Journal of Biomedical Informatics 45 (2012) 1191–1198

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Synergistic effect of different levels of genomic data for cancer clinical outcome prediction

Dokyoon Kim [a,b,1], Hyunjung Shin [c,*,1], Young Soo Song [a,c], Ju Han Kim [a,b,*]

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

ARTICLE INFO

Article history:
Received 31 October 2011
Accepted 19 July 2012
Available online 15 August 2012

Keywords:
Multi-layers of genomic data
Data integration
Clinical outcome prediction
Glioblastoma multiforme
Serous cystadenocarcinoma

ABSTRACT

There have been many attempts in cancer clinical-type classification by using a dataset from a number of molecular layers of biological system. Despite these efforts, however, it still remains difficult to elucidate the cancer phenotypes because the cancer genome is neither simple nor independent but rather complicated and dysregulated by multiple molecular mechanisms. Recently, heterogeneous types of data, generated from all molecular levels of 'omic' dimensions from genome to phenome, for instance, *copy number variants* at the genome level, *DNA methylation* at the epigenome level, and *gene expression* and *microRNA* at the transcriptome level, have become available. In this paper, we propose an integrated framework that uses multi-level genomic data for prediction of clinical outcomes in brain cancer (glioblastoma multiforme, GBM) and ovarian cancer (serous cystadenocarcinoma, OV). From empirical comparison results on individual genomic data, we provide some preliminary insights about which level of data is more informative to a given clinical-type classification problem and justify these perceptions with the corresponding biological implications for each type of cancer. For GBM, all clinical outcomes had a better the area under the curve (AUC) of receiver operating characteristic when integrating multi-layers of genomic data, 0.876 for survival to 0.832 for recurrence. Moreover, the better AUCs were achieved from the integration approach for all clinical outcomes in OV as well, ranging from 0.787 to 0.893. We found that the opportunity for success in prediction of clinical outcomes in cancer was increased when the prediction was based on the integration of multi-layers of genomic data. This study is expecting to improve comprehension of the molecular pathogenesis and underlying biology of both cancer types.

1. Introduction

An understanding of the molecular basis of cancer brings many benefits for predicting clinical outcomes of cancer and for determining the corresponding best treatment. Since cancer is related to alterations in the genes that control normal cell growth and death, molecular-based diagnostics are promising in that they may provide more opportunities for objective, precise, and systematic predictions on cancer. Data at the multiple molecular levels, generated from all levels of 'omic' dimensions from genome to phenome (Fig. 1), have recently become more available. At the genome level, copy number variants have attracted considerable attentions, since alterations of genomic DNA can be explored by expanding the scope of view to a larger region of the genome or to chromosomes. At the epigenome level, data from DNA methylation, which plays a crucial role in the control of gene activity, is of interest, while at the level of the transcriptome, gene expression and microRNA (miRNA) are the most representative datasets. DNA microarrays have already been widely used for the classification of tumor subtypes or clinical outcomes for the diagnosis, treatment, or prognosis of cancer for many years [3,13,15,19,34,47]. More recently, miRNA has become available for understanding the inhibition of expression on target mRNAs in gene regulatory networks.

There have been attempts at cancer classification using a set of miRNA, copy number alterations (CNAs), and DNA methylation [4,6,27,30,51]. Despite these efforts, however, it still remains difficult to elucidate the cancer phenotypes because the cancer genome is neither simple nor independent but rather complicated and dysregulated by multiple molecular mechanisms [9,18]. For example, the cancer genome is related to mutations in coding and non-coding sequences, changes in the DNA structure and copy number, DNA methylation and histone modification, and miRNA regulation. Those possibilities lead to many alternative forms of cause-and-result in

* Corresponding authors. Address: Seoul National University Biomedical Informatics (SNUBI), Div. of Biomedical Informatics, Seoul National University College of Medicine, Seoul 110-799, Republic of Korea. Fax: +82 31 219 1610 (J.H. Kim), fax: +82 2 747 8928 (H. Shin).
E-mail addresses: shin@ajou.ac.kr (H. Shin), juhan@snu.ac.kr (J.H. Kim).
[1] These authors are contributed equally to this work.

1532-0464/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2012.07.008

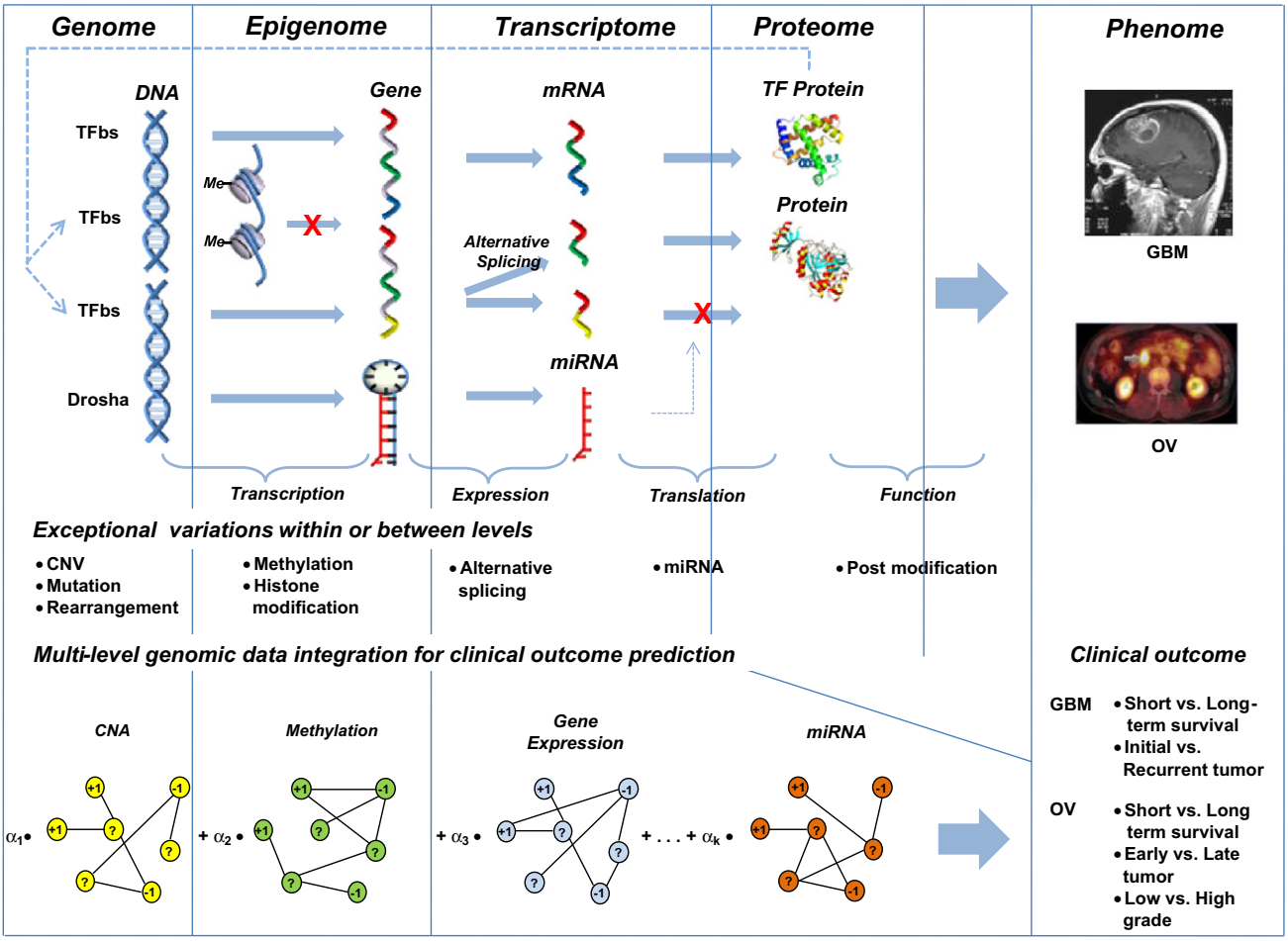

Fig. 1. Multi-layers of genomic data in biological system from genome, epigenome, transcriptome and proteom to phenome. There are many exceptional variations within or between levels: copy number variant (CNV), sequence mutation, and genomic rearrangement in genome level; DNA methylation and histone modification in epigenome level; alternative splicing and miRNA regulation in transcriptome level; post translational modification in proteom level. Multiple graphs given from different genomic levels are integrated into one by finding an optimum value of the linear combination coefficient α_k for the individual graph. TF, transcription factor; TFbs, transcription factor binding site; Me, methylation.

transcription, translation, post-translational modification, and eventually, gene and protein functions (Fig. 1). Thus, no single level of genomic data will be sufficient to comprise all of the information in the mechanism, and hence, a consideration of the layered process of biological systems through incorporation of multiple levels of genomic data will provide a much more reasonable prediction of cancer phenotypes.

The Cancer Genome Atlas (TCGA) is a collaborative initiative to improve understanding of cancer using existing large-scale whole-genome technologies. The TCGA research network lately published two notable papers on glioblastoma patients concerning an interim analysis of DNA sequencing, copy number, gene expression, and DNA methylation data [45], and discovery of links between cancer subtypes and different neural lineages with gene expression [48]. While the TCGA opens many opportunities to researchers to deepen the knowledge of the molecular basis of cancer [16,24], it is particularly important to access multiple data sources as we propose here.

In this paper, we propose an integrated framework that uses multi-level genomic data sources for the molecular-based classification of clinical outcomes in brain cancer (glioblastoma multiforme; GBM) and ovarian cancer (serous cystadenocarcinoma; OV). GBM is the most common and aggressive primary brain tumor in adults [14], and notorious for its tendency to recur [37]. Despite recent advances in the molecular pathology of GBM, the underling molecular mechanisms associated with clinical outcome are still poorly understood [14,38]. OV is one of the most common gynecological malignancies, and is the 5th leading cause of cancer mortality in women in the United States [21]. Understanding the molecular pathogenesis and underlying biology for both types of cancer is expected to provide guidance for improved prognostic indicators and effective therapies.

In computational biology, this work will be a pioneering attempt to predict the cancer phenotype based on the underlying complex biological mechanisms. From individual TCGA data sources, we conducted empirical comparisons at each level of genomic data; to deduce possible biological implications based on the results of the relative contribution of each piece of data to increase prediction accuracy. We show that accuracy of prediction increases because of incorporation of information fused over heterogeneous biological data sources, providing an enhanced global view on cancer mechanisms.

Several approaches to multiple data integration have been applied to protein function prediction such as the kernel-based integration framework [2,22,33], the Relevance Vector Machine (RVM) approach [52], and a Bayesian model [20]. In recent years, we developed an integration method of protein networks based on graph-based semi supervised learning (SSL) [41,46]. SSL is a halfway method between unsupervised learning and supervised learning, which takes an advantage of both unlabeled data and labeled

data. Here, we focus on graph-based SSL due to their solid mathematical background, model visualization, sparseness properties as well as the close relationship with kernel methods. In terms of integration issues, learning with other methods such as support vector machine (SVM) and Markov random field (MRF) may not be finished in a reasonable time when the data is large-scale, e.g. TCGA dataset (Appendix A). Although, any of the above mentioned methods could be used to implement the proposed idea, the graph-based SSL is employed in this study, taking advantage of computational efficiency and representational ease for the biological system. The learning time of graph-based SSL is nearly linear with the number of graph edges, which in most biological networks is few, while the accuracy remains comparable to the kernel-based methods that suffer from the relative disadvantage of a longer learning time [42,46]. In addition, the interpretation of biological phenomena can be improved because of the graph data structure [31,40,44], which naturally fits into the graph based SSL.

There have been several studies for integrative analysis of different levels of genomic data such as pairs of 'CNA-gene expression' [7,29,39], 'miRNA-gene expression' [23,32,50], and 'methylation-gene expression' [12,25]. These approaches are generally based on regression or correlation analysis, which fit into combining two levels of genomic data and are not designed to accommodate another level of genomic data. Thus, these methods have a difficulty to integrate more than three types of genomic data because of non-scalability. On the other hand, many integrative approaches have been proposed to combine more than three types of genomic data through the gene-based integration; that is, heterogeneous genomic data from different sources were analyzed sequentially, then each genomic data mapped into the gene level for the integration [26,35,36]. However, information loss might occur by such gene-based integration from the different levels of genomic data because there is no guarantee to be a simple one-to-one relationship between each feature of multi-layers of genomic data and a specific gene such as relations of exon-gene, miRNA-gene, CNA-gene, and methylation-gene. Therefore, here we propose a data integration framework for different levels of genomic data, not only with a scalable formalism to extend another level of genomic data but preserving level-specific properties from the multi-layers and heterogeneous genomic data.

The manuscript is organized as follows. Data description and methods for graph-based SSL and the integration approach are explained in Section 2. In Section 3, experimental results in GBM and OV are provided to demonstrate the validity and effectiveness of our integrative approach. Finally, we discuss the meaning of our study and future works in the last section.

2. Materials and methods

2.1. Data

Datasets were retrieved from the TCGA data portal (http://www.tcga-data.nci.nih.gov/) (Supplementary Table 2). Table 1 shows the data description of the multi-level genomic datasets in GBM and OV. CNA belongs to the genome level, methylation to the epigenome level, gene expression and miRNA to the level of the transcriptome. The fourth column shows the number of features for each type of genomic data. Five sets of binary classification problems were set using the phenotype information from patients depending on the types of clinical outcomes (Table 2). Using the clinical outcome from GBM, the two sets of problems are defined: (1) *short-term or long-term survival* and (2) *initial or recurrent tumor*. In the classification of *short-term or long-term survival*, 'short-term' represents the samples from patients who survived less than 9 months, whereas 'long-term' means samples derived from patients who survived longer than 24 months [28]. In the classification of *initial or recurrent tumor*, 'initial tumor' indicates samples from surgical resections with no pretreatment history, while samples from secondary surgeries for tumor recurrence are defined as 'recurrent tumor.' Similarly, the remaining three sets of classifications are defined using OV clinical outcomes, which are as follows: (3) *early stage (T1–T2) or late stage (T3–T4)*, (4) *low grade (G1–G2) or high grade (G3–G4)*, and (5) *short-term (<3 y) or long-term (⩾3 y) survival* [5]. The last column of Table 2 summarizes the number of available (positive/negative) samples for each of these problems.

2.2. Clinical outcome classification

We used a graph-based semi-supervised learning as a classification algorithm, which is a halfway learning scheme between supervised and unsupervised learning [1,8,53,54]. If two patients' samples were more closely related than to others, we assumed that the clinical outcomes of those two patients were more likely to be similar. In other words, clinical outcome prediction can be done by considering relationships between patient samples. A natural method of analyzing relationships between samples is a graph, where nodes depict patient samples and edges represent their possible relations. Fig. 2 presents a cartoon graph of patient samples. An annotated sample is labeled either by '−1' or '1', indicating the two possible clinical outcomes, either 'normal' or 'cancer'. To predict the label of the unannotated sample '?', the edges connected from/to the sample play an important role in influencing propagation between the sample and its neighbors. This idea can be easily formulated using graph-based SSL [53]. Edges represent relations, more specifically similarities between samples that may be extracted from different genomic sources of CNA, methylation, gene expression, miRNA, etc. Different data sources produce different graphs. Clinical outcome prediction can benefit by integrating diverse graphs from diverse genomic data sources, rather than relying only on single sources that may have possible limitations, (i.e. incomplete information and noise). When data sources are presented as a graph form, combining multiple data sources can be done by employing a graph integration method [41,42,46].

Table 1
Data description.

Cancer type	Data type	Platform	# Features (d)
GBM	CNA	Agilent Human Genome CGH Microarray 244A	235,829
	Methylation	Illumina DNA Methylation OMA003 Cancer Panel 1	1498
	Gene expression	Affymetrix HT Human Genome U133 Array Plate Set	12,043
	miRNA	Agilent 8 × 15 K Human miRNA-specific microarray	534
OV	CNA	Agilent SurePrint G3 Human CGH Microarray Kit 1 × 1 M	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

Table 2
Clinical outcomes.

Cancer type	Clinical outcome	# Samples (n)[a] (Neg/Pos)
GBM	Short-term survival (survived less than 9 months) vs. long-term survival (survived longer than 24 months)	82 (54/28)
	Initial tumor (initial diagnosis) vs. recurrent tumor (tumor recurrence)	159 (39/120)
OV	Short-term survival (survived less than 3 years) vs. long-term survival (survived longer than 3 years)	348 (150/198)
	Early stage (T1 or T2) vs. late stage (T3 or T4)	503 (39/464)
	Low grade (G1 or G2) vs. high grade (G3 or G4)	496 (65/431)

[a] Solid tumor samples from each type of cancer were only considered.

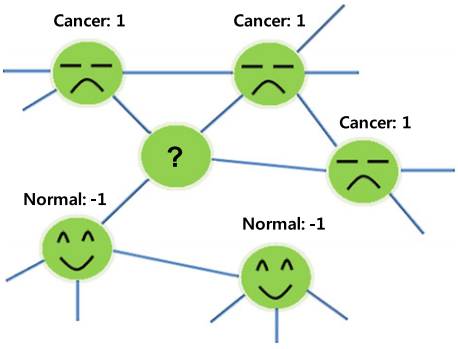

Fig. 2. A graph model of relationships between patient samples. Nodes represent patient samples and edges depict relations between samples. An annotated sample is labeled either by −1 or +1. In this example, the negative labels indicate samples from 'normal' patients. On the contrary, the positive labels indicate the samples from 'cancer' patients. The clinical outcome of the unannotated sample marked as '?' is predicted by employing graph-based semi-supervised learning.

2.2.1. Graph-based semi-supervised learning

In the graph-based SSL algorithm [53], a sample x_i ($i = 1,\ldots,n$) is represented as a node i in a graph, and the relationship between samples is represented by an edge. The edge strength from each node j to each other node i is encoded in element w_{ij} of a $n \times n$ symmetric weight matrix W. A Gaussian function of Euclidean distance between samples, with length scale hyperparameter σ, is used to specify connection strength[2]:

$$w_{ij} = \begin{cases} \exp\left(-\frac{(x_i - x_j)^T (x_i - x_j)}{\sigma^2}\right) & \text{if } i \sim j, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Nodes i, j are connected by an edge if i is in j's k-nearest-neighborhood or vice versa. Therefore, nearby samples in Euclidean spaces are assigned large edge weights. The labeled nodes have labels $y_l \in \{-1, 1\}$, while the unlabeled nodes have zeros $y_u = 0$. SSL will output an n-dimensional real-valued vector $f = [f_l^T f_u^T]^T = (f_1, \ldots, f_l, f_{l+1}, \ldots, f_{n=l+u})^T$, which can be thresholded to make label predictions on $f_{l=1},\ldots,f_n$ after learning. We require (a) the score f_i should not be too different from the scores of adjacent vertices, and (b) the scores should be close to the given label y_i in training nodes. One can obtain f by minimizing the following quadratic functional [1,8,53]:

$$\sum_{i=1}^{l} (f_i - y_i)^2 + c\sum_{i=l+1}^{n} f_i^2 + \mu \sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2 \quad (2)$$

[2] x_i and x_j are d-dimensional column vector and $(\cdot)^T$ means the transpose of $(\cdot)$.

The first term corresponds to the *loss function* in terms of condition (b), and the third term describes the *smoothness* of the scores in terms of condition (a). The parameter μ trades off loss vs. smoothness. The second term is a regularization term to keep the scores of unlabeled nodes in a reasonable range. Alternative choices of *smoothness* and *loss functions* can be found in Chapelle et al. [8]. From later on, we focus on the special case $c = 1$ [53]. Then, the three terms reduce to the following two terms in matrix notation,

$$\min_f (f - y)^T (f - y) + \mu f^T L f \quad (3)$$

where $y = (y_1,\ldots,y_l, 0,\ldots,0)^T$, and the matrix L, called the graph Laplacian matrix [10], is defined as $L = D - W$ where $D = \mathrm{diag}(d_i)$, $d_i = \sum_j w_{ij}$. The parameter μ trades off loss vs. smoothness. The solution of this problem is obtained as

$$f = (I + \mu L)^{-1} y \quad (4)$$

where I is the identity matrix.

2.3. Integration of multi-level genomic data sources

From multi-level genomic data sources, multiple graphs are generated. Information from each graph is regarded as partially independent from and partly complementary to others. Therefore, it is not accurate enough to elucidate phenotype using only a single genomic data source belonging to a specific single layer. Reliability may be enhanced by integrating all available information sources using the method proposed by Tsuda et al. [46] and Shin and Tsuda [42], which has been re-validated on the extended problem of protein function prediction [41]. According to the method, the integration of multiple graphs is used to find an optimum value of the linear combination coefficient for the individual graphs. This corresponds to finding the combination coefficients α for the individual Laplacians of the following mathematical formulation:

$$\min_{\alpha} y^T \left(I + \sum_{k=1}^{K} \alpha_k L_k\right)^{-1} y, \quad \sum_k \alpha_k \leqslant \mu \quad (5)$$

where K is the number of graphs (data sources) and L_k is the corresponding graph-Laplacian of graph G_k. One can perceive the formulation by synchronizing the schematic descriptions shown in Fig. 1. Similar to the output prediction for single graphs, the solution is obtained by

$$f = \left(I + \sum_{k=1}^{K} \alpha_k L_k\right)^{-1} y \quad (6)$$

3. Results

From the multi-levels of genomic data available from TCGA, the following five sets of clinical outcome classifications were defined: for GBM, (1) short-term or long-term survival and (2) initial or recurrent tumor and for OV (3) early or late stage; (4) low or high grade; (5) short-term or long-term survival. For the five sets of binary classification problems, the proposed approach, *prediction from integration of multi-level genomic data sources*, was compared with the four *individual predictions* obtained from CNA, methylation, gene expression, and miRNA. For each problem, we calculated the area under the curve (AUC) of receiver operating characteristic (ROC) [17] and the true positive rate yielding an 1% false positive rate (TP1FP) as performance measurements [22]. The dataset was randomly splited so that 80% of the samples are assigned to the training set and 20% to the validation set to find the model parameters. The overall performance was measured through three times of fivefold cross validation. The Wilcoxon signed-rank test was used to validate the significance of the difference in performance for all the combinations of the comparisons [11]. Details on the

Table 3
AUC results on GBM clinical outcomes.

Clinical outcome	Data type	AUC (p-value[a])	TP1FP
Short-term survival vs. long-term survival	CNA	0.8160 (2.19e−26)	0.30
	Methylation	0.7408 (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**

[a] The p-value of the pairwise Wilcoxon signed-rank test in AUCs between the multi-level integration approach and the single data approaches.

model parameter selection are available in Supplementary methods.

3.1. Glioblastoma multiforme

Table 3 shows the AUC performance on the two sets of classifications of GBM clinical outcomes. The AUCs of the four *individual sources* (CNA, methylation, gene expression, and miRNA) are shown in the first four rows and the AUC of the proposed approach—*integration with multi-level genomic data sources* is shown in the fifth row. For the short-term survival vs. long-term survival classification, the SSL with gene expression data performed best with 0.8560 AUC (underlined) among the four genomic data sources, and CNA data showed comparable performance with an AUC of 0.8160. However, none of the AUC values from single data sources could outperform the AUC of 0.8760 (boldface) generated by the multi-level genomic data. The p-values of the pairwise comparisons in AUC between the proposed approach and the single data approaches demonstrate that a statistically significant difference exists in performance. The superior performance of the integration approach is also found in terms of the value of the TP1FP—of 0.8, the highest among the five. Furthermore, for the initial tumor vs. recurrent tumor classification, the SSL with CNA data showed the best performance with an AUC value of 0.8131 (underlined) among the four individual data sources, but again the performance of the integration approach was superior to that of the best individual approach with an AUC of 0.8369 and TP1FP of 0.75 (boldface) was superior to all of the individual approaches.

Table 4
AUC results on OV clinical outcomes.

Clinical outcome	Data type	AUC (p-value[a])	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 (2.31e−05)	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

[a] The p-value of the pairwise Wilcoxon signed-rank test in AUCs between the multi-level integration approach and the single data approaches.

3.2. Serous cystadenocarcinoma

Table 4 shows the AUC performance of the three sets of classification problems of OV clinical outcomes. For the short-term survival vs. long-term survival classification, the SSL with gene expression data performed best with an AUC of 0.7651 when compared with the other single genomic data sources. Note that a similar result was obtained for GBM, which might imply that the information from gene expression data plays a critical role in the classification of short-term vs. long-term survival. Among the four individual sources, the best performing data differed for each classification; for the early stage vs. late stage classification, the prediction from CNA data showed the best performance (0.8767 AUC) while for the low grade vs. high grade classification, methylation data performed best (0.8161 AUC). However, the AUC of the proposed approach consistently outperforms the individual best for all of the three sets of problems, attaining values of 0.7867, 0.8932, and 0.8678. There is a slight degradation of the proposed approach in TP1FP for the low grade vs. high grade classification, but the magnitude in difference is negligible (0.57 in methylation data vs. 0.54 in multi-level integrated data).

3.3. Integration effect

We found that the integration of various genomic data sources increases the performance of clinical outcome prediction. As a next step, we investigated the effect of integration by considering all possible combinations of the four multi-level genomic data sources. Fig. 3 shows a gradual increase in AUC by integration: for the short-term survival vs. long-term survival classification of GBM (Fig. 3A) and for the low grade vs. high grade classification of OV (Fig. 3B), where C stands for CNA, M for methylation, E for gene expression, and R for miRNA, and the combinations are represented as MR, CMR, and so on. AUC consistently increases as more data sources are added to the combination in both GBM and OV. In Fig. 3A, for instance, AUC increases in the order of mass of combinations, C < CR < CMR < CMER. These findings suggest that biological information may be fused to various data sources from different genomic levels; therefore, integration of those independent or complementary pieces of information may elevate the opportunity of success in prediction of clinical outcomes in cancer.

3.4. Biological implication

On the basis of the results of our computational experiments, some biological and clinical implications may be cautiously drawn. Fig. 4 illustrates the following observations that show the level of contributions of multi-level genomic data sources for the five classification problems. As of yet, there has been no clear-cut definition on boundaries between the different genomic levels; but, it is naturally conceived that the *structural changes* in the chromosome or chromatin will lead to the changes on data sources obtained from the genome or the epigenome level (CNA and methylation in our experiment) before the influence reaches to the transcriptome level. On the other hand, the *functional changes* caused by the by-products of DNA will be more directly related to the changes on data sources generated from the transcriptome level (mRNA and miRNA in our experiment).

First, the CNA data performed best in the initial vs. recurrent tumor classification in GBM and the early vs. late stage classification in OV. Both problems concern the structural changes in chromosome by the elapsed amount of time since tumor initiation [43,49]. Therefore CNA data might have provided appropriate information for classifying the alternative clinical outcomes.

Second, the performance of the gene expression data was superior to those of others in the short-term vs. long-term survival

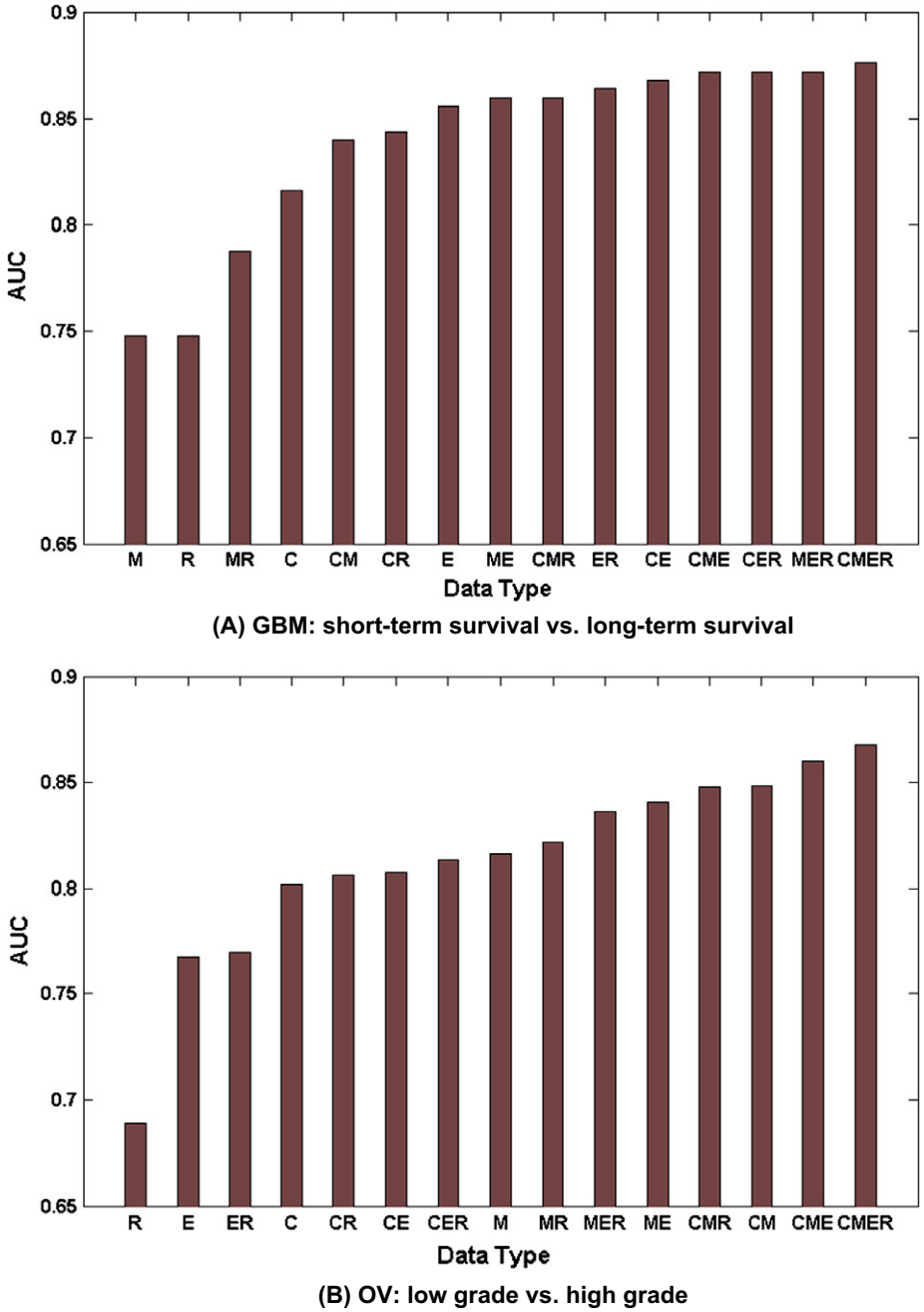

Fig. 3. Gradual increase in AUC by integration: C stands for CNA, M for methylation, E for gene expression, and R for miRNA, and the combinations are represented as MR, CMR, and so on. (A) The short-term vs. long-term survival classification of GBM. (B) The low vs. high grade classification of OV.

classification in both GBM and OV. The strength of the current malignant behavior of the tumor is related to the functional changes of genes or proteins [3] which can be detected by gene expression data in our experimental setting. Interestingly, the gene expression data performed as good as the other three dataset (CMR), and almost as good as the full dataset (CMER) (Fig. 3A). This calls for additional bioinformatical analyses. One intriguing possibility suggests that the same genomic loci contribute clinical information in more than one domain – the same genes that change in their copy number and methylation patterns also present predictive powers based on mRNA expression levels.

Third, the methylation data performed best for the low vs. high grade classification of OV. Despite the lack of understanding of epigenomic characteristics in cancer, we suggest that structural changes may be worthy of further study.

Fourth, even though we made an ad hoc separation on genomic data to structural changes or functional changes, the phenotype of clinical outcome is not influenced by only one of them. As shown in Fig. 4, the integration of all genomic data sources can be helpful to unveil the relationship from genome to phenome.

4. Discussion

Since cancer is the phenotypic end-point of events cumulated through multiple levels of the biological system from genome to proteome, a single layer of biological information will not be sufficient to fully understand tumor behavior or the underlying biological mechanisms [18]. In the present study, a pilot framework of integration of multiple levels of genomic data sources, CNA, DNA

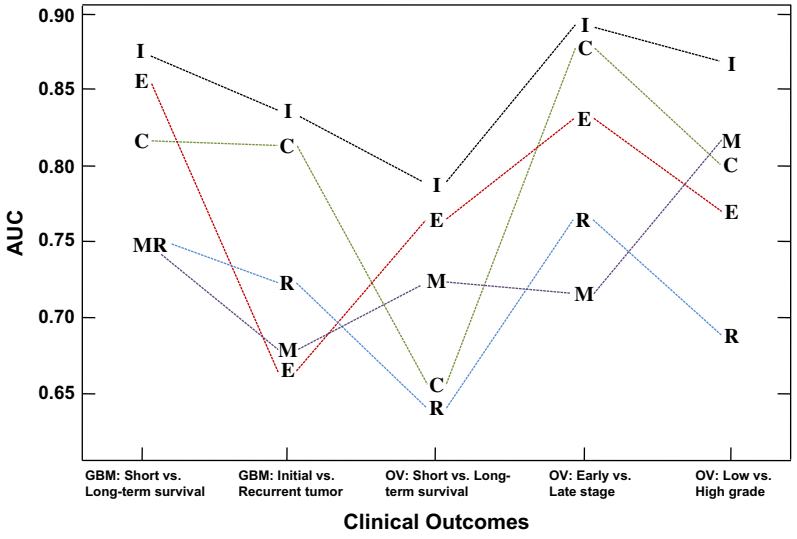

Fig. 4. Performance comparison of genomic data over the five sets of clinical outcome classification problem: C stands for CNA, M for methylation, E for gene expression, R for miRNA, and I for the integration of the four data sources, which corresponds to CMER of Fig. 3.

methylation, gene expression, and miRNA expression, has been applied to the problem of prediction of clinical outcomes in GBM and OV.

When comparing individual genomic data sources, we suggested some preliminary insights that medical experts or biologists may consider. The main result of our study is that integration of multi-layers of genomic data sources increases the opportunity of success in prediction of clinical outcomes in cancer. To the best of our knowledge, this study is the first effort in molecular-based classification of clinical outcomes from cancer patients, by unfolding and integrating the genomic, the epigenomic, and the transcriptomic features in their samples. In addition, the proposed framework would be easily extended when novel genomic data from different levels is available.

This study underpins our on-going work. There are possible relationships between the sample features (attributes) belonging to different layers of genomic data such as 'miRNA-target genes,' 'copy number alteration region-genes located in the alteration region,' and 'DNA methylation site-specific genes regulated by promoter regions.' Therefore, when integrating multiple genomic data, it will be desirable that a framework will be capable of containing the inter-relationships between sample features belonging to different layers of the biological system.

Recently, there has been an announcement from TCGA that additional cancer genomic data for about 20–25 tumor types will be generated in the next few years as the second phase of the project. With abundance in multi-layers of genomic and clinical data, our proposed integrative framework will be valuable for elucidating the underlying tumor behavior, eventually leading to more effective screening strategies and therapeutic targets in many types of cancer. The Matlab code will be available upon request.

Acknowledgments

This work was supported by Post Brain Korea 21 and the National Research Foundation of Korea (NRF) Grant funded by the Korea government (MEST) (2009-0065043/2012-0000994). We gratefully acknowledge the TCGA Consortium and all its members for the TCGA Project initiative, for providing samples, tissues, data processing and making data and results available.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2012.07.008.

References

[1] Belkin M. Regularization and semi-supervised learning on large graphs. In: Proceedings of the 17th annual conference on learning theory (COLT) 3120. Lecture notes in computer science; 2004. p. 624–38.
[2] Ben-Hur A, Noble WS. Kernel methods for predicting protein–protein interactions. Bioinformatics 2005;21(1):i38–46.
[3] Berchuck A, Iversen ES, Lancaster JM, Pittman J, Luo J, Lee P, et al. Patterns of gene expression that characterize long-term survival in advanced stage serous ovarian cancers. Clin Cancer Res 2005;11:3686–96.
[4] Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, et al. The landscape of somatic copy-number alteration across human cancers. Nature 2010;463:899–905.
[5] Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 2006;439:353–7.
[6] Boeri M, Verri C, Conte D, Roz L, Modena P, Facchinetti F, et al. MicroRNA signatures in tissues and plasma predict development and prognosis of computed tomography detected lung cancer. Proc Natl Acad Sci USA 2011;108:3713–8.
[7] Bussey KJ, Chin K, Lababidi S, Reimers M, Reinhold WC, Kuo WL, et al. Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. Mol Cancer Ther 2006;5:853–67.
[8] Chapelle O, Weston J, Scholkopf B. Cluster kernels for semi-supervised learning. Adv Neur Inform Process Syst (NIPS) 2003;15:585–92.
[9] Chin L, Gray JW. Translating insights from the cancer genome into clinical practice. Nature 2008;452:553–63.
[10] Chung FRK. Spectral graph theory. In: Number 92 in regional conference series in mathematics; 1997.
[11] Demsar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 2006;7:1–30.
[12] Fan SC, Zhang XG. CpG island methylation pattern in different human tissues and its correlation with gene expression. Biochem Biophys Res Commun 2009;383:421–5.
[13] Fan X, Shi L, Fang H, Cheng Y, Perkins R, Tong W. DNA microarrays are predictive of cancer prognosis: a re-evaluation. Clin Cancer Res 2010;16:629–36.
[14] Furnari FB, Fenton T, Bachoo RM, Mukasa A, Stommel JM, Stegh A, et al. Malignant astrocytic glioma: genetics, biology, and paths to treatment. Genes Dev 2007;21:2683–710.
[15] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.
[16] Gravendeel LA, Kouwenhoven MC, Gevaert O, de Rooi JJ, Stubbs AP, Duijm JE, et al. Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. Cancer Res 2009;69:9065–72.
[17] Gribskov M, Robinson NL. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Comput Chem 1996;20:25–33.

[18] Hanash S. Integrated global profiling of cancer. Nat Rev Cancer 2004;4:638–44.
[19] Huang E, Cheng SH, Dressman H, Pittman J, Tsou MH, Horng CF, et al. Gene expression predictors of breast cancer outcomes. Lancet 2003;361:1590–6.
[20] Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, et al. A Bayesian networks approach for predicting protein–protein interactions from genomic data. Science 2003;302:449–53.
[21] Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. Cancer statistics, 2009. CA Cancer J Clin 2009;59:225–49.
[22] Lanckriet GR, De Bie T, Cristianini N, Jordan MI, Noble WS. A statistical framework for genomic data fusion. Bioinformatics 2004;20:2626–35.
[23] Lanza G, Ferracin M, Gafa R, Veronese A, Spizzo R, Pichiorri F, et al. MRNA/microRNA gene expression profile in microsatellite unstable colorectal cancer. Mol Cancer 2007;6:54.
[24] Li A, Walling J, Ahn S, Kotliarov Y, Su Q, Quezado M, et al. Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes. Cancer Res 2009;69:2091–9.
[25] Li M, Balch C, Montgomery JS, Jeong M, Chung JH, Yan P, et al. Integrated analysis of DNA methylation and gene expression reveals specific signaling pathways associated with platinum resistance in ovarian cancer. Bmc Med Genom 2009;2.
[26] Louhimo R, Hautaniemi S. CNAmet: an R package for integrating copy number, methylation and expression data. Bioinformatics 2011;27:887–8.
[27] Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, et al. MicroRNA expression profiles classify human cancers. Nature 2005;435:834–8.
[28] Marko NF, Toms SA, Barnett GH, Weil R. Genomic expression patterns distinguish long-term from short-term glioblastoma survivors: a preliminary feasibility study. Genomics 2008;91:395–406.
[29] Monni O, Barlund M, Mousses S, Kononen J, Sauter G, Heiskanen M, et al. Comprehensive copy number and gene expression profiling of the 17q23 amplicon in human breast cancer. Proc Natl Acad Sci USA 2001;98:5711–6.
[30] Myllykangas S, Tikka J, Bohling T, Knuutila S, Hollmen J. Classification of human cancers based on DNA copy number amplification modeling. BMC Med Genom 2008;1:15.
[31] Ohn JH, Kim J, Kim JH. Genomic characterization of perturbation sensitivity. Bioinformatics 2007;23:i354–8.
[32] Peng X, Li Y, Walters KA, Rosenzweig ER, Lederer SL, Aicher LD, et al. Computational identification of hepatitis C virus associated microRNA-mRNA regulatory modules in human livers. BMC Genomics 2009;10:373.
[33] Qiu J, Noble WS. Predicting co-complexed protein pairs from heterogeneous data. PLoS Comput Biol 2008;4:e1000054.
[34] Roepman P, Wessels LF, Kettelarij N, Kemmeren P, Miles AJ, Lijnzaad P, et al. An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas. Nat Genet 2005;37:182–6.
[35] Sadikovic B, Yoshimoto M, Al-Romaih K, Maire G, Zielenska M, Squire JA. In vitro analysis of integrated global high-resolution DNA methylation profiling with genomic imbalance and gene expression in osteosarcoma. Plos One 2008;3.
[36] Sadikovic B, Yoshimoto M, Chilton-MacNeill S, Thorner P, Squire JA, Zielenska M. Identification of interactive networks of gene expression associated with osteosarcoma oncogenesis by integrated molecular profiling. Hum Mol Genet 2009;18:1962–75.
[37] Salcman M, Kaplan R. Intracranial tumors in adults. In: Salcman M, editor. Neurology of brain tumors. Baltimore: Williams & Wilkins; 1991. p. 1339–52.
[38] Saxena A, Robertson JT, Ali IU. Abnormalities of p16, p15 and CDK4 genes in recurrent malignant astrocytomas. Oncogene 1996;13:661–4.
[39] Schafer M, Schwender H, Merk S, Haferlach C, Ickstadt K, Dugas M. Integrated analysis of copy number alterations and gene expression: a bivariate assessment of equally directed abnormalities. Bioinformatics 2009;25:3228–35.
[40] Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 2003;34:166–76.
[41] Shin H, Lisewski AM, Lichtarge O. Graph sharpening plus graph integration: a synergy that improves protein functional classification. Bioinformatics 2007;23:3217–24.
[42] Shin H, Tsuda K. Prediction of protein function from networks. In: Olivier Chapelle, Bernhard Schölkopf, Alexander Zien, editors. Semi-supervised learning. MIT press; 2006. p. 339–52 [Chapter 20].
[43] Shridhar V, Lee J, Pandita A, Iturria S, Avula R, Staub J, et al. Genetic analysis of early- vs. late-stage ovarian tumors. Cancer Res 2001;61:5895–904.
[44] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol Biol Cell 1998;9:3273–97.
[45] Network TCGA. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008;455:1061–8.
[46] Tsuda K, Shin H, Scholkopf B. Fast protein classification with multiple networks. Bioinformatics 2005;21(2):i59–65.
[47] van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–6.
[48] Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 2010;17:98–110.
[49] Waldman FM, DeVries S, Chew KL, Moore 2nd DH, Kerlikowske K, Ljung BM. Chromosomal alterations in ductal carcinomas in situ and their in situ recurrences. J Natl Cancer Inst 2000;92:313–20.
[50] Wang YP, Li KB. Correlation of expression profiles between microRNAs and mRNA targets using NCI-60 data. BMC Genomics 2009;10.
[51] Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. Science 2007;318:1108–13.
[52] Wu CC, Asgharzadeh S, Triche TJ, D'Argenio DZ. Prediction of human functional genetic networks from heterogeneous data using RVM-based ensemble learning. Bioinformatics 2010;26:807–13.
[53] Zhou D, Bousquet O, Weston J, Scholkopf B. Learning with local and global consistency. Adv Neur Inform Process Syst (NIPS) 2004;16:321–8.
[54] Zhu X, Ghahramani Z, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions. In: Proceedings of the 21st international conference on machine learning (ICML), Washington, DC: AAAI Press; 2003. p. 912–19.