Empirical Comparison on Heterogeneous Genomic Data: CNV, Methylation, miRNA and Gene Expression

Dokyoon Kim[1,2], Young Soo Song[1,3], Ju Han Kim[1,2,*], Hyunjung Shin[3,*]
[1] Seoul National University Biomedical Informatics (SNUBI), Div. of Biomedical Informatics, Seoul National University College of Medicine, Seoul 110-799, Korea
[e-mail: dkkim@snu.ac.kr]
[2] Systems Biomedical Informatics Research Center, Seoul National University, Seoul 110-799, Korea
[e-mail: juhan@snu.ac.kr]
[3] Department of Industrial & Information Systems Engineering, Ajou University
San 5, Wonchun-dong, Yeoungtong-gu, 443-749, Suwon, Korea
[e-mail: teshenawa@naver.com]
[e-mail:shin@ajou.ac.kr]
*Corresponding author: Ju Han Kim and Hyunjung Shin

Abstract

Thanks to the recent collaborative initiative against cancer, heterogeneous types of genomic data from cancer patient become available. The aim of the present study is to compare different types of genomic data for Glioblastoma multiforme (GBM) recurrence prediction. The four types of genomic data, Copy Number Variation (CNV), methylation, miRNA, and gene expression data, are employed and tested on 159 GBM patients using the state-of-the-art machine learning algorithm, semi-supervised learning.

Keywords: Bioinformatics, Microarray, Brain Cancer, Glioblastoma Multiforme, Semi-Supervised Learning

Acknowledgment: The authors would like to gratefully acknowledge support from Post Brain Korea 21 and research grant National Research Foundation of the Korean Government (2009-0065043/2010-0028631).

Empirical comparison on heterogeneous genomic data: CNV, Methylation, miRNA, and Gene Expression

Dokyoon Kim[1,2], Young Soo Song[1,3], Ju Han Kim[1,2,*], Hyunjung Shin[3,*]

[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 110799, Korea
[3]Department of Industrial & Information Systems Engineering, Ajou University, San 5, Wonchun-dong, Yeoungtong-gu, 443-749, Suwon, Korea

Abstract

Thanks to the recent collaborative initiative against cancer, heterogeneous types of genomic data from cancer patient become available. The aim of the present study is to compare different types of genomic data for classification of clinical outcomes in Glioblastoma multiforme (GBM) . The four types of genomic data, Copy Number Variation (CNV), methylation, miRNA, and gene expression data, are employed and tested on 159 GBM patients using the state-of-the-art machine learning algorithm, semi-supervised learning.

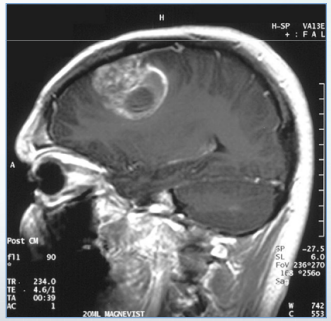
Acknowledgement: The authors would like to gratefully acknowledge support from Post Brain Korea 21 and the research grant from National Research Foundation of the Korean Government (2009-0065043/2010-0028631)

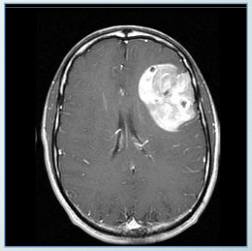
Contents

Introduction

Glioblastoma Multiforme (GBM)

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





Classification in Cancer Research

- Why do we need to classify cancers?
  - The general way of treating cancer is to:
    - Categorize the cancers in different classes
    - Use specific treatment for each of the classes
- Traditional ways to classify cancers
  - Morphological appearance
    Not accurate!
  - Enzyme-based histochemical analyses
  - Immunophenotyping
  - Cytogenetic analysis
    Complicated & need highly specialized laboratories

Classification in Cancer Research (cont'd)

- Microarray-based cancer diagnosis
  - Cancer is caused by changes in the genes that control normal cell growth and death
  - Molecular diagnostics offer the promise of precise, objective, and systematic cancer classification
  - Molecular-based classification of cancer subtypes or clinical outcomes using microarray

Microarray

- **A multiplex technology used in molecular biology and in medicine**
  - Microarray techniques will lead to a more complete understanding of the molecular variations among tumors or clinical outcomes, hence to a more reliable classification

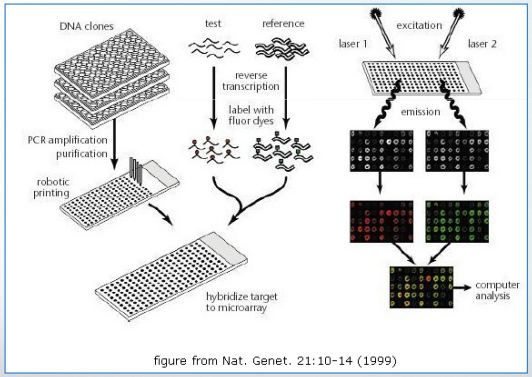


figure from Nat. Genet. 21:10-14 (1999)

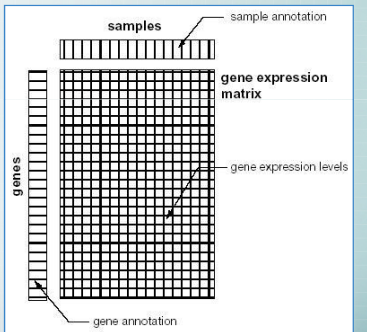
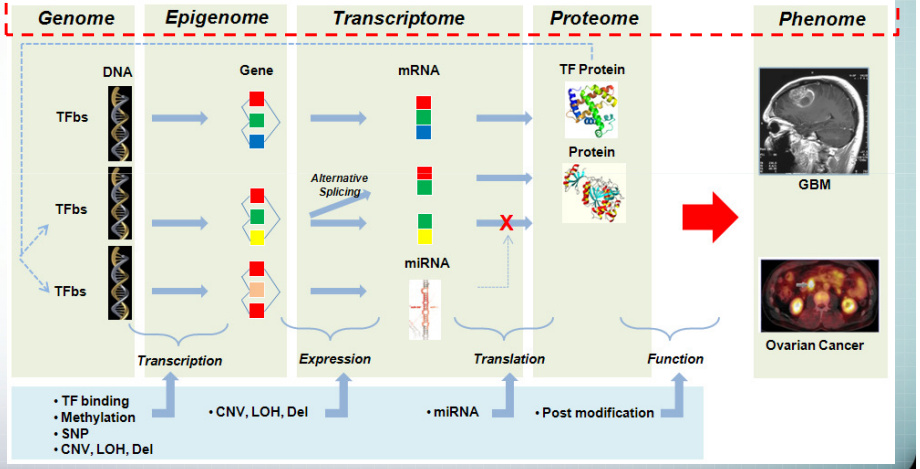


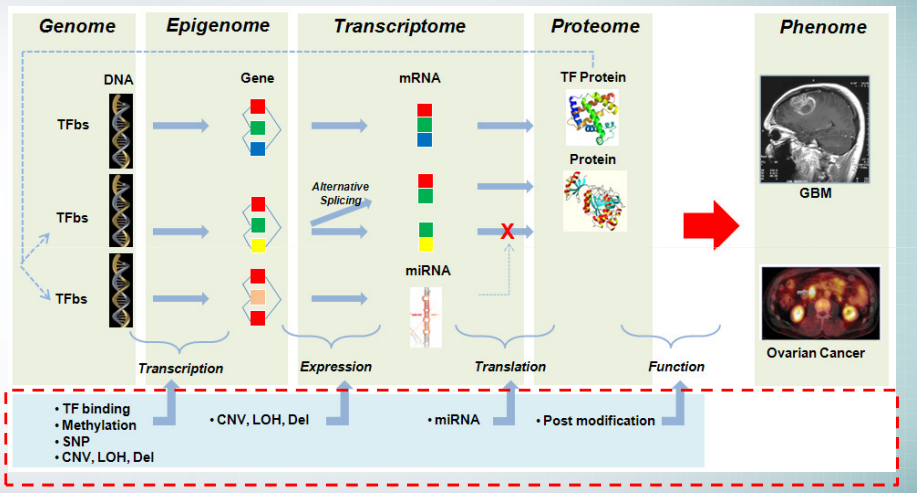
figure from Nat. Genet. 29:365-71 (2001)

The Complex Mechanism of Biological Organization

There are multiple levels in biological system !

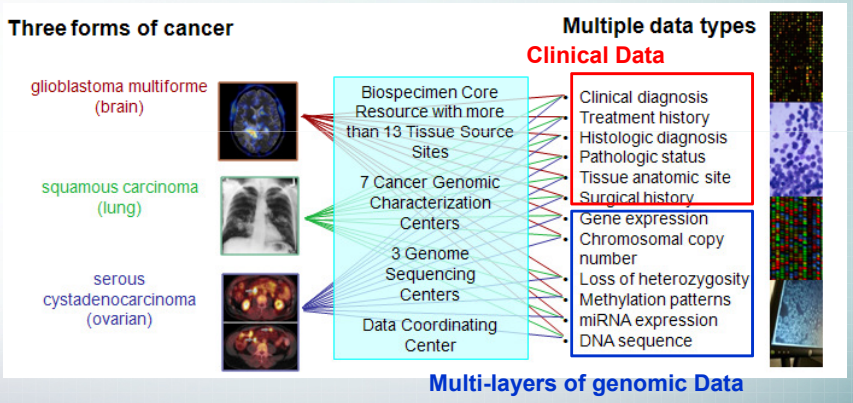


The Complex Mechanism of Biological Organization



There are many exceptional variations between levels such as CNVs, DNA methylation, alternative splicing, miRNA regulation, post translational modification, etc

TCGA: Connecting multiple sources, experiments, and data types



Motivation

- **Cancer can be dysregulated through multiple mechanisms**
  - Modifications to the DNA and the histones
  - Changes in the DNA structure and copy number
  - Mutations in the coding and non-coding sequences

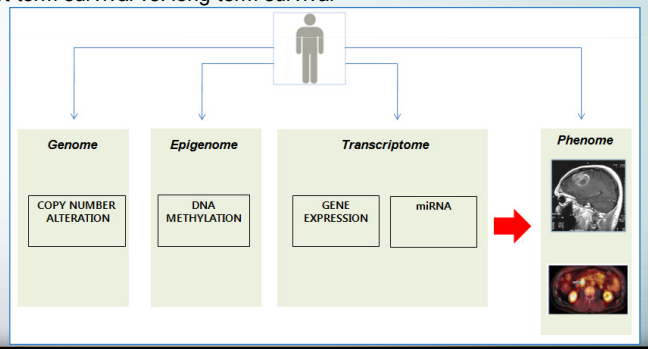
- **These changes can lead to alterations in**
  - Transcription
  - Translation
  - Post-translational modification
  - Ultimately gene and protein function

Motivation (cont'd)

- **With abundance in genomic/clinical data in cancer research**
  - The question that bioinformaticians often encounter is which genomic data is more informative?
- **To wet-lab analysts**
  - It concerns data generation that requires highly cost/time-demanding work and experienced facilities
- **To dry-lab analysts**
  - It concerns selection of appropriate data source for more accurate prediction, avoiding unnecessary waste of computational resource

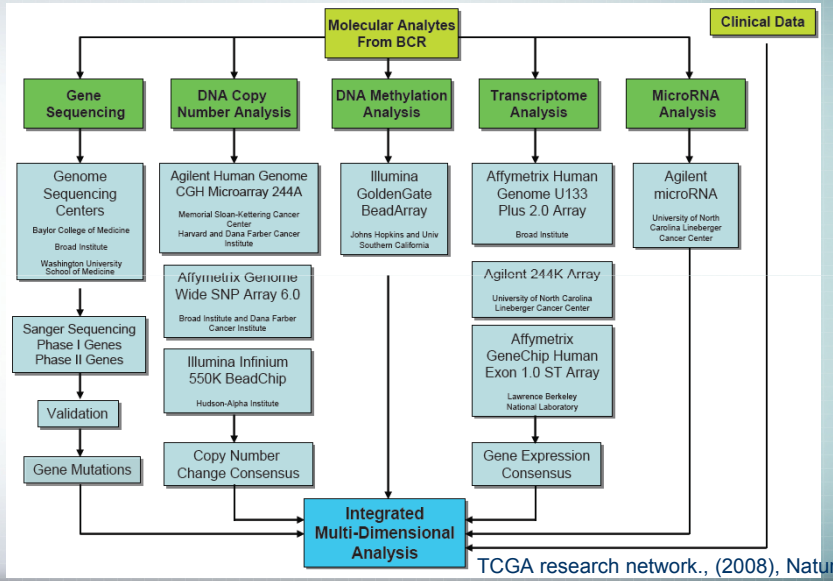
Purpose of the Study

- **To provide a preliminary insight on the question**
  - This study compares different types of genomic data in GBM using the state-of-the-art machine learning algorithm, Semi-Supervised Learning (SSL)
- **Clinical outcomes**
  - Initial GBM vs. recurrent GBM
  - Short-term survival vs. long-term survival



Data

TCGA Data



TCGA research network., (2008), Nature

Download Multi-level Genomic Data

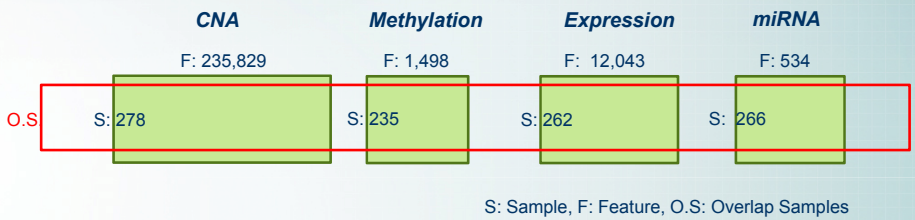
- Available raw and normalized different types of genomic data were retrieved from the TCGA data portal
- Tissue: GBM
- Size: about 230 GBs
- Databasing each level of data for further analysis

Data Description

Data type	Platform	Num of Samples *	Num of Features
CNA	Agilent Human Genome CGH Microarray 244A	278	235,829
Methylation	Illumina DNA Methylation OMA003 Cancer Panel 1	235	1,498
Gene Expression	Affymetrix HT Human Genome U133 Array Plate Set	262	12,043
miRNA	Agilent 8x15K Human miRNA-specific microarray	266	534

*** Samples with tumor type = 'solid tumor'**

Data: Input



S: Sample, F: Feature, O.S: Overlap Samples

- Select overlap samples among multi-level of genomic datasets as an input

Data: Output Variable

Clinical outcome	Num of samples (Neg / Pos)
Disease Recurrence (yes vs. no)	159 (39 / 120)
Survival status (short-term vs. long-term)	82 (54 / 28)

- Define classes
  - Disease recurrence
    - Initial GBM: Procedure_Type = 'Surgical Resection' & Pretreatment_History = 'No'
    - Recurrent GBM: Procedure_Type = 'Secondary Surgery for tumor recurrence'
  - Survival status
    - Short-term survival: Survival < 9 months
    - Long-term survival: Survival > 24 months

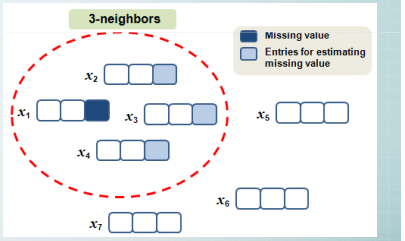
Methods

Data preprocessing: Missing Value Imputation

- Missing values were estimated using K-nearest neighbors with *K* = 15 (Troyanskaya *et al*., 2001)
  - Assume a M x N matrix $G = (g_{i,j})_{i,j=1}^{M,N}$
  - Missing entry $G_{i,l}$ as the weighted average of neighboring genes

$$G_{i,l} = \frac{\sum_{j=1}^{k} G_{j,l} / d_{i,j}}{\sum_{j=1}^{k} 1/d_{i,j}} (j \neq i)$$

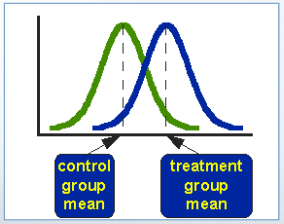
  - , where the k neighbor genes are those with the closest distances to the target gene

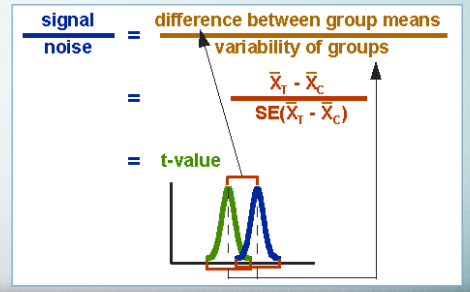


Data preprocessing: Feature Selection

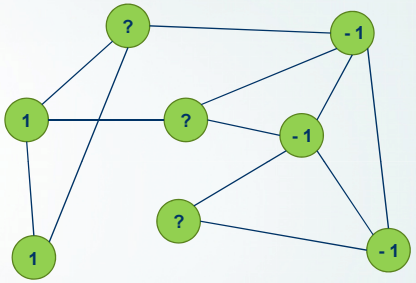
- Identify differential expressed genes from two conditions or phenotypes
  - t-test

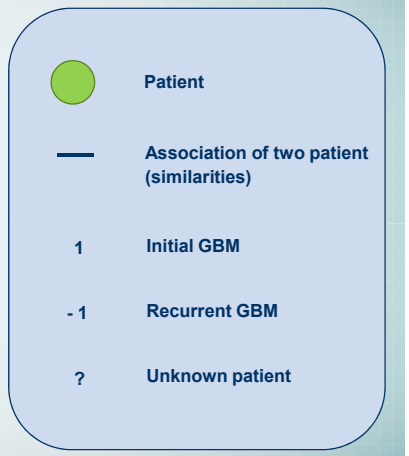
$$t_j = \frac{\overline{X}_{j1} - \overline{X}_{j2}}{\sqrt{\frac{S_{j1}^2}{n_1} + \frac{S_{j2}^2}{n_2}}}, j = 1,\ldots,p \qquad S_{jk}^2 = \frac{1}{n_i - 1}\sum_{i=1}^{n_k}(X_{ijk} - \overline{X}_{jk})^2, k = 1,2$$





Graph-based Semi-Supervised Learning (SSL)





- The goal of SSL is to classify unknown patient into the right class

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

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

Input for SSL: Weight Matrix (*W*)

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

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

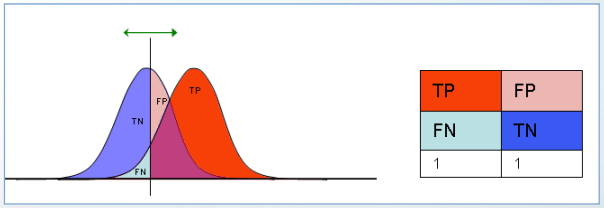
  - Nodes *i, j* are connected by an edge if *i* is in *j*'s *K*-nearest-neighborhood or vice versa
  - d: Euclidean distance
  - Hyperparameter α controls the decay rate

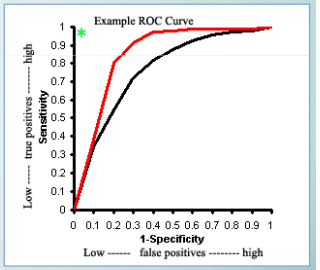
Model Parameter Selection

- Parameters should be selected by user when learning with SSL
  - *K* : *K*NN
  - 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}

Experiment: Measurement (AUC)

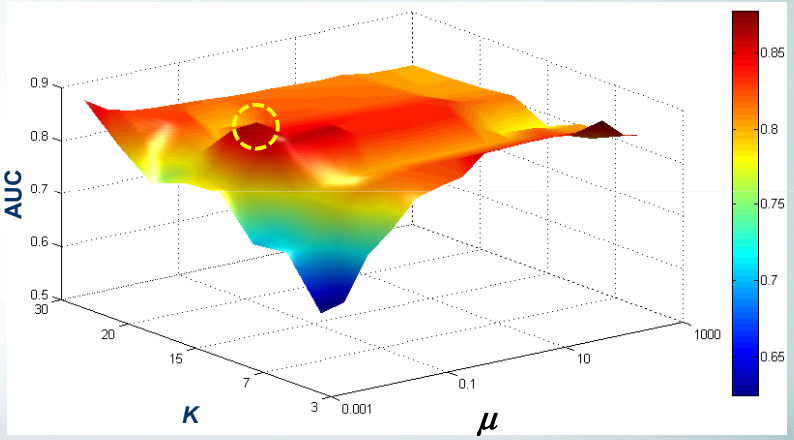
- AUC (Area under the ROC curve)
- 5-fold cross validation





Results

Model Parameter Selection



Survival Status: Gene expression ($p < 0.001$)

Experiment Results: Recurrence

Gene Expression

P_value <	Num of Features	BEST AUC	Avg AUC with Std	K	Mu
1.000	12,043	0.4095	0.3992 ±0.0086	15	0.010
0.100	545	0.4976	0.4842 ±0.0097	15	0.010
0.050	209	0.6583	0.5334 ±0.0405	10	1,000
0.010	17	**0.6667**	0.6098 ±0.0281	15	0.550
0.005	8	0.6369	0.5720 ±0.0327	30	0.650

miRNA

P_value <	Num of Features	BEST AUC	Avg AUC with Std	K	Mu
1.000	534	0.5083	0.4768 ± 0.0205	20	1,000
0.100	58	0.5738	0.5120 ± 0.0289	15	0.600
0.050	29	0.5988	0.4711 ±0.0345	30	1,000
0.010	5	0.7131	0.5879 ±0.0414	30	0.900
0.005	4	0.7107	0.5953 ±0.0459	9	100.0
0.001	3	**0.7226**	0.5900 ±0.0427	30	1.000

Experiment Results: Recurrence

Methylation

P_value <	Num of Features	BEST AUC score	Avg AUC with Std	K	Mu
1.000	1,498	0.6071	0.4220 ±0.0378	3	1,000
0.100	131	**0.6774**	0.5722 ±0.0437	30	0.400
0.050	68	0.6226	0.5454 ±0.0381	15	0.350
0.010	16	0.5536	0.4393 ±0.0405	30	0.050
0.005	10	0.5631	0.4888 ±0.0310	30	0.050

CNA

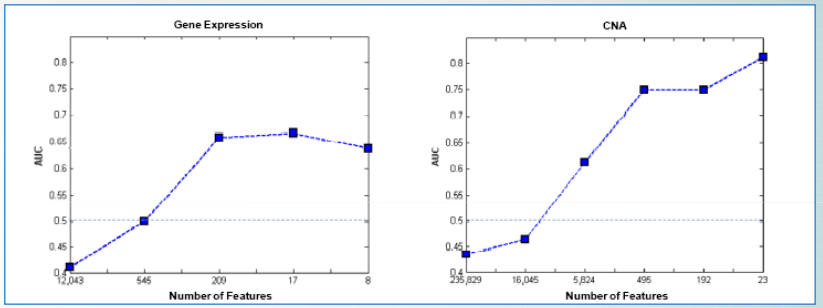
P_value <	Num of Features	BEST AUC score	Avg AUC with Std	K	Mu
1.000	235,829	0.4345	0.4231 ±0.0046	3	0.001
0.100	16,045	0.4631	0.4376 ±0.0099	3	0.001
0.050	5,824	0.6119	0.5845 ±0.0244	7	0.001
0.010	495	0.7488	0.7051 ±0.0197	10	1,000
0.005	192	0.7500	0.6895 ±0.0396	3	0.900
0.001	23	**0.8131**	0.7498 ±0.0241	30	0.300

Best AUC Comparison

Outcome	Data type	AUC
Recurrence	CNA	**0.8131**
	Methylation	0.6774
	Gene Expression	0.6667
	miRNA	0.7226
Survival Status	CNA	0.8160
	Methylation	0.7480
	Gene Expression	**0.8560**
	miRNA	0.7480

- Recurrence: CNA data showed the best performance (AUC: 0.8131)
- Survival status: Gene expression data showed the best performance (AUC: 0.8560)

AUC Changes after Feature Selection



Recurrence

- Increasing tendency of AUC through feature selection

Conclusion

Biological Implication

- Disease recurrence in GBM
  - CNA data showed the best performance among multi-level of genomic data sets
  - These findings suggest that tumor progression from initial to recurrent tumor has high probability to be associated with an increase of genetic changes
  - Therefore, recurrences in GBM are more advanced than initial GBM
  - An increasing amount of DNA copy number alterations is a dominant feature between initial and recurrent GBM
- Survival status in GBM
  - Even though CNA data showed good performance, gene expression data was the most dominant feature in survival status
  - These findings suggest that functional level is relatively better than structural level to distinguish between short-term and long-term survival in GBM

Biological Implication

- ***Why did CNA data show the best performance in disease recurrence in GBM?***
- ***Why did gene expression show the best performance in survival status in GBM?***
- **We could get these meaningful questions from our approach for further study**

Importance of Feature Selection

- Molecular biology data from high-throughput technologies
  - High-dimension
  - Noisy
  - Missing value
  - These factors listed above affect performance
  - Importance of feature selection in bioinformatics is getting increased
- AUCs were increased dramatically through feature selection using simple t-test

Conclusion

- Classification of clinical outcomes in GBM was performed as a base task in order to provide a preliminary insight on the question:
  - Which genomic data is more informative when multiple genomic dataset are available
- Specific genomic data with high performance could be solely used in classification tasks where other genomic dataset are unavailable
  - Due to the high cost associated with the experimental procedure and sample availability
- However, data integration with multi-level of genomic data is needed to better explain phenotype
  - Different genomic data contain partly independent and partly complementary pieces of information

Multi-level Genomic Data Integration

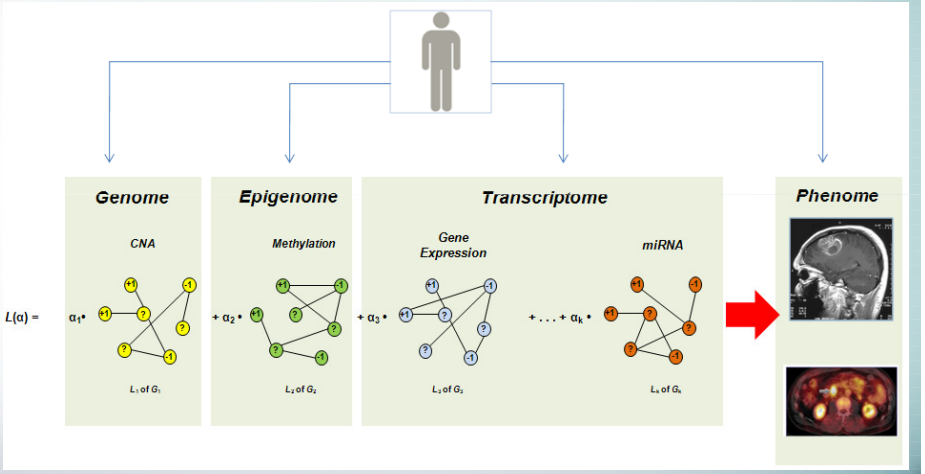
- Multiple graphs from heterogeneous genomic data can be combined

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

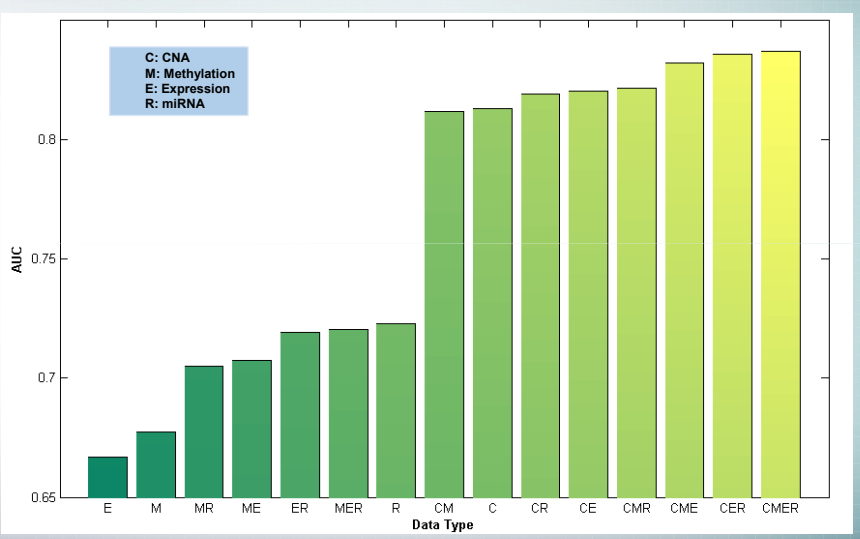
$$f = (I + \sum_{k=1}^{K} \alpha_k L_k)^{-1} y$$

Shin *et al*., 2007

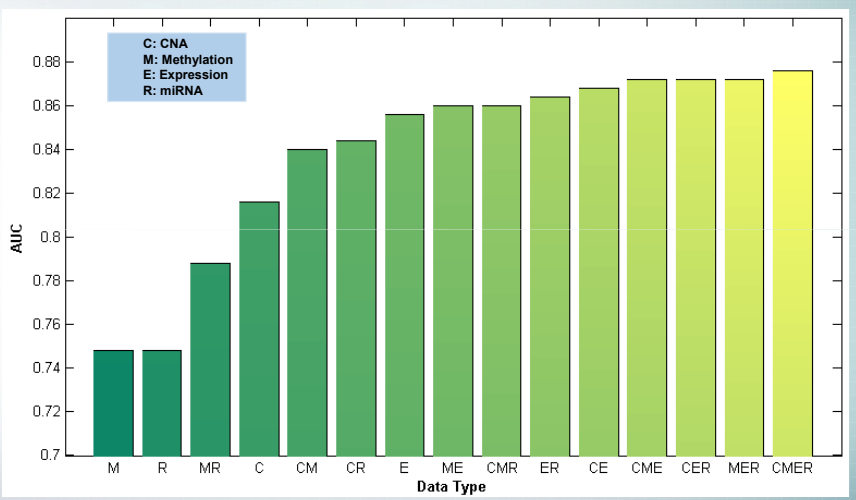
Multi-level Genomic Data Integration



Integration of Multi-level Genomic Data: Recurrence



Integration of Multi-level Genomic Data: Survival Status



The Second phase of TCGA Project

References

- Ben-Hur, A. and Noble, W.S. (2005) Kernel methods for predicting protein-protein interactions, *Bioinformatics*, 21 Suppl 1, i38-46.
- Chari, R., *et al.* (2008) SIGMA2: a system for the integrative genomic multi-dimensional analysis of cancer genomes, epigenomes, and transcriptomes, *BMC Bioinformatics*, 9, 422.
- Daemen, A., *et al.* (2009) A kernel-based integration of genome-wide data for clinical decision support, *Genome Med*, 1, 39.
- Fan, X., *et al.* (2010) DNA microarrays are predictive of cancer prognosis: a re-evaluation, *Clin Cancer Res*, 16, 629-636.
- Furnari, F.B., *et al.* (2007) Malignant astrocytic glioma: genetics, biology, and paths to treatment, *Genes Dev*, 21, 2683-2710.
- Golub, T.R., *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 286, 531-537.
- Jansen, R., *et al.* (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data, *Science*, 302, 449-453.
- Lanckriet, G.R., *et al.* (2004) A statistical framework for genomic data fusion, *Bioinformatics*, 20, 2626-2635.
- Ohn, J.H., Kim, J. and Kim, J.H. (2007) Genomic characterization of perturbation sensitivity, *Bioinformatics*, 23, i354-358.
- Qiu, J. and Noble, W.S. (2008) Predicting co-complexed protein pairs from heterogeneous data, *PLoS Comput Biol*, 4, e1000054.
- Roepman, P., *et al.* (2005) An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas, *Nat Genet*, 37, 182-186.
- Sadikovic, B., *et al.* (2008) In vitro analysis of integrated global high-resolution DNA methylation profiling with genomic imbalance and gene expression in osteosarcoma, *PLoS One*, 3, e2834.
- Saxena, A., Robertson, J.T. and Ali, I.U. (1996) Abnormalities of p16, p15 and CDK4 genes in recurrent malignant astrocytomas, *Oncogene*, 13, 661-664.
- Segal, E., *et al.* (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data, *Nat Genet*, 34, 166-176.
- Shin, H., Lisewski, A.M. and Lichtarge, O. (2007) Graph sharpening plus graph integration: a synergy that improves protein functional classification, *Bioinformatics*, 23, 3217-3224.

References (cont'd)

- Spellman, P.T., *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Mol Biol Cell*, 9, 3273-3297.
- TCGA Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways, *Nature*, 455, 1061-1068.
- Tsuda, K., Shin, H. and Scholkopf, B. (2005) Fast protein classification with multiple networks, *Bioinformatics*, 21 Suppl 2, ii59-65.
- van 't Veer, L.J., *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer, *Nature*, 415, 530-536.
- Wu, C.C., *et al.* (2010) Prediction of human functional genetic networks from heterogeneous data using RVM-based ensemble learning, *Bioinformatics*, 26, 807-813.

- Sadikovic, B., *et al.* (2008) In vitro analysis of integrated global high-resolution DNA methylation profiling with genomic imbalance and gene expression in osteosarcoma, *PLoS One*, 3, e2834.
- Saxena, A., Robertson, J.T. and Ali, I.U. (1996) Abnormalities of p16, p15 and CDK4 genes in recurrent malignant astrocytomas, *Oncogene*, 13, 661-664.
- Segal, E., *et al.* (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data, *Nat Genet*, 34, 166-176.
- Shin, H., Lisewski, A.M. and Lichtarge, O. (2007) Graph sharpening plus graph integration: a synergy that improves protein functional classification, *Bioinformatics*, 23, 3217-3224.
- Spellman, P.T., *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Mol Biol Cell*, 9, 3273-3297.
- TCGA Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways, *Nature*, 455, 1061-1068.
- Tsuda, K., Shin, H. and Scholkopf, B. (2005) Fast protein classification with multiple networks, *Bioinformatics*, 21 Suppl 2, ii59-65.
- van 't Veer, L.J., *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer, *Nature*, 415, 530-536.
- Wu, C.C., *et al.* (2010) Prediction of human functional genetic networks from heterogeneous data using RVM-based ensemble learning, *Bioinformatics*, 26, 807-813.