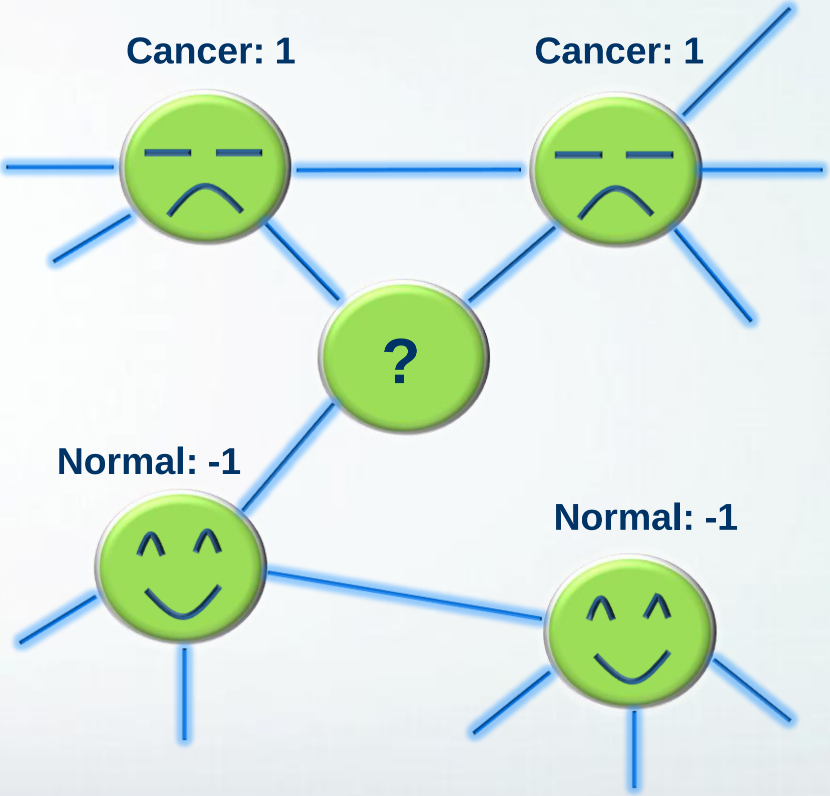

(A)
1
1
1
?
-1
-1
?
Gene expression (G_O)
(B)
1
1
X
1
X
?
-1
-1
X
?
Gene expression (G_{D50})
(C)
Gene expression (G_R)
1
1
1
?
-1
-1
?
Inter-relationship
miRNA
1
1
1
?
-1
?
-1
(D)
+
=
G_{D50}
G_R
G_A

Expected Results



Prediction based on intra-relation from mRNAs

- Graph-based semi-supervised learning (SSL)



●	Patient
—	Association of two patients (similarities)
- 1	Normal
1	Cancer
?	Patient in question

Graph-based Semi-Supervised Learning (SSL)

- Objective function

$$\min_f = \underbrace{(f-y)^T(f-y)}_{\text{Loss}} + \underbrace{\mu f^T L f}_{\text{Smoothness}}$$

  - **Loss condition:** In labeled nodes, final output should be closed to the given label
  - **Smoothness condition:** final output should not be too different from the adjacent node's output
  - *L* is called the graph Laplacian matrix where

$$L = D - W, \qquad D = diag(d_i), \quad d_i = \sum_j w_{ij}$$

- Solution

$$f = (I + \mu L)^{-1} y$$

Tsuda *et al*., 2005 Shin *et al*., 2007

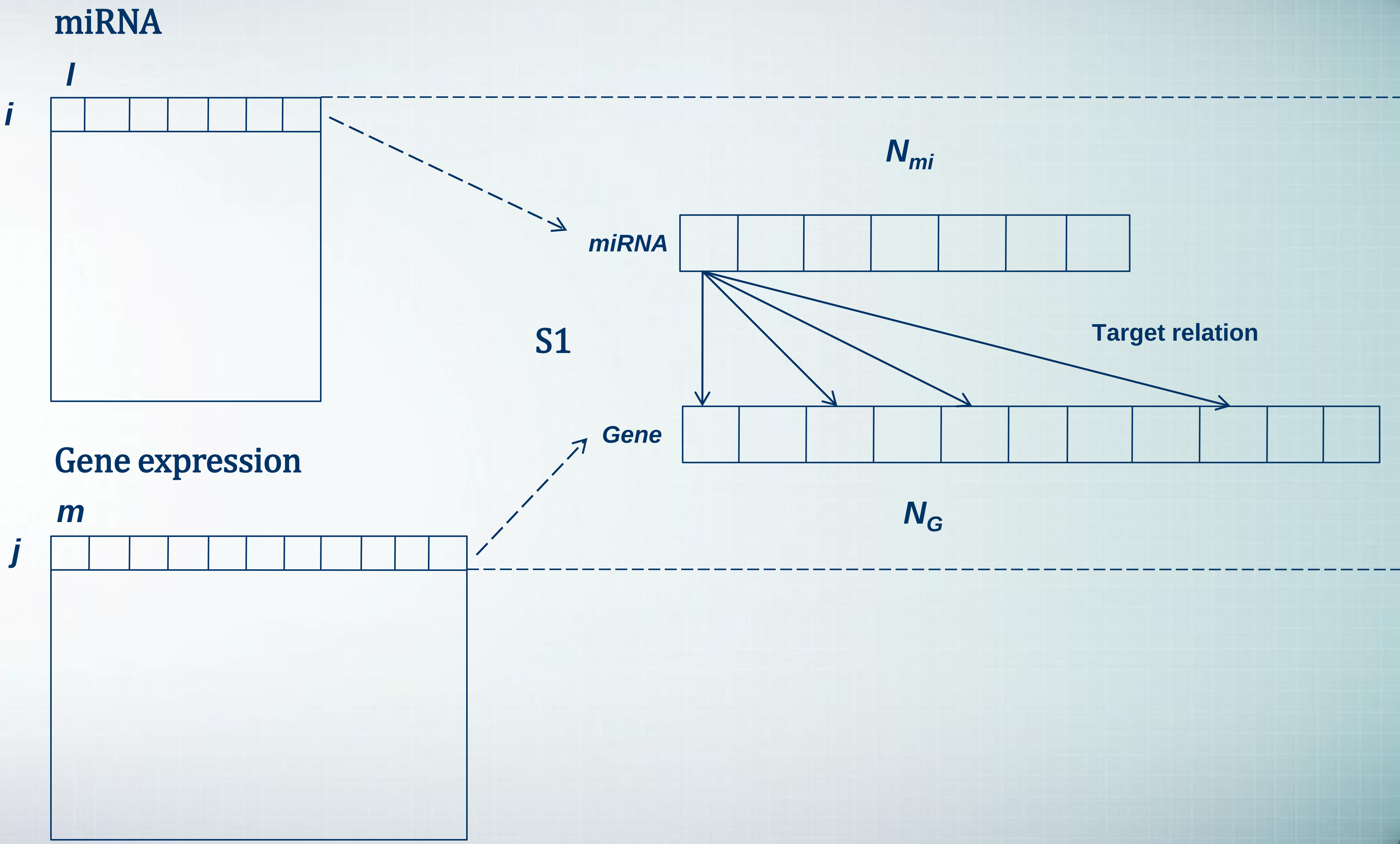
Input for SSL: Weight Matrix (*W*)

- **Exp-weighted *K*-NN graphs**

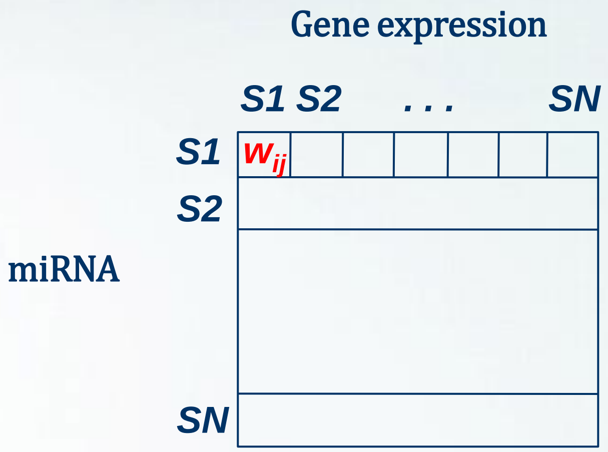
$$W_{ij} = \exp(-\frac{d(i,j)^2}{\alpha^2})$$

  - Nodes *i, j* are connected by an edge if *i* is in *j*'s *K*-nearest-neighborhood or vice versa
  - d: Euclidean distance
  - Hyperparameter α controls the decay rate

Data Description



Prediction based on inter-relation from miRNA to mRNA



Gene expression

S1 S2 . . . SN

miRNA

S1 w_{ij}

S2

SN

Inter-relationship weight matrix

$$f_{ij} = \sum_{l=1}^{N_{mi}} \sum_{m=1}^{N_G} miRNA(i,l) \bullet gene(j,m)$$

where miRNA and gene are target relation

$$Z_{ij} = \frac{f_{ij} - \overline{f}}{std(f)}$$

$$w_{ij} = \frac{1}{1 + e^{-Z_{ij}}}$$

Integration of multiple networks

- Two graphs can be integrated from finding optimum combination coefficients

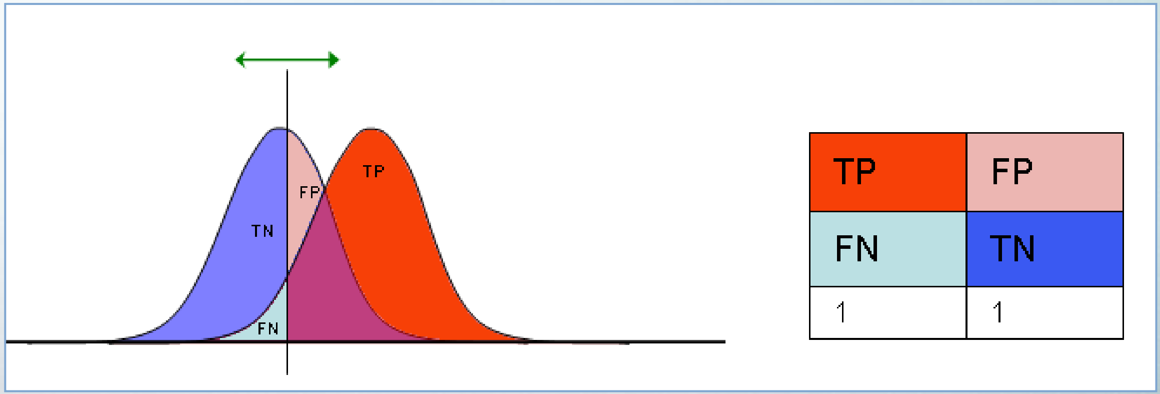
$$\min_{\alpha} y^T (I + \sum_{k=1}^{K} \alpha_k L_k)^{-1} y \qquad \sum_k \alpha_k \leq \mu$$

$$f = (I + \sum_{k=1}^{K} \alpha_k L_k)^{-1} y$$

Tsuda *et al*., 2005 Shin *et al*., 2007

Experiment setting

- G_O: Original graph from gene expression
- G_D: Gene expression graph with damages (10% ~ 90%)
- G_R: Reconstructed graph via inter-relationship between miRNA and gene expression
- G_A: Augmented graph by damaged graph and reconstructed graph
- Performance measure: AUC (Area under the ROC curve)



TP	FP
FN	TN
1	1

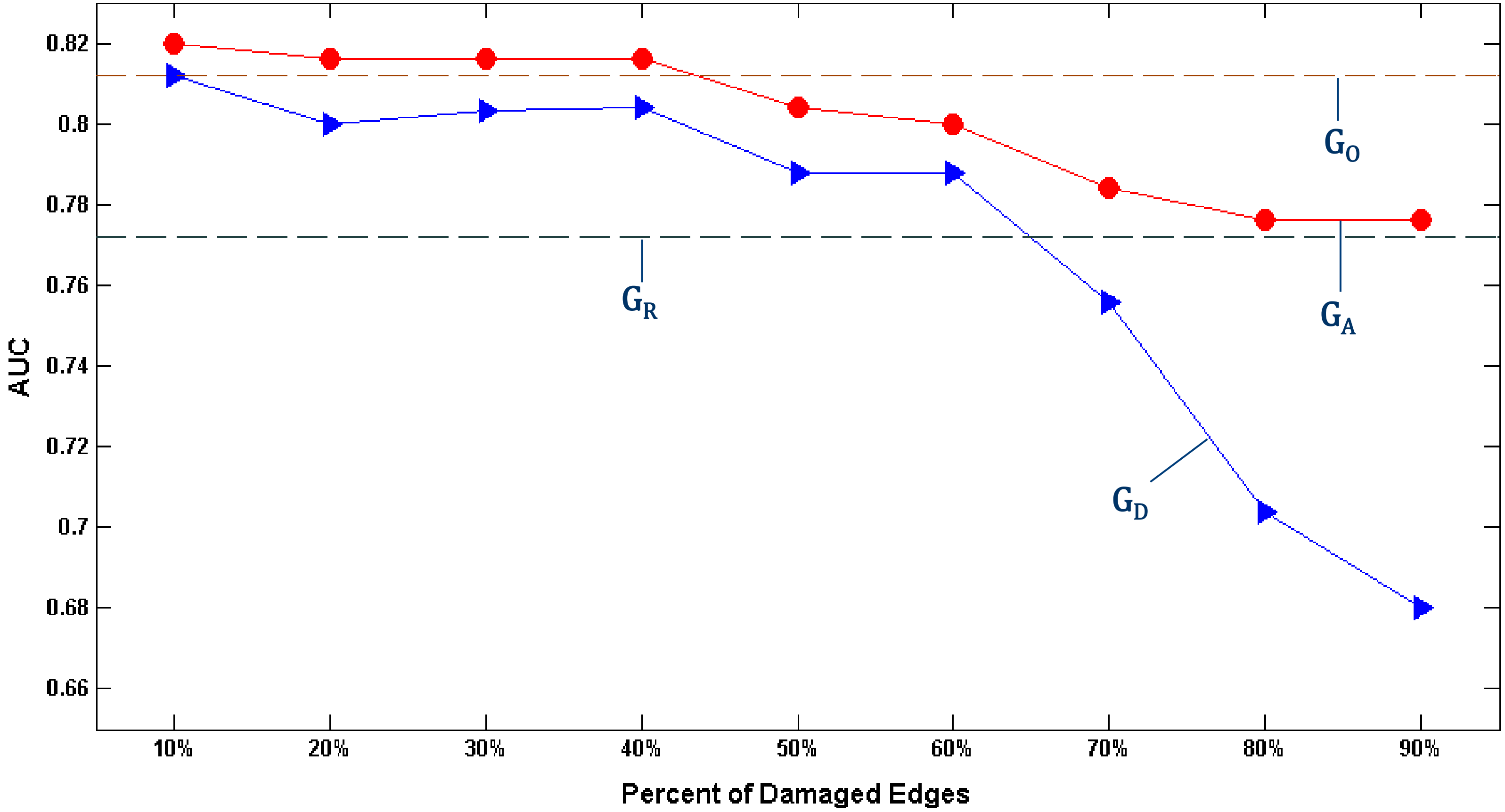
Model Parameter Selection

- Parameters should be selected by user when learning with SSL
  - K : KNN
  - μ : SSL

- Combination of parameters
  - K = {3, 4, 5, 6, 7, 8, 9, 10, 20, 30}
  - μ = {0.001, 0.01, 0.05, 0.1, 0.2, 0.5, 0.7, 1.0, 10.0, 100.0, 1000.0}

Results

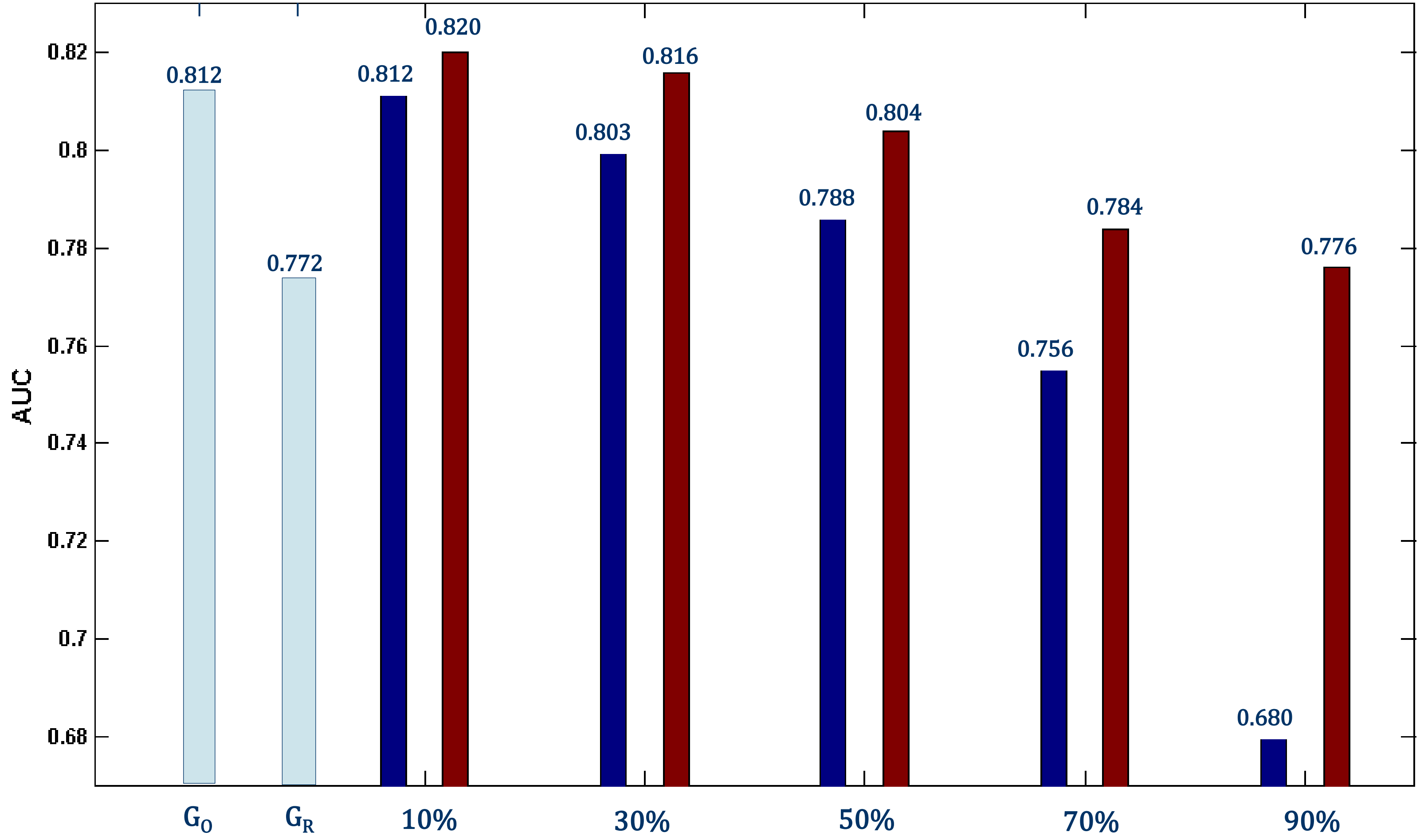
Result comparison



Significance test of the performance differences

Percent of damaged edges	AUC of G_D	AUD of G_A	P-value
10%	0.812	0.820	1.87e-02
30%	0.803	0.816	2.09e-03
50%	0.788	0.804	3.43e-05
70%	0.756	0.784	9.59e-08
90%	0.680	0.776	1.24e-13

Improving performance from augmented knowledge based on inter-relation between miRNA and miRNA



Conclusion

Discussion & Conclusion

- **Proposed an integrated framework that combines genomic dataset and genomic knowledge**
  - **In order to provide a preliminary insight on the question that is how informative is inter-relationship between and gene expression**

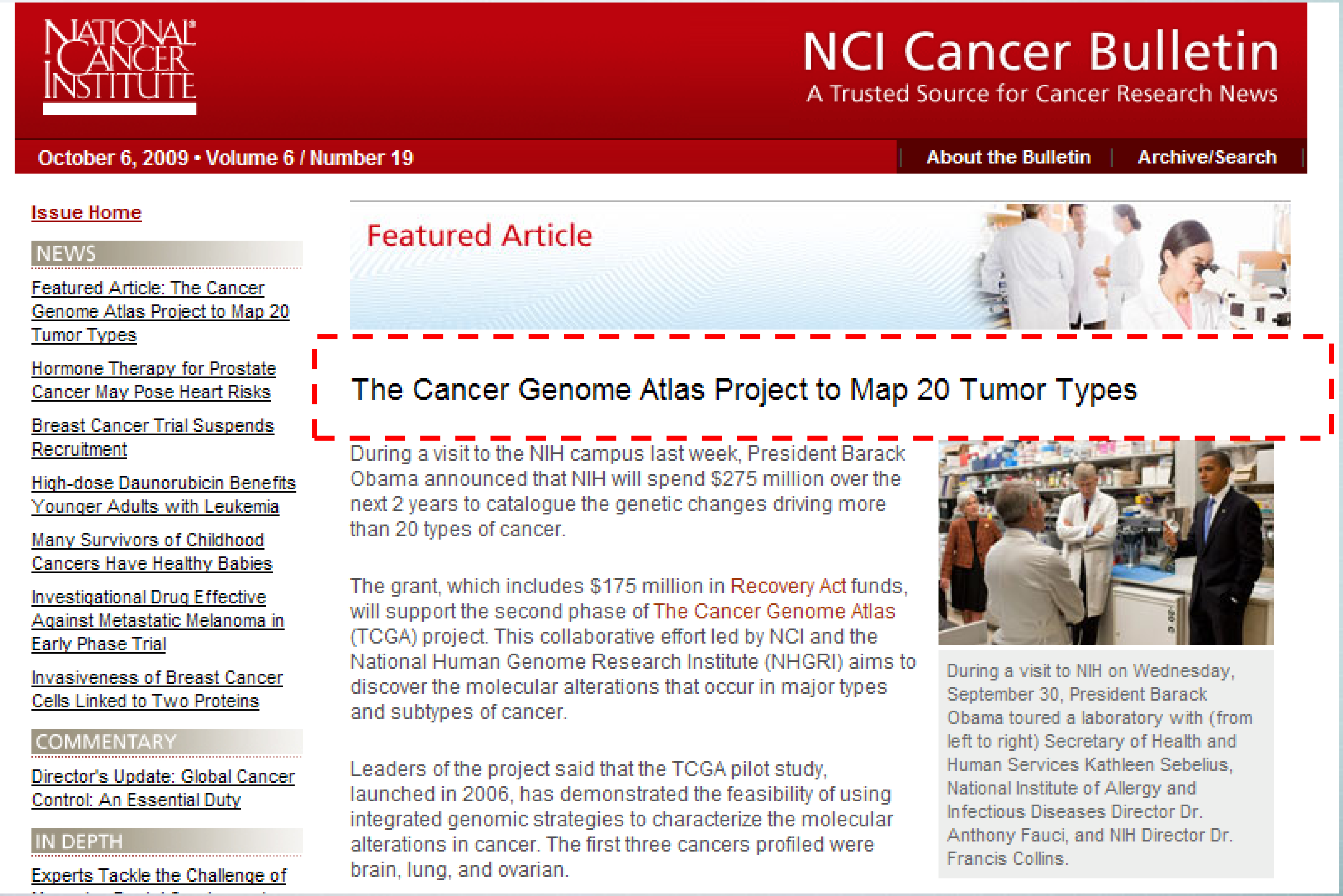
- **Inter-relation from miRNA and target gene could help constructing intra-relation from gene expression for better cancer clinical outcome prediction**

- **Our results suggests that genomic knowledge is complementary to the prediction power of explaining cancer phenotype**
  - **Even though genomic data such as gene expression has incomplete information**

Future work

- **Gene expression regulation through mechanisms that involve miRNAs is valid knowledge for elucidating the cancer phenotype**
  - **Because miRNAs regulate many genes associated with different biological processes**
- **Reconstructing intra-relation from miRNA**
- **Combining gene expression, miRNA, and inter-relation**

The Second phase of TCGA Project



NCI Cancer Bulletin

A Trusted Source for Cancer Research News

October 6, 2009 • Volume 6 / Number 19

NEWS

- Featured Article: The Cancer Genome Atlas Project to Map 20 Tumor Types
- Hormone Therapy for Prostate Cancer May Pose Heart Risks
- Breast Cancer Trial Suspends Recruitment
- High-dose Daunorubicin Benefits Younger Adults with Leukemia
- Many Survivors of Childhood Cancers Have Healthy Babies
- Investigational Drug Effective Against Metastatic Melanoma in Early Phase Trial
- Invasiveness of Breast Cancer Cells Linked to Two Proteins

COMMENTARY

- Director's Update: Global Cancer Control: An Essential Duty

IN DEPTH

- Experts Tackle the Challenge of

Featured Article

The Cancer Genome Atlas Project to Map 20 Tumor Types

During a visit to the NIH campus last week, President Barack Obama announced that NIH will spend $275 million over the next 2 years to catalogue the genetic changes driving more than 20 types of cancer.

The grant, which includes $175 million in Recovery Act funds, will support the second phase of The Cancer Genome Atlas (TCGA) project. This collaborative effort led by NCI and the National Human Genome Research Institute (NHGRI) aims to discover the molecular alterations that occur in major types and subtypes of cancer.

Leaders of the project said that the TCGA pilot study, launched in 2006, has demonstrated the feasibility of using integrated genomic strategies to characterize the molecular alterations in cancer. The first three cancers profiled were brain, lung, and ovarian.

During a visit to NIH on Wednesday, September 30, President Barack Obama toured a laboratory with (from left to right) Secretary of Health and Human Services Kathleen Sebelius, National Institute of Allergy and Infectious Diseases Director Dr. Anthony Fauci, and NIH Director Dr. Francis Collins.

Thank You !

- **Any Question?**