Heterogeneous Multiple Genomic Data Integration for Translational Bioinformatics: the TCGA

Hyunjung (Helen) Shin

Dept. of Industrial & Information Systems Engineering
Ajou University, Korea

shin@ajou.ac.kr
http://www.alphaminers.net
http://www.kyb.tuebingen.mpg.de/~shin

Abstract

Motivation: Cancer is a complex disease, which is can be dysregulated through multiple mechanisms. In the past several years, DNA microarrays have been widely used for the classification of tumor subtypes or clinical outcomes for the diagnosis, treatment or prognosis of cancer. However, no single level of genomic data fully elucidates tumour behavior since there are many exceptional variations within or between levels in biological system such as copy number variants, DNA methylation, alternative splicing, miRNA regulation, post translational modification, *etc*.

Results: In the present study, the integrated framework has been proposed for the classification of several clinical outcomes in different cancer types, glioblastoma multiforme and ovarian cancer, using multi-layers of genomic data: copy number alteration; DNA methylation; gene expression; miRNA expression. By the empirical comparison on heterogeneous genomic data, our results showed that the level of contribution from each genomic data to various cancer clinical outcomes was relatively different as either structural changes or functional changes. However, through multi-level genomic data integration approach, our results indicate that the integration with multi-layers of genomic data is better for elucidating the cancer clinical outcomes than the model with only single level of genomic data. With abundance in multi-layers of genomic data and clinical data from many types of cancer in the near future, our proposed integrative framework will be valuable for better understanding the underlying tumor behavior, leading to more effective screening strategies and therapeutic targets.

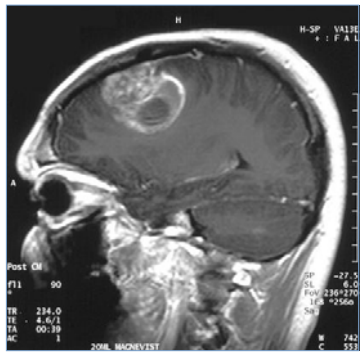
Outline

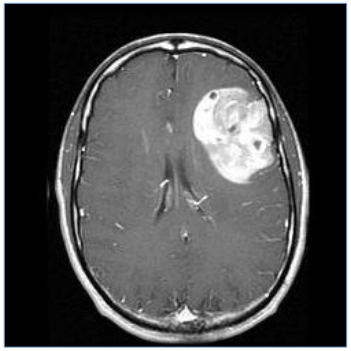
Introduction

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**



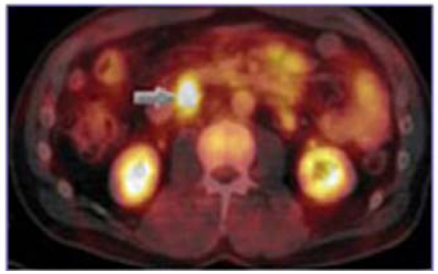


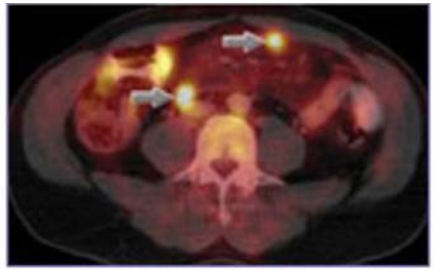
Serous Cystadenocarcinoma

Ovarian cancer (OV)

- One of the most common gynecologic malignancies
- 5th leading cause of cancer mortality in women in the United States

Jemal, et al. Cancer statistics, 2009c





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, object, and systematic cancer classification**

The studies about molecular-based classification of cancer subtypes or clinical outcomes using microarray are getting increased

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

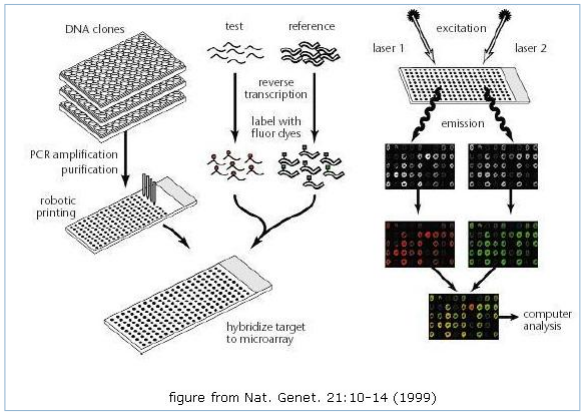


figure from Nat. Genet. 21:10-14 (1999)

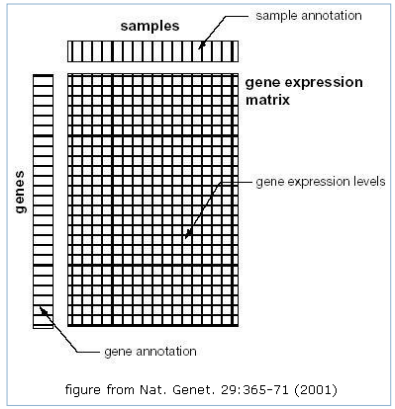
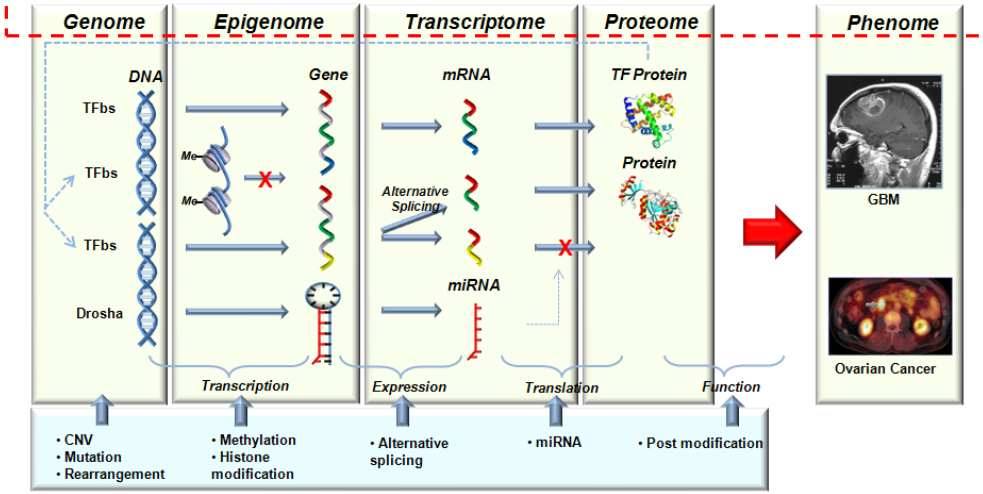


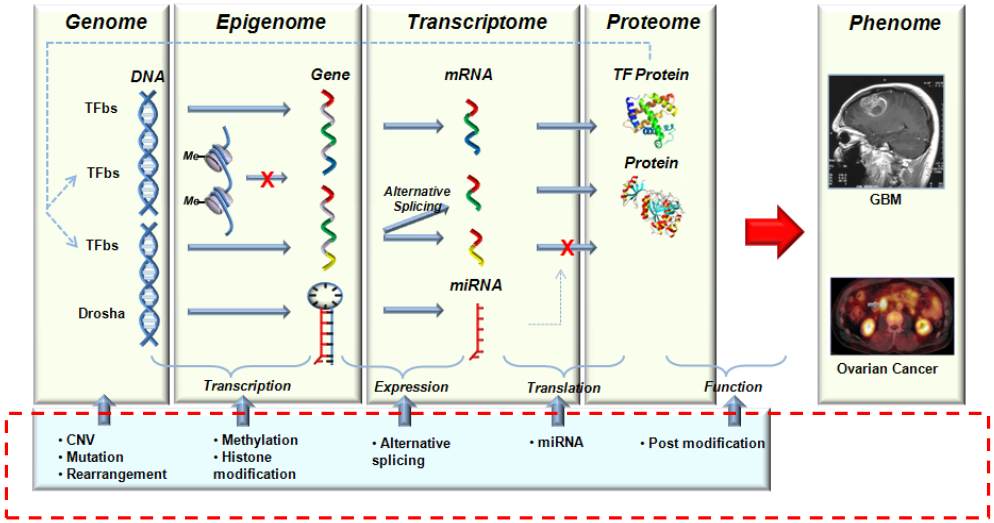
figure from Nat. Genet. 29:365-71 (2001)

Multi-layers of Genomic Data in Biological System

There are multiple levels in biological system !



Multi-layers of Genomic Data in Biological System



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

Multiple Mechanisms in Cancer

Cancer can be dysregulated through multiple mechanisms

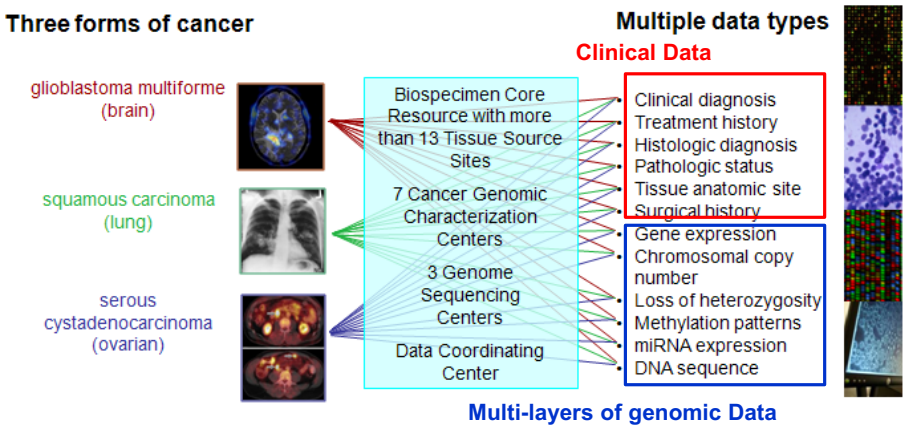
- Mutations in the coding and non-coding sequences
- Changes in the DNA structure and copy number
- Modifications to the DNA and histones

These changes can lead to alterations in

- Transcription
- Translation
- Post-translational modification
- Ultimately gene and protein function

The Cancer Genome Atlas (TCGA)

Connecting multiple sources, experiments, and data types



Motivation

Genomic data comparison

- **Which genomic data is more informative?**

Genomic data integration

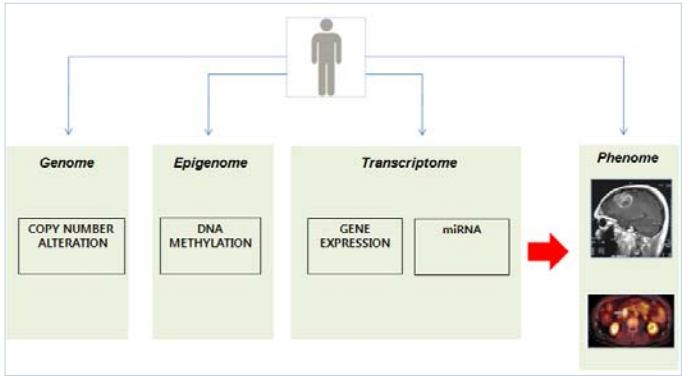
- Increase the importance of **integration more than one source of genome-wide data**, such as genome, epigenome, transcriptome, and proteome
- **Different genomic data contain partly independent and partly complementary pieces of biological information**
- The current increase in the amount of available omics data emphasizes the **need for a methodological integration framework**

Purpose of the Study

Integrative molecular-based classification of cancer clinical outcomes

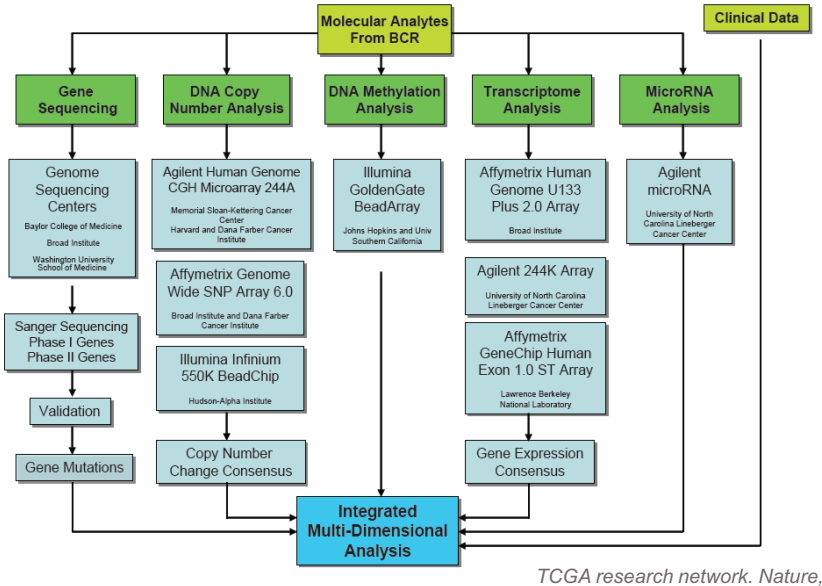
This study provide an **integrated methodological framework for analyzing multi-layers of genomic data**

- CNA, DNA methylation, gene expression, and miRNA



Data

TCGA Data



TCGA research network. Nature, 2008

Retrieving Multi-level Genomic Data

- Available raw and normalized different types of genomic data were retrieved from the TCGA data portal
- Cancer type
  - **Glioblastoma multiforme (GBM)**
  - **Serous cystadenocarcinoma (OV)**
- Size
  - About 500 GBs
- Databasing each level of genomic data for further analysis

Data Description

GBM

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

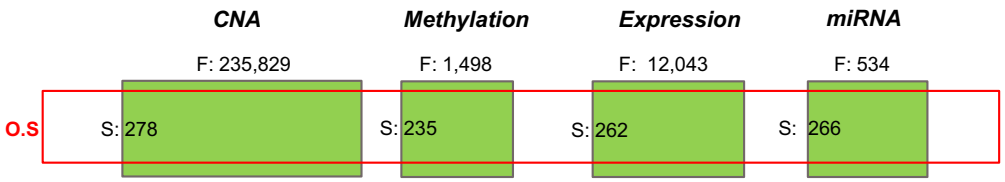
Data Description

OV

Data type	Platform	# Features
CNA	Agilent SurePrint G3 Human CGH Microarray Kit 1x1M	962,434
Methylation	Infinium humanmethylation27 BeadChip	27,578
Gene Expression	Affymetrix HT Human Genome U133 Array Plate Set	12,043
miRNA	Agilent Human miRNA Microarray Rel12.0	799

Data: Input

Select overlap samples among multi-level genomic dataset as an input



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

Data: Output Variables

Cancer type	Clinical outcomes	Binary classes	# Overlap samples* (*Neg/Pos*)
GBM	Survival	Short-term survival (survived less than nine m onths) vs. long-term survival (survived more than 24 months)	82 (54 / 28)
	Recurrence	Initial GBM (Initial diagnosis) vs. recurrent GBM (tumor recurrence)	159 (39 / 120)
OV	Survival	Short-term survival (survived less than three years) vs. long-term survival (survived more than three years)	348 (150 / 198)
	Stage	Early stage (T1 or T2) vs. late stage (T3 or T4)	503 (39 / 464)
	Grade	Low grade (G1 or G2) vs. high grade (G3 or G4)	496 (65 / 431)

*** Solid tumor samples from each type of cancer were only considered**

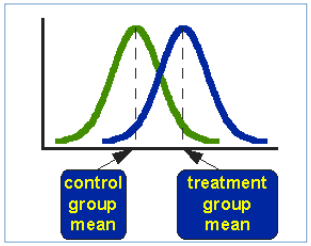
Methods

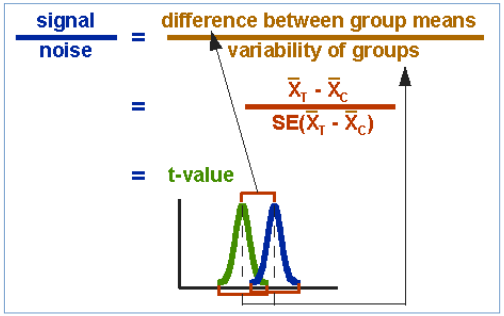
Data Preprocessing

Feature selection

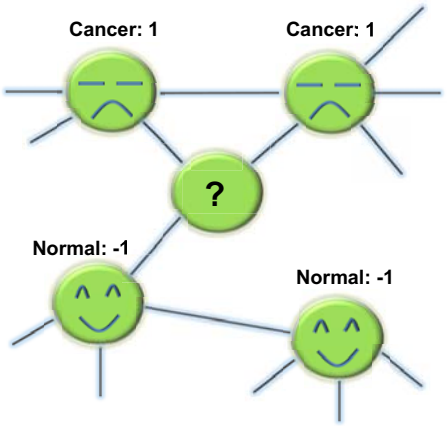
- Student t-test based feature selection method was used

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





Graph-based Semi-Supervised Learning (SSL)



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

The goal of SSL is to classify unlabeled sample 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

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

Final solution

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

Input for SSL: Weight Matrix (*W*)

Exp-weighted *K*-NN graphs

- Nodes *i, j* are connected by an edge if *i* is in *j*'s *K*-nearest-neighborhood or vice versa

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

- *d*: Euclidean distance
- Hyperparameter α controls the decay rate

Multi-level Genomic Data Integration

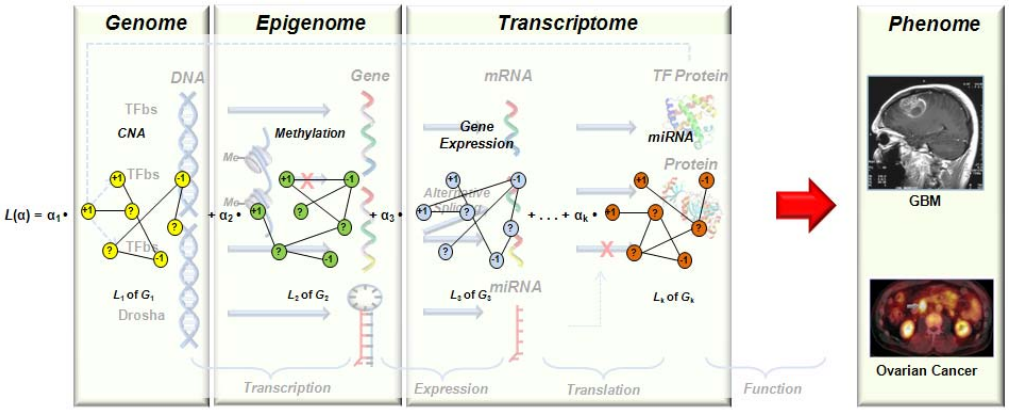
Multiple graphs from heterogeneous genomic data can be combined

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

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

Shin et al. Bioinformatics, 2007

Multi-level Genomic Data Integration

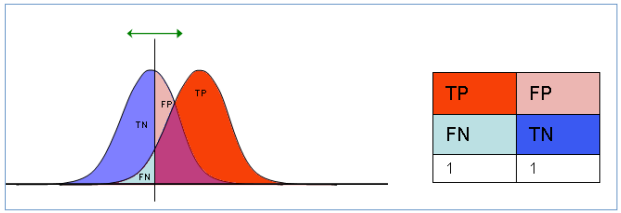


Model Parameter Selection

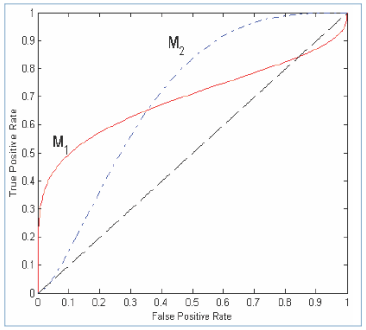
- **Parameters should be selected by user when learning with SSL**
  - *K*: *K*-NN
  - μ: SSL
- **Combination of parameters**
  - *K* = {3, 4, 5, 6, 7, 8, 9, 10, 20, 30}
  - μ = {0.001, 0.01, 0.05, 0.1, 0.2, 0.5, 0.7, 1.0, 10.0, 100.0, 1000.0}

Experiment: Measurement

- **AUC (Area Under the ROC Curve)**
- **TP1FP**
- **5-fold cross validation**



TP	FP
FN	TN
1	1



Results

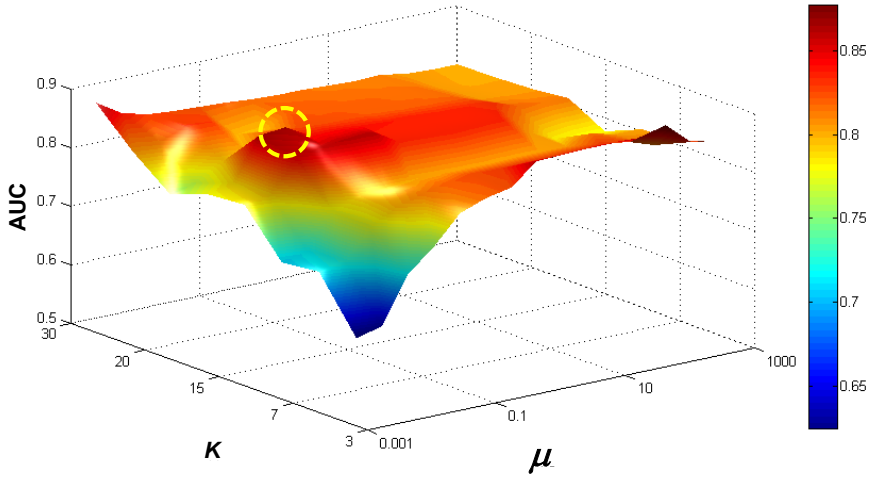
Preprocessing: Feature Selection Results

CNA

P_value <	Num of Features	BEST AUC	Avg AUC (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

Initial tumor vs. Recurrent tumor (GBM)

Model Parameter Selection



Survival in GBM: Gene expression (p < 0.001)

Best AUC Comparison: GBM

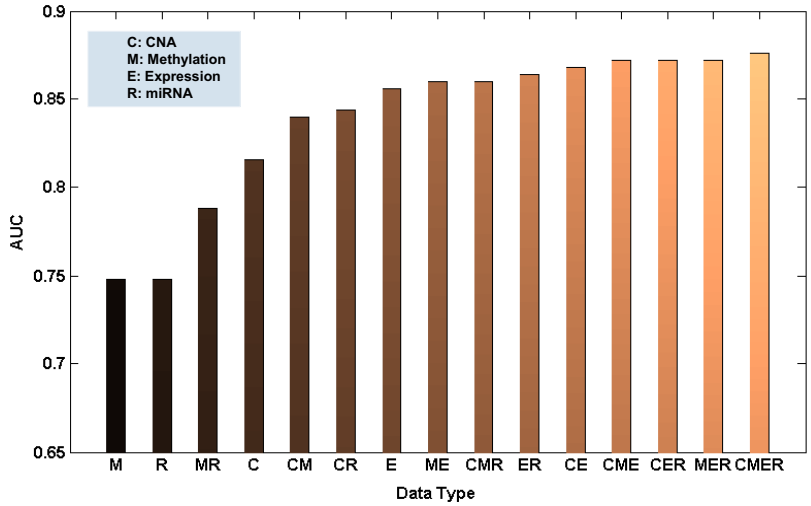
Outcome	Data type	AUC (*P-value*)	TP1FP (%)
Short-term survival **vs.** Long-term survival	CNA	0.8160 (2.19e-26)	0.30
	Methylation	0.7480 (1.19e-28)	0.60
	Gene Expression	0.8560 (1.22e-11)	0.72
	miRNA	0.7480 (1.07e-28)	0.40
	Multi-level data	**0.8760**	**0.80**
Initial tumor **vs.** Recurrent tumor	CNA	0.8131 (3.04e-04)	0.65
	Methylation	0.6774 (3.30e-33)	0.20
	Gene Expression	0.6667 (2.09e-34)	0.15
	miRNA	0.7226 (1.15e-33)	0.43
	Multi-level data	**0.8369**	**0.75**

Best AUC Comparison: OV

Outcome	Data type	AUC (*P-value*)	TP1FP (%)
Short-term survival **vs.** Long-term survival	CNA	0.6547 (1.24e-28)	0.17
	Methylation	0.7251 (1.34e-27)	0.14
	Gene Expression	0.7651 (8.96e-10)	0.26
	miRNA	0.6403 (1.24e-28)	0.17
	Multi-level data	**0.7867**	**0.40**
Early stage **vs.** Late stage	CNA	0.8767 (1.87e-05)	0.74
	Methylation	0.7149 (1.51e-28)	0.61
	Gene Expression	0.8332 (2.31e05)	0.53
	miRNA	0.7661 (1.39e-21)	0.78
	Multi-level data	**0.8932**	**0.80**
Low grade **vs.** High grade	CNA	0.8014 (3.43e-05)	0.37
	Methylation	0.8161 (4.63e-09)	**0.57**
	Gene Expression	0.7676 (2.59e-06)	0.39
	miRNA	0.6887 (9.61e-15)	0.16
	Multi-level data	**0.8678**	0.54

Integration Effect

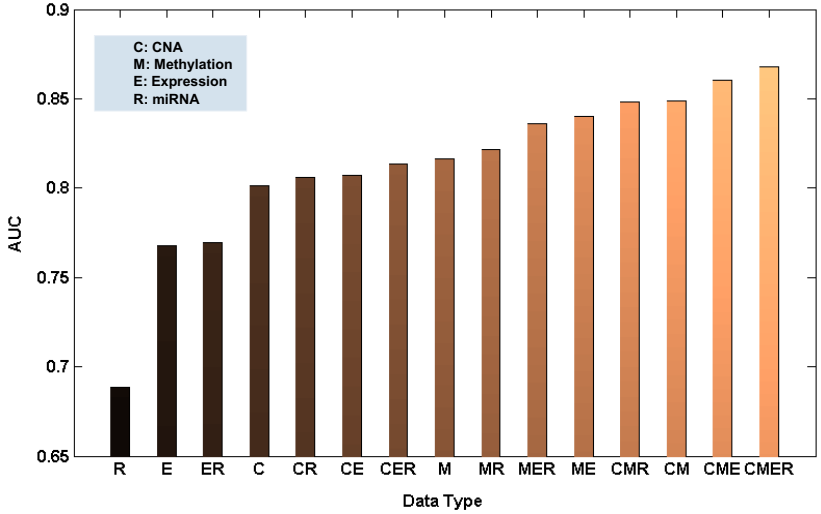
GBM: Survival



(A) GBM: Survival

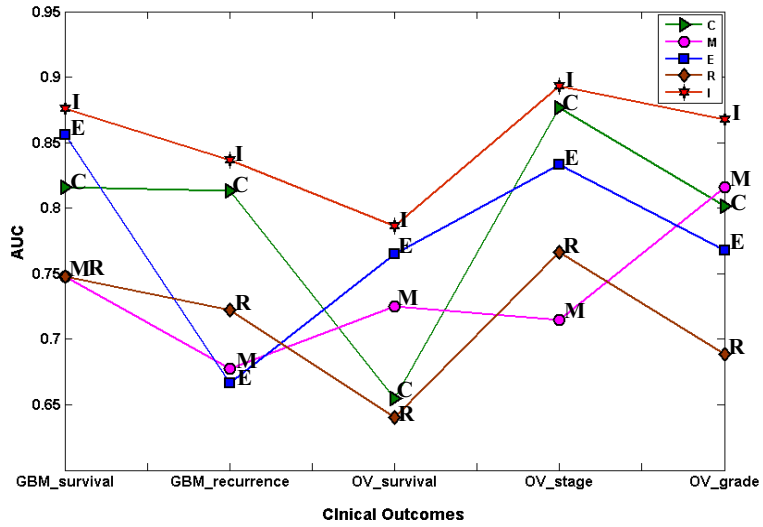
Integration Effect

OV: Grade



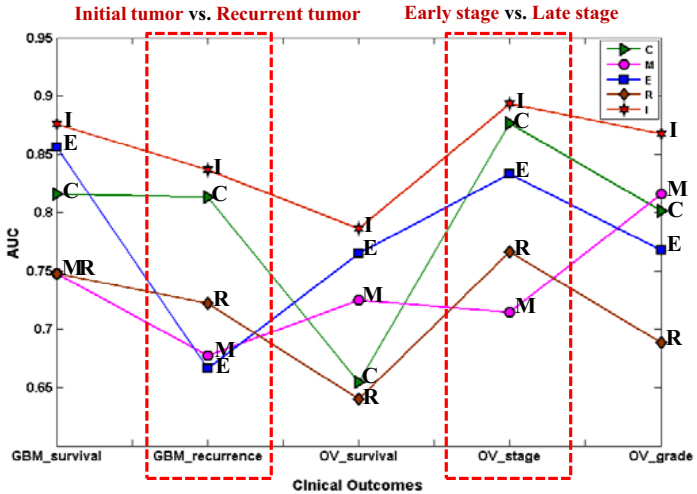
(B) OV: Grade

Biological Implication



Performance comparison of genomic data over the five sets of clinical outcome classification problem

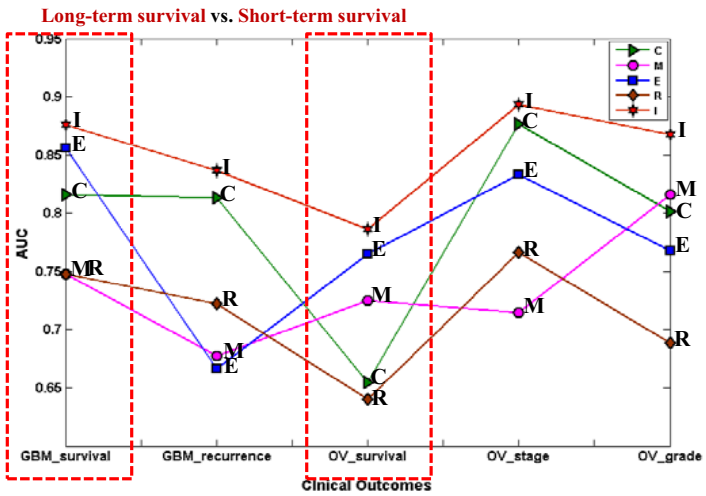
Biological Implication



Both problems concern the **structural changes in chromosome by the elapsed amount of time** since tumor initiation

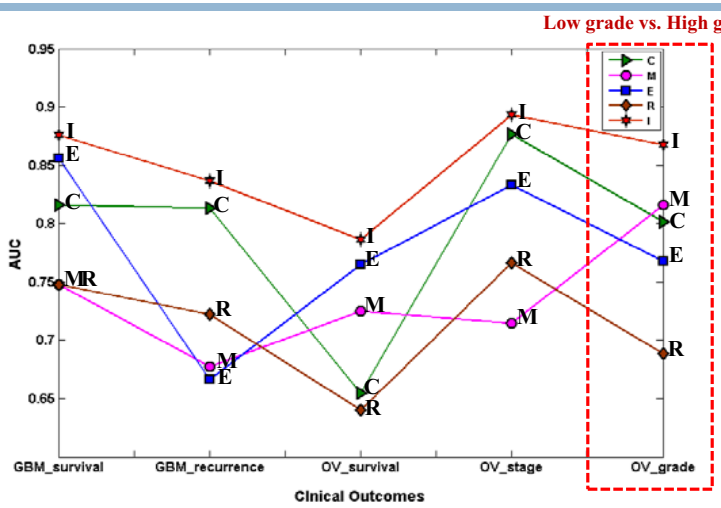
Therefore, **CNA data might have provided an appropriate information** for classifying the alternative clinical outcomes

Biological Implication (cont'd)



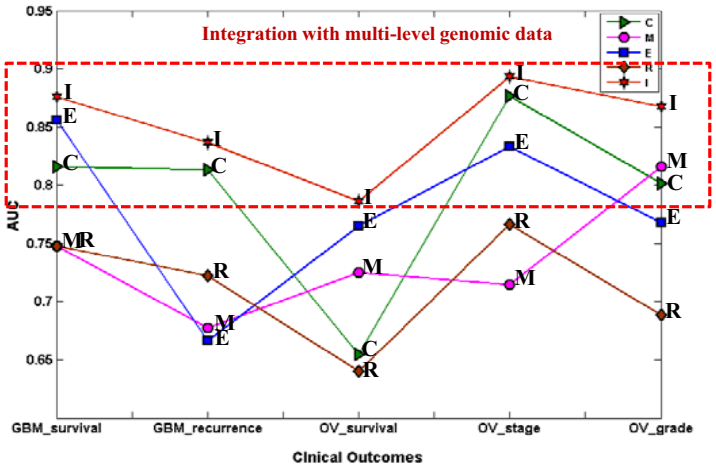
The strength of current malignant behavior of tumor is related to the functional changes of genes or proteins which can be detected by **gene expression data** in our experimental setting

Biological Implication (cont'd)



Despite lack of understanding of **epigenomic characteristics** in cancer, we could suggest the **structural changes may be worthy of further study**

Biological Implication (cont'd)



Integration of all genomic data sources can be helpful to unveil the relationship from genome to phenome

Conclusion

Conclusion

Cancer can be dysregulated through multiple mechanisms

The **integrative molecular-based classification of clinical outcomes** has been applied to two cancer types: GBM, OV

Genomic data comparison

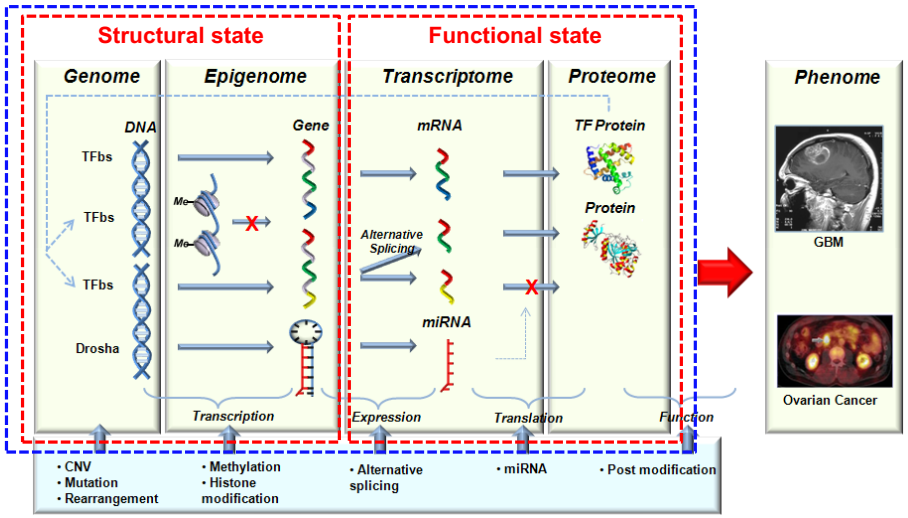
In order to provide a preliminary insight on the question:
Which genomic data is more informative?

- Various cancer types
- Various clinical outcomes

Genomic data integration

For both cancer types, **combining multi-level genomic dataset outperformed the models based on data from a single layer of biological information**

Conclusion



Our results emphasize the **need for an integrated methodological framework for analyzing multi-layers of genomic data for better understanding underlying tumor behavior**

The Second Phase of TCGA Project

Acknowledgements

Ju Han Kim
Professor and Chairman, Div. of Biomedical Informatics, Seoul National University
Director, Systems Biomedical Informatics Research Center

Dokyoon Kim
Ph.D. candidate, Seoul National University Biomedical Informatics (SNUBI), Div. of Biomedical Informatics, Seoul National University

Young Soo Song
Research professor , Dept. of Industrial and Information Systems Engineering, Ajou University

Kanghee Park
Ph.D. candidate, Dept. of Industrial Engineering, Ajou University

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)

References

•Bach, F., Lanckriet, G. and Jordan, M. (2004) Multiple kernel learning, conic duality, and the SMO algorithm, *In Proceedings of the Twenty-first I nternational Conference on Machine Learning (ICML), Banff, Canada, ACM Press*, pp. 6-13.
•Belkin, M. (2004) Regularization and Semi-supervised Learning on Large Graphs, *In Proceedings of the 17th Annual Conference on Learning T heory (COLT) 3120. Lecture Notes in Computer Science*, 624-638.
•Ben-Hur, A. and Noble, W.S. (2005) Kernel methods for predicting protein-protein interactions, *Bioinformatics*, **21 Suppl 1**, i38-46.
•Berchuck, A., *et al.* (2005) Patterns of gene expression that characterize long-term survival in advanced stage serous ovarian cancers, *Clin Can cer Res*, **11**, 3686-3696.
•Beroukhim, R., *et al.* (2010) The landscape of somatic copy-number alteration across human cancers, *Nature*, **463**, 899-905.
•Bild, A.H., *et al.* (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies, *Nature*, **439**, 353-357.
•Chapelle, O., Weston, J. and Scholkopf, B. (2003) Cluster kernels for semi-supervised learning, *Advances in Neural Information Processing Sys tems (NIPS)*, **15**, 585-592.
•Chin, L. and Gray, J.W. (2008) Translating insights from the cancer genome into clinical practice, *Nature*, **452**, 553-563.
•Chung, F.R.K. (1997) Spectral Graph Theory, *Number 92 in Regional Conference Series in Mathematics*.
•Demsar, J. (2006) Statistical comparisons of classifiers over multiple data sets, *Journal of Machine Learning Research*, **7**, 1-30.
•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**, 5 31-537.
•Gribskov, M. and Robinson, N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching, *Comput Chem*, **20**, 25-33.
•Hanash, S. (2004) Integrated global profiling of cancer, *Nat Rev Cancer*, **4**, 638-644.
•Huang, E., *et al.* (2003) Gene expression predictors of breast cancer outcomes, *Lancet*, **361**, 1590-1596.
•Jafari, P. and Azuaje, F. (2006) An assessment of recently published gene expression data analyses: reporting experimental design and statistic al factors, *BMC Med Inform Decis Mak*, **6**, 27.
•Jansen, R., *et al.* (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data, *Science*, **302**, 449-453.
•Jemal, A., *et al.* (2009) Cancer statistics, 2009, *CA Cancer J Clin*, **59**, 225-249.
•Kondor, I. and Lafferty, J. (2002) Diffusion kernels on graphs and other discrete structures, *In Sammut,C. and Hoffmann,A.G. (eds), Proceeding s of the Nineteenth International Conference on Machine Learning (ICML 2002), Sydney, Australia, Morgan Kaufmann*, pp. 315-322.
•Lanckriet, G.R., *et al.* (2004) A statistical framework for genomic data fusion, *Bioinformatics*, **20**, 2626-2635.
•Lu, J., *et al.* (2005) MicroRNA expression profiles classify human cancers, *Nature*, **435**, 834-838.
•Marko, N.F., *et al.* (2008) Genomic expression patterns distinguish long-term from short-term glioblastoma survivors: a preliminary feasibility stu dy, *Genomics*, **91**, 395-406.

References

•Mischel, P.S., Cloughesy, T.F. and Nelson, S.F. (2004) DNA-microarray analysis of brain cancer: molecular classification for therapy, *Nat Rev N eurosci*, **5**, 782-792.
•Myllykangas, S., *et al.* (2008) Classification of human cancers based on DNA copy number amplification modeling, *BMC Med Genomics*, **1**, 15.
•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 carcinoma s, *Nat Genet*, **37**, 182-186.
•Salcman, M. and Kaplan, R. (1991) Intracranial tumors in adults, *In : Salcman M (ed) Neurology of brain tumors. Williams & Wilkins, Baltimore*, 1339-1352.
•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 G enet*, **34**, 166-176.
•Shin, H., Lisewski, A.M. and Lichtarge, O. (2007) Graph sharpening plus graph integration: a synergy that improves protein functional classificati on, *Bioinformatics*, **23**, 3217-3224.
•Shin, H. and Tsuda, K. (2006) Prediction of Protein Function from Networks, *in Book: Semi-Supervised Learning, Edited by Olivier Chapelle, Ber nhard Sch ¨olkopf, Alexander Zien, MIT press*, **Chapter 20**, 339-352.
•Shridhar, V., *et al.* (2001) Genetic analysis of early- versus late-stage ovarian tumors, *Cancer Res*, **61**, 5895-5904.
•Spellman, P.T., *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray h ybridization, *Mol Biol Cell*, **9**, 3273-3297.
•TCGA Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways, *Nature*, **455**, 1061-106 8.
•Troyanskaya, O., *et al.* (2001) Missing value estimation methods for DNA microarrays, *Bioinformatics*, **17**, 520-525.
•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.
•Verhaak, R.G., *et al.* (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1, *Cancer Cell*, **17**, 98-110.
•Waldman, F.M., *et al.* (2000) Chromosomal alterations in ductal carcinomas in situ and their in situ recurrences, *J Natl Cancer Inst*, **92**, 313-320.
•Wu, C.C., *et al.* (2010) Prediction of human functional genetic networks from heterogeneous data using RVM-based ensemble learning, *Bioinfor matics*, **26**, 807-813.
•Zhou, D., *et al.* (2004) Learning with local and global consistency, *Advances in Neural Information Processing Systems (NIPS)*, **16**, 321-328.