Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach

Muxuan Liang, Zhizhong Li, Ting Chen, and Jianyang Zeng

Abstract—Identification of cancer subtypes plays an important role in revealing useful insights into disease pathogenesis and advancing personalized therapy. The recent development of high-throughput sequencing technologies has enabled the rapid collection of multi-platform genomic data (e.g., gene expression, miRNA expression, and DNA methylation) for the same set of tumor samples. Although numerous integrative clustering approaches have been developed to analyze cancer data, few of them are particularly designed to exploit both deep intrinsic statistical properties of each input modality and complex cross-modality correlations among multi-platform input data. In this paper, we propose a new machine learning model, called multimodal deep belief network (DBN), to cluster cancer patients from multi-platform observation data. In our integrative clustering framework, relationships among inherent features of each single modality are first encoded into multiple layers of hidden variables, and then a joint latent model is employed to fuse common features derived from multiple input modalities. A practical learning algorithm, called contrastive divergence (CD), is applied to infer the parameters of our multimodal DBN model in an unsupervised manner. Tests on two available cancer datasets show that our integrative data analysis approach can effectively extract a unified representation of latent features to capture both intra- and cross-modality correlations, and identify meaningful disease subtypes from multi-platform cancer data. In addition, our approach can identify key genes and miRNAs that may play distinct roles in the pathogenesis of different cancer subtypes. Among those key miRNAs, we found that the expression level of miR-29a is highly correlated with survival time in ovarian cancer patients. These results indicate that our multimodal DBN based data analysis approach may have practical applications in cancer pathogenesis studies and provide useful guidelines for personalized cancer therapy.

Index Terms—Multi-platform cancer data analysis, restricted Boltzmann machine, multimodal deep belief network, identification of cancer subtypes, genomic data, clinical data

1 Introduction

Cancer tumors are often caused by different genetic mutations and generally display considerable phenotypic heterogeneity in cancer cells. Identification of individual cancer subtypes can reveal useful insights into disease pathogenesis and facilitate personalized cancer therapy. Clustering cancer patient data is an initial and important step to achieve this goal. By clustering cancer patients into different groups according to their genetic profiles and clinical symptoms, we can have a better view on the pathogenic mechanisms of cancer diseases, and thus find better anticancer treatment for individual disease subtypes.

The recent advent of high-throughput experimental technologies, especially next-generation DNA sequencing methods, has enabled us to rapidly collect multi-platform genomic profiles of tumor samples, such as gene expression (GE), miRNA expression (ME) and DNA methylation (DM). In particular, The Cancer Genome Atlas (TCGA) pilot project has made tremendous efforts and generated a large amount of cross-platform genomic data for exploring the complex landscapes of human cancers.

The cross-platform cancer data provide important opportunities to characterize different disease subtypes, gain insights into disease pathogenesis, and advance personalized cancer therapy. Unfortunately, analyzing such cross-platform genomic data also poses new computational challenges [1], [2] for traditional data analysis approaches such as K-means clustering methods [3] or principal component analysis [4]. On the one hand, input data from different platforms usually display distinct modalities, which typically have different representations, intrinsic correlational structures and statistical properties. For example, gene expression data measure the abundance of transcripts, i.e., messenger RNAs (mRNAs), while miRNAs expression profiles reflect the levels of miRNAs involving in globally post-transcriptional regulation of mRNAs. It is impossible to identify meaningful cancer subtypes without exploiting the intrinsic statistical features within these individual input

- *M. Liang is with the Department of Mathematical Sciences, Tsinghua University, Beijing 100084, P. R. China, and the Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706. E-mail: lmx1992413@hotmail.com.*
- *Z. Li is with the Drug Discovery Oncology Group, Genomics Institute of the Novartis Research Foundation, 75 John Jay Hopkins Drive, San Diego, CA 92121. E-mail: zhizhongli@gmail.com.*
- *T. Chen is with the Bioinformatics Division, TNLIST and Department of Computer Science and Technology, Tsinghua University, Beijing 100084, P. R. China, and the Program in Computational Biology and Bioinformatics, University of Southern California, LA, CA 90089. E-mail: tingchen@mail.tsinghua.edu.cn.*
- *J. Zeng is with the Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, P. R. China. E-mail: zengjy321@tsinghua.edu.cn.*

Manuscript received 16 May 2014; revised 28 Sept. 2014; accepted 21 Nov. 2014. Date of publication 4 Dec. 2014; date of current version 4 Aug. 2015.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2014.2377729

modalities. On the other hand, cross-platform genomic data for the same tumor sample are unlikely to be independent. For example, high correlations are often observed between gene expression and DNA methylation, as DNA methylation generally controls gene expression by affecting the interactions between DNAs and transcription factors or chromatin proteins. In addition, genetic information is typically highly correlated with clinical data, such as survival time and time to recurrence, for cancer patients. For instance, [5] has shown that breast cancer patients with a good-prognosis genetic signature can have longer survival time than those with a poor-prognosis genetic profile. These statistical correlations across multiple input modalities can also yield crucial factors in accurately differentiating distinct cancer subtypes. Thus, in practice, to perform an effective integrative clustering of cancer patients using multi-platform data, we need to take into account both intrinsic statistical properties within a single input modality and cross-platform correlations over different input modalities.

Although recently numerous integrative clustering approaches have been proposed to identify cancer subtypes from cross-platform genomic data [1], [2], [6], very few of them are particularly designed to exploit both intra-modality statistical properties and cross-modality correlations from multi-platform data. In this paper, we propose a generative model based on a multimodal deep belief network (DBN) framework [7], [8], to capture both intra-modality and cross-modality relationships and identify cancer subtypes from multi-platform genomic and clinical data. In our integrative data analysis framework, we first use a probabilistic graphical model, called restricted Boltzmann machine (RBM) [9], to encode latent features defined by each input modality. Then a joint representation of hidden variables is used to fuse cross-platform modalities and capture the common features resulting from multi-platform input data. The states of hidden variables representing the intra- and cross-modality features of multi-platform data are learned using a practical learning algorithm, called contrastive divergence (CD), in an unsupervised manner. The final states of the joint representation of latent features are then used to define the cancer subtypes based on cross-platform input data. Tests on two cancer datasets demonstrate that our integrative clustering approach can identify meaningful cancer subtypes from cross-platform genomic and clinical data, and discover key genes and miRNAs that may play different roles in the pathogenic mechanisms of these clustered cancer subtypes. These results indicate that our integrative cancer data analysis approach can provide a useful tool for studying cancer pathogenesis and advancing personalized cancer treatment.

The rest of this paper is organized as follows. In Section 2.1, we describe a restricted Boltzmann machine model. In Section 2.2, we present a multimodal deep belief network framework to address our integrative clustering problem. Section 2.3 describes a statistical inference approach to learn our proposed multimodal DBN model. Section 3 describes the test results of our integrative clustering approach on two available cancer datasets. In Section 4, we discuss the difference between our approach and other existing methods for integrative cancer data analysis.

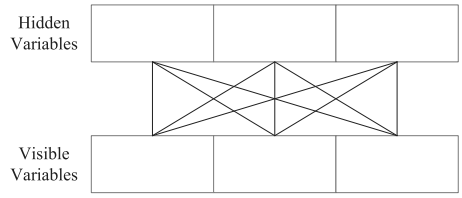

Fig. 1. A restricted Boltzmann machine in which connections only exist between visible and hidden layers, and no connection is allowed between any two variables within the same layer.

2 METHODS

2.1 Restricted Boltzmann Machines

In our cancer data analysis problem, for each patient, we call the measured genomic data from each platform (e.g., gene expression, miRNA expression and DNA methylation) the *genomic profile*. Let n be the total number of cancer patients, and let m be the total number of genomic profiles measured for each patient. A *restricted Boltzmann machine* is an undirected graphical model which consists a layer of *visible* variables $v_i, i = 1, \ldots, m$, and a layer of *hidden* variables $h_j, j = 1, \ldots, g$, where m is the number of visible variables and g is the number of hidden variables. Here, the number of visible variables is equal to the number of genomic profiles per patient. In an RBM model, each visible variable is connected to every hidden variable, but no connection is allowed between any two variables within the same layer (Fig. 1). Meanwhile, each connection between visible and hidden layers is associated with a specific weight. Let $\mathbf{W} = (W_{ij})_{m \times g}$ be an $m \times g$ matrix representing the parameter setting of weights between two layers of variables, where each element W_{ij} stands for the weight of the corresponding connection between visible variable v_i and hidden variable h_j. Let $\mathbf{a} = (a_1, \ldots, a_m)$ and $\mathbf{b} = (b_1, \ldots, b_g)$ be the bias vectors, where a_i and b_j stand for the biases of visible variables v_i and hidden variable h_j, respectively. Let $\mathbf{v} = (v_1, \ldots, v_m)$ be a vector representing the *configuration* of all visible variables, and let $\mathbf{h} = (h_1, \ldots, h_g)$ be a vector representing the *configuration* of all hidden variables. Let $(\mathbf{v}, \mathbf{h})$ be a vector representing the *joint configuration* of an RBM model. Given its joint configuration $(\mathbf{v}, \mathbf{h})$, the energy function of an RBM model can be defined as

$$E(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta}) = \mathbf{a}^T\mathbf{v} + \mathbf{b}^T\mathbf{h} + \mathbf{v}^T\mathbf{W}\mathbf{h}, \tag{1}$$

where $\boldsymbol{\theta} = (\mathbf{a}, \mathbf{b}, \mathbf{W})$ stands for the parameter setting of the model. Then the probability density function of a joint configuration $(\mathbf{v}, \mathbf{h})$ can be defined as

$$f(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(-E(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta})\right), \tag{2}$$

where $Z(\boldsymbol{\theta})$ is called the *normalizing constant*.

Based on the structure of an RBM model, we can derive the following conditional density distribution:

$$P(v_i = 1|\mathbf{h}) = \frac{1}{1 + \exp\left(-\sum_{j=1}^{g} W_{ij}h_j - b_i\right)}, \tag{3}$$

$$P(h_j = 1|\mathbf{v}) = \frac{1}{1 + \exp\left(-\sum_{i=1}^{m} W_{ij}v_i - a_j\right)}. \tag{4}$$

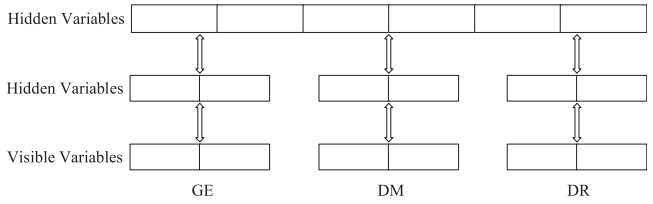

Fig. 2. An example of a multimodal DBN model which consists of three layers. The bottom two layers constitute three separate RBMs, which take gene expression, DNA methylation and drug response respectively as input data. The joint hidden variables in the top layer are connected to all hidden variables in the middle layer simultaneously.

Conventional RBM models [10] typically use binary observations as input visible data. In practice, we usually need to normalize the real-valued genomic data in advance. Hence, it is better to formulate genomic information (e.g., gene expression) as Gaussian distribution rather than binary values. Thus, ordinary RBMs hardly satisfy our needs. Instead, we use a Gaussian RBM model [9], in which visible variables given hidden values follow a Gaussian distribution, to represent observed genomic data. In a Gaussian RBM [9], the energy function in Eq. (1) can be now rewritten as

$$E(\mathbf{v},\mathbf{h};\boldsymbol{\theta}) = \sum_{i=1}^{m}\frac{(v_i-a_i)^2}{2\sigma_i^2} + \sum_{i=1}^{m}\sum_{j=1}^{g}\frac{v_i}{\sigma_i}W_{ij}h_j + \sum_{j=1}^{g}b_j h_j, \quad (5)$$

where $\boldsymbol{\theta} = (\mathbf{a},\mathbf{b},\mathbf{W},\boldsymbol{\sigma})$ stands for the parameter setting of the model, and vector $\boldsymbol{\sigma} = (\sigma_1,\ldots,\sigma_m)$ stands for the Gaussian noise in input visible data. Also, the original conditional density functions in Eqs. (3) and (4) can be rewritten as

$$P(v_i|\mathbf{h}) = N\left(b_i + \sigma_i\sum_{j=1}^{g}W_{ij}h_j, \sigma_i^2\right), \quad (6)$$

$$P(h_j = 1|\mathbf{v}) = \frac{1}{1+\exp\left(-\sum_{i=1}^{m}W_{ij}v_i - a_j\right)}, \quad (7)$$

where $N(\mu,\sigma^2)$ stands for a Gaussian distribution with mean μ and standard deviation σ.

2.2 Multimodal Deep Belief Networks

We aim to integrate multi-platform genomic and clinical data, which are generally in different forms (e.g., binary, real-valued and categorical values) and have distinct statistical properties, for cancer data analysis and identification of disease subtypes. We apply a deep learning framework [11], called *multimodal deep belief network* [7], [8], to achieve this goal. A multimodal DBN is a network of stacked RBMs, in which the separate RBMs at the bottom level take multimodal data as input, and the top-level RBMs contain hidden variables that represent the common features across different modalities from multi-platform data. To further illustrate a multimodal DBN framework, we use a specific example (see Fig. 2) of cancer data analysis, in which input data involve three modalities, including gene expression, DNA methylation and drug response (DR). In this example, gene expression and DNA methylation are both represented as real-valued vectors, while drug response is represented as a binary vector. As shown in Fig. 2, the multimodal DBN model in this example consists of three layers of variables. The bottom two layers constitute three separate stacked RBMs, which take gene expression, DNA methylation and drug response information respectively as input data. The hidden variables in the top two layers are only connected to the variables in the adjacent layers. The hidden variables in the top layer are connected to the hidden variables in the middle layer from three separate RBMs simultaneously. We call the hidden variables in the middle layer the *modality-specific hidden variables*, which encode the modality-dependent features extracted from individual single-platform input data. We call the hidden variables in the top layer the *common hidden variables*, which encode the modality-independent features across multi-platform input data. In our multimodal DBN model, the modality-specific hidden variables encode the intrinsic correlations within each input modality, while the common hidden variables fuse these intra-modality features and form a joint representation of cross-platform features. The multimodal DBN model given in Fig. 2 only contains a single layer of modality-specific hidden variables and a single layer of common hidden variables. In practice, we can also add more layers for individual types of hidden variables.

As the hidden variables are binary, we can take each configuration of all hidden variables in the top layer as a cluster. For the cancer data analysis task, we use all possible combinations of the common hidden variables in the top layer to represent distinct subtypes of cancer learned from multimodal input data. For example, suppose that there are three hidden variables in the top layer. Then we have at most $2^3 = 8$ groups of patients which represent different subtypes of cancer.

2.3 Learning

The parameters of an RBM or multimodal DBN model can be learned from data using a standard maximum likelihood estimation approach. Given a dataset $D = \{\mathbf{v}^{(i)}\}_{i=1}^{n}$, where n is the total number of patients, each data point $\mathbf{v}^{(i)}$ has a probability distribution function defined based on the energy function in Eq. (2):

$$\begin{aligned} P(\mathbf{v}^{(i)};\boldsymbol{\theta}) &= \int f(\mathbf{v}^{(i)},\mathbf{h}|\boldsymbol{\theta})\mathrm{d}\mathbf{h} \\ &= \int \frac{1}{Z(\boldsymbol{\theta})}\exp\left(-E(\mathbf{v}^{(i)},\mathbf{h};\boldsymbol{\theta})\right)\mathrm{d}\mathbf{h}, \end{aligned} \quad (8)$$

where $\boldsymbol{\theta}$ stands for the parameter setting of the model. Assuming each data point is independent, we can derive the average log-likelihood function by averaging all data points $\mathbf{v}^{(i)}$

$$\begin{aligned} l(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\log P(\mathbf{v}^{(i)};\boldsymbol{\theta}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\log\int\exp(-E(\mathbf{v}^{(i)},\mathbf{h};\boldsymbol{\theta}))\mathrm{d}\mathbf{h} - \log Z(\boldsymbol{\theta}). \end{aligned} \quad (9)$$

In a standard maximum likelihood estimation approach, we want to maximize $l(\boldsymbol{\theta})$. In general, it is impossible to derive the analytical solution to maximize the likelihood function. In practice, we can approximate the maximum likelihood solution by calculating the gradient of the log-likelihood function

$$\frac{\partial l}{\partial \boldsymbol{\theta}} = E_{f_0}\left[\frac{\partial E(\mathbf{v},\mathbf{h};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] - E_{f_\infty}\left[\frac{\partial E(\mathbf{v},\mathbf{h};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right], \tag{10}$$

where f_0 means the measure defined by original observation data, f_∞ means the measure defined by the model, and $E_M[x]$ means the expectation of x under measure M. In practice, approximate learning approaches such as mean-field inference and Markov Chain Monte Carlo (MCMC), are often used to learn the parameters of an RBM model [12]. Alternatively, Hinton (2002) proposed an practically efficient algorithm, called *contrastive divergence*, for learning the parameters of an RBM or deep learning framework. The CD algorithm approximates the result of maximizing the log likelihood function of the data by minimizing the Kullback-Leibler divergence [13], and has been proved practically useful in many cases [14]:

$$\frac{\partial l}{\partial \boldsymbol{\theta}} \approx E_{\mu_0}[\mathbf{v}\mathbf{h}^T] - E_{\mu_1}[\mathbf{v}\mathbf{h}^T], \tag{11}$$

where μ_0 is the measure defined by visible data in a Gaussian RBM, and μ_1 indicates distribution after a number of alternating Gibbs sampling steps. Basically, the Kullback-Leibler divergence is computed over the empirical distribution function of the visible data and the model [13], [14].

We use the CD algorithm in a greedy layer-wise fashion [11] to learn the parameters of our multimodal RBM model. In principle, the up-down version of the greedy layer-wise method [9], [11] can also be applied to learn our model.

2.4 Identification of Key Genes and miRNAs

In addition to clustering cancer patients into different groups, we also want to identify key biomarkers, such as crucial genes and miRNAs, that play distinct roles in the pathogenesis of different cancer subtypes. We apply a two-sample t test with pooled variance to find essential genes or miRNAs that characterize individual disease subtypes. Here we illustrate the test procedure using an example of selecting key genes. Identification of key miRNAs can be performed similarly. Suppose that data follow a Gaussian distribution and we want to check whether gene X plays significantly different roles in two distinct cancer subtypes (say Subtypes 1 and 2). We consider the null hypothesis H_0: gene X is not a key gene; and the alternative hypothesis H_1: gene X is a key gene. Suppose that the populations of cancer patients in Subtypes 1 and 2 are n_1 and n_2, respectively, and the average expression values of gene X in Subtypes 1 and 2 are z_1 and z_2, respectively. Suppose that the sample mean of expression value of gene X is z with pooled variance estimate S^2. Under null hypothesis H_0, no significant difference can be found between z_1 and z_2. We then construct the following test statistic:

$$T = \frac{z1 - z2}{\sqrt{S^2(1/n_1 + 1/n_2)}}.$$

Under null hypothesis H_0, we have $T \sim t_{n_1+n_2-2}$. We call $z_1 - z$ the *Z-score*. If the p-value is smaller than a threshold $\gamma = 10^{-10}$ (FDR $< 10^{-5}$), we select gene X as a key gene (we choose threshold FDR$< 10^{-2}$ for the selection of a key miRNA).

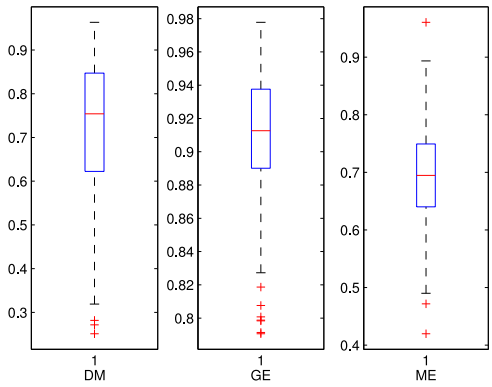

Fig. 3. Box plots of the correlations between original and reconstructed values of the GE, ME and DM modalities for ovarian cancer data.

3 Results

3.1 Data Sources

We tested our integrative data analysis method on two sets of cancer data, including an ovarian cancer dataset and a breast cancer dataset. The ovarian cancer dataset contained gene expression, DNA methylation and miRNA expression data across 385 patients which were downloaded from The Cancer Genome Atlas. The same dataset was also used in [2] to identify the common modules of multi-dimensional cancer genomic data. In addition to genomic information, we also downloaded available clinical data, such as survival time and drug response data, for ovarian cancer patients. The breast cancer dataset included GE data and corresponding clinical information, such as survival time and time to recurrence data, which were collected by the Netherlands Cancer Institute [15]. The GE data in the breast cancer dataset were also used in [15] to predict survival time of cancer patients.

3.2 Analysis of Ovarian Cancer Data

We first established a gradually shrinkage multimodal DBN model which took GE (including approximate 16,000 genes), DM (including approximate 12,000 genes) and ME (including approximate 800 miRNAs) as input data. For the GE and DM dimensions, we set two hidden layers, which included 400 and 40 modality-specific hidden variables from bottom to top respectively. For the ME dimension, we set a single hidden layer, which contained 40 modality-specific hidden variables above the visible layer. On the top of modality-specific hidden variables, we set another two hidden layers, which contained 24 and 3 common hidden variables from bottom to top, respectively.

After the learning process of the multimodal DBN model, the values of GE, DM and ME reconstructed by the CD algorithm agreed well with experimental input data, with the average correlations 0.91, 0.73 and 0.69, and standard deviations 0.037, 0.147 and 0.081 for the GE, DM and ME dimensions, respectively. The box plots of the correlations between original and constructed data in three separate modalities are shown in Fig. 3. The high correlations between reconstructed and input observation data indicate that the hidden variables in our multimodal DBN model can accurately represent the intrinsic features across multimodal input data.

TABLE 1
Populations of Different Groups over Ovarian Cancer Data Clustered by Our Multimodal DBN Model

	Group 1	Group 2	Group 3	Group 4
Population	164	9	12	21
	Group 5	Group 6	Group 7	Group 8
Population	13	23	19	124

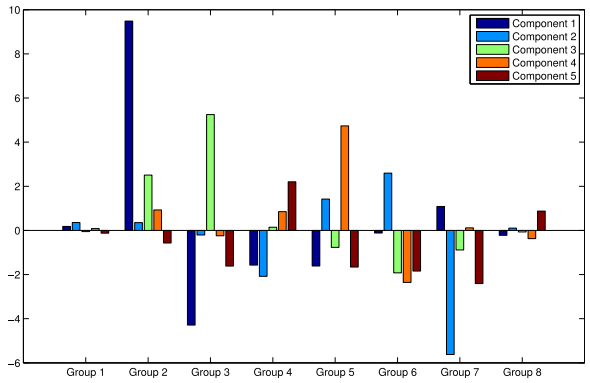

Fig. 4. The first five principal components of the mean ME values from the principle component analysis (PCA) for eight groups of ovarian cancer patients clustered by our multimodal DBN model.

The ultimate output (i.e., the states of hidden variables in the top-most layer) of the multimodal DBN model revealed that there were eight groups representing different disease subtypes based on multi-platform ovarian cancer data. These eight groups clustered by our model had biased population distribution (Table 1). The principal component analysis (PCA) of the mean ME values indicates that these eight groups displayed significantly different genomic signatures in the ME dimension. In particular, for the first principal component, Group 2 exhibited the largest value while Group 3 had the smallest value (Fig. 4).

We also downloaded corresponding clinical data, including survival time and drug record information, from TCGA to investigate clinical discrepancy among the above identified subtypes of ovarian cancer. In total, the survival time data of 372 patients were available. Table 2 summarizes survival time for eight subtypes of ovarian cancer identified by our multimodal DBN model. On average, cancer patients in Groups 2, 4, 7 and 8 lived 20 to 50 percent longer than those in other four groups. In Groups 1, 2, 4 and 7, more than 70 percent of patients had survival time longer than 900 days. In Groups 2, 4, 5, 6, 7 and 8, more than 50 percent of cancer patients lived more than 1,500 days, while the percentage was reduced to 31.1 percent in Group 1. In addition, the Kaplan-Meier plots, which have been commonly used in medical studies to estimate the survival function from survival time data, showed that distinct survival functions existed among different cancer subtypes (Figs. 5 and 6). For example, the Kaplan-Meier estimator showed that patients in Group 8 tended to have higher probability to survive than those in Group 1 (Fig. 5). This survival time analysis implies the soundness of our clustering scheme. If we merge Groups 1, 3 into one super group (say Super-group A) and the remaining groups into another super group (say

TABLE 2
Survival Time of Different Subtypes of Ovarian Cancer Identified by Our Multimodal DBN Model

	Group 1	Group 2
Mean survival time (days)	1,255.02 (62.39)	1,839.33 (156.91)
900-day survival rate	72.2% (3.7%)	88.9% (10.5%)
1500-day survival rate	31.1% (4.0%)	59.3% (25.2%)
	Group 3	Group 4
Mean survival time (days)	1,028.29 (150.28)	2,728.09 (551.93)
900-day survival rate	68.8% (15.7%)	80.4% (11.1%)
1500-day survival rate	41.3% (17.8%)	60.3% (19.3%)
	Group 5	Group 6
Mean survival time (days)	1,418.75 (277.59)	1,575.91 (207.36)
900-day survival rate	56.8% (16.5%)	69.7% (11.5%)
1500-day survival rate	56.8% (16.5%)	69.7% (11.5%)
	Group 7	Group 8
Mean survival time (days)	2,374.41 (277.90)	1,885.96 (151.74)
900-day survival rate	86.3% (9.2%)	65.6% (5.5%)
1500-day survival rate	86.3% (9.2%)	54.2% (6.2%)

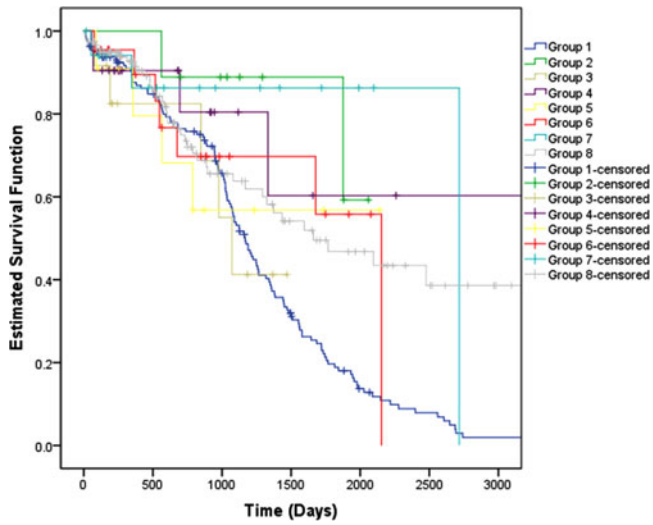

Fig. 5. The Kaplan-Meier plots for all ovarian cancer patients in eight different groups clustered by our multimodal DBN model.

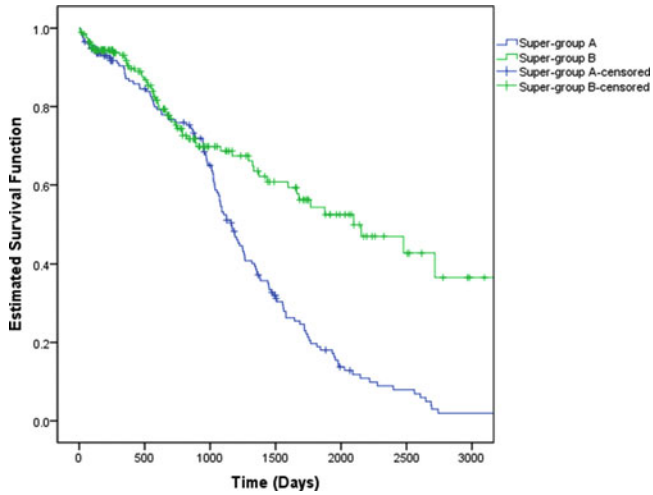

Fig. 6. The Kaplan-Meier plots for ovarian cancer patients in Super-groups A and B, which merged Groups 1, 3 and Groups 2, 4, 5, 6, 7, 8, respectively from the original clustering scheme produced by our multimodal DBN model.

TABLE 3
Survival Time for Ovarian Cancer Patients in Super-Groups A and B

	Super-group A	Super-group B
Mean survival time (days)	1,251.68 (61.28)	2,109.63 (170.88)
1500-day survival rate	31.2% (3.9%)	60.8% (4.7%)
3000-day survival rate	2.0% (1.4%)	36.5% (8.1%)

Super-groups A and B merged Groups 1, 3 and Groups 2, 4, 5, 6, 7, 8, respectively from the original clustering scheme produced by our multimodal DBN model.

Super-group B), the difference in survival time can be more noticeable. Both 1,500-day and 3,000-day survival rates of Super-group B were nearly as twice as those of Super-group A (Fig. 6 and Table 3). These results indicate that our multimodal DBN model can distinguish clinical difference among subtypes of ovarian cancer using multi-platform genomic data.

We also checked drug use information for different subtypes of ovarian cancer identified by our model. Overall, drug record data for 72 patients were collected with 36 in Group 1 and the remaining 36 in Group 8. We examined the difference in drug use rates for patients between Groups 1 and 8. In particular, our analysis mainly focused on seven drugs that were commonly used in cancer treatment, including carboplatin, cisplatin, doxil, gemcitabine, taxol, taxotere and topotecan (Fig. 7). We noticed a significantly wider use of all seven usual drugs in Group 8 than in Group 1, which may be due to the possible poor diagnosis for patients in Group 8. The phenotypes of Group 8 might be more complex, which induced patients to use a wider range of drugs in cancer treatment. The potentially worse disease severity in Group 8 corresponded to the fact that survival time of Group 8 was generally shorter than that of Group 1. The above analysis indicates the possibility of personalized therapy for different subtypes of cancer based on multi-platform genomic information.

Using the two-sample t test procedure as described in Section 2.4, only 41 genes in the GE dimension and 55 miRNAs in the ME dimension were identified with significant difference between Groups 1 and 8 (FDR is smaller than 10^{-5} for 41 genes, and 10^{-2} for 55 miRNAs). One of the distinct key genes between Group 1 and Group 8 is human Telomerase Reverse Transcriptase (hTERT) which encodes the catalytic component of the enzyme telomerase. While normal tissues have low telomerase activity, more than 90 percent of human cancers re-activate this enzyme, suggesting that hTERT is crucial for cancer immortality and progression. The gene hTERT can be activated by various mechanisms, including stimulation by oncogenes such as c-Myc or direct mutations in its promoter region [16], [17]. The different expression levels of hTERT between Groups 1 and 8 indicate that the telomerase activity may play distinct roles in the pathogenesis of different subtypes of ovarian cancer. In addition, recently multiple inhibitors against hTERT have been in clinical development [18]. As our integrative clustering approach is able to distinguish the discrepancy of telomerase activity for different subtypes of ovarian cancer, it may provide the power to predict the sensitivity

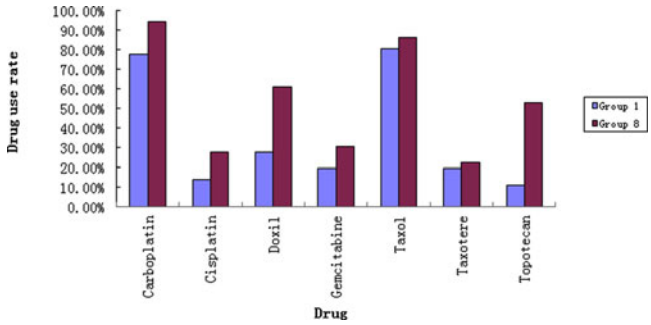

Fig. 7. Drug use records for all available ovarian cancer patients in Groups 1 and 8 clustered by our multimodal DBN model.

of anti-cancer drugs and have potential applications in personalized therapy in cancer treatment. Among all miRNAs markers that can be used to distinguish Groups 1 and 8, miR-29a is particularly interesting because it is a known tumor suppressor in multiple systems [19], [20]. Consistent with its role at different stages of tumor growth, its expression level is highly correlated with survival time in mantle cell lymphoma (MCL) patients [21]. Combined with the survival analysis as summarized in Table 2, we found that patients in Group 1 with low expression in miR-29a (Z-score $= -3.14$) lived at least 50 percent longer than those in Group 8 with high expression in miR-29a (Z-score $= 3.42$). This observation agreed with that in the analysis of MCL patients [21]. Our result therefore established a potential factor to determine survival time of ovarian cancer patients [21].

3.3 Analysis of Breast Cancer Data

In our analysis of breast cancer data, we applied our multimodal DBN framework to cluster cancer patients by integrating genetic information with clinical data, such as survival time and time to recurrence. We established a gradually shrinkage multimodal DBN which took both gene expression and clinical data (e.g., survival time and time to recurrence) as input data. In particular, for genetic information, our multimodal DBN framework contained four hidden layers, which included 800, 80, 8 and 2 modality-specific hidden variables from downwards to upwards, respectively. For clinical data, our multimodal DBN framework contained one hidden layer with a single hidden variable. On the top of these two modalities which encoded genetic and clinical data separately, we set another hidden layer, which included two modality-independent hidden variables.

This multimodal DBN model clustered all breast cancer patients into four groups, which had a clear discrepancy in survival time (Table 4 and Fig. 8). Note that in [15], the breast cancer patients with the same data were clustered into two groups based on a 70-gene prognosis profile, namely poor-prognosis and good-prognosis groups. To examine the difference between individual groups identified by our multimodal DBN framework, we also checked the correlations of the gene expression of the 70-gene prognosis signature established in [15] and the percentage of ESR1 mutation (Table 4). The ESR1 mutation data was obtained from the original dataset downloaded from the Netherlands Cancer Institute.

TABLE 4
Different Subtypes of Breast Cancer Identified by Our Multimodal DBN Model Using Both Genetic and Clinical Data

	Group A	Group B
Mean survival time (years)	17.21 (0.34)	5.52 (0.42)
10-year survival rate	95.2% (2.1%)	30.5% (6.6%)
5-year survival rate	100.0% (0%)	48.7% (6.1%)
Mean correlation	0.34	0.00
Percentage of ESR1 mutation	80.82%	47.14%
Population	146	70
	Group C	Group D
Mean survival time (years)	7.37 (0.33)	3.57 (0.45)
10-year survival rate	0.0% (0.0%)	0.0% (0.0%)
5-year survival rate	86.8% (4.1%)	0.0% (0.0%)
Mean correlation	0.37	0.02
Percentage of ESR1 mutation	97.14%	77.78%
Population	70	9

"Mean correlation" means the mean of correlations with good-prognosis signature, which is the average correlation coefficient of the 70-gene prognosis profile between patients in each group identified by our algorithm and those with the good-prognosis signature established in [15]. "Percentage of ESR1 mutation" is the fraction of patients with ESR1 mutation in individual groups.

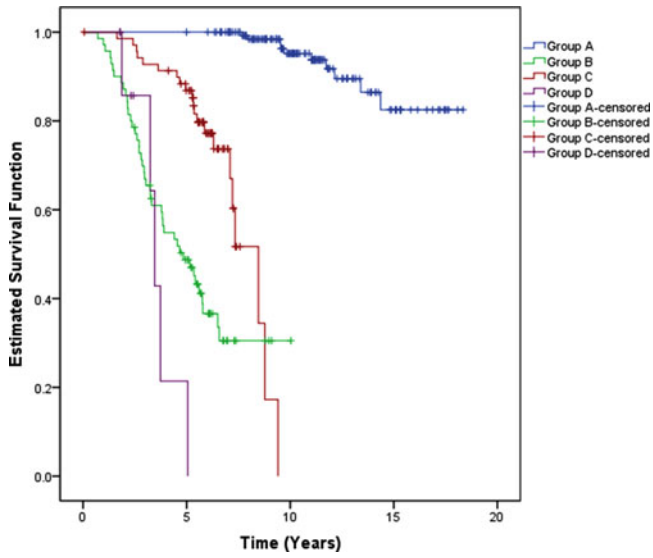

Fig. 8. The Kaplan-Meier plots of different subtypes of breast cancer identified by our multimodal DBN model using both genetic and clinical data.

We first calculated the correlation coefficients of the 70-gene prognosis profile between individual patients in each group identified by our approach and those associated with a good-prognosis signature established in [15]. The average correlation coefficients over all patients in individual subtypes of cancer were then reported to examine the clustering scheme. As shown in Table 4, we can easily draw a line to separate subtypes of breast cancer using the 70-gene prognosis signature. In particular, Groups A and C had a better agreement with the good-prognosis group defined in [15] based on the 70-gene profile than the other two groups. This observation can also be supported from clinical evidence, that is, patients in Groups A and C had closer mean survival time and 5-year survival rates than those in Groups B and D. Among these four groups identified by our model discrepancy and connections can also be found both in genetic information and clinical information. In particular, Groups A and C shared similar genetic characteristic in the mean correlation with the 70-gene prognosis signature, and so did Groups B and D (Table 4). Groups A and C tended to have similar 5-year survival rates, but with quite different 10-year survival rates.

Next, we examined the percentage of ESR1 mutation in different subtypes of breast cancer identified by our model. As shown in Table 4, Groups A and C had a higher ESR1 mutation frequency than other two groups (i.e., Groups B and D). This gap was also consistent with evidence derived from genetic (i.e., the 70-gene prognosis profile) and clinical (i.e., survival time) data. The estrogen receptor (ER), including both ESR1 and ESR2, is a family of transcription factors that are activated by hormone estrogen. It has been observed that ER is aberrantly over-expressed in more than 70 percent of breast cancer patients, i.e., ER-positive breast cancer patients [22]. ER protein immunohistochemistry (IHC) has been used as a clinical diagnostic marker for breast cancer [23] and our results suggest that ESR1 can also be used as a key factor to distinguish different subtypes of breast cancer. Compared to the results in [15], we combined both genomic and clinical data in a different way, that is, clinical information was used in our model to integrate with genomic data to cluster cancer patients rather than being used to validate the classification results in [15].

All these results show that, in our multimodal DBN modal, hidden variables can effectively capture intrinsic features within each input modality and cross-modality correlations across multi-platform data. The states of the hidden variables in the final top layer can be used to determine different subtypes of cancer, which are mainly defined based on the intra- and cross-modality statistical properties defined by multimodal input data.

4 Discussion

With a large number of parameters, the deep learning model could overfit the data. In practice, we apply several empirical rules to avoid overfitting the data. Here, we use the choice of the number of hidden variables as an example to demonstrate how we choose the parameters in our deep learning model in principle. We mainly use two principles to determine the number of hidden variables, including that in the top layer. First, according to the empirical studies in the deep learning literature, the number of hidden variables is usually around one tenth of that of visible variables [7], [8], [9], [10], [11], [24], [25], [26]. Second, we choose the number of hidden variables by measuring how it influences the mean squared error (MSE) between the original visible data and corresponding reconstructed values. By configuring the deep learning model with different numbers of hidden variables, we can observe how the MSE changes with respect to the number of hidden variables. For example, in Fig. 9, we can easily obtain a reasonable choice near the change point for the number of hidden variables, which achieves a small MSE value but with less overfitting (i.e., the MSE is not too small). Before the change point, the MSE drops dramatically as the number of hidden variables increases, which indicates that the model is at lack of fitness. After the change point, the MSE only achieves small improvement as the number of hidden variables increases. This indicates that

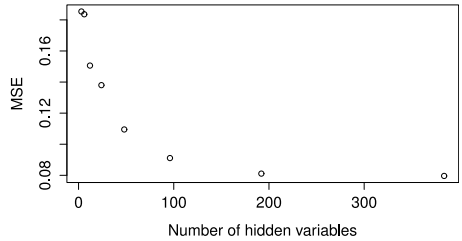

Fig. 9. A plot example of the number of hidden variables vs. the mean squared error between the original visible data and corresponding reconstructed values.

probably the model is overfitting the data. Thus, we are inclined to choose such a change point or pick the number of hidden variables near the change point (considering other principles).

We have compared the clustering results of our method with those of K-means on the same dataset for ovarian cancer patients. The clustering results of K-means are shown in Table 5 and Fig. 10. The comparison results show that our multimodal DBN model had better clustering outcome than the K-means approach. As shown in Table 5 and Fig. 10, K-means had worse performance in separating the patients into different groups than our approach. For example, the difference between the groups with the longest and shortest survival time was nearly 1,700 days in our clustering results (Table 2 and Fig. 5), while the corresponding difference in the K-means results was only about 1,000 days (Table 5 and Fig. 10). In Fig. 10, curves of different groups were closer to each other than those identified by our method. Among all eight groups identified by K-means, four of them had a trend of long-term survival with the probability below 40 percent and were quite similar to each other. On the other hand, in the subtypes identified by our method, only one of them showed a trend of long-term survival with the probability below 40 percent and the highest probability of long-term survival was above 60 percent. These results indicate that our method is more capable of capturing the intrinsic relationship in different modalities and has better clustering performance.

TABLE 5
Survival Time of Different Subtypes of Ovarian Cancer Identified by K-Means

	Group 1	Group 2
Mean survival time (days)	1,384.82 (199.94)	1,172.00 (0.00)
900-day survival rate	60.7% (9.2%)	100.0% (0.0%)
1500-day survival rate	46.4% (9.4%)	0.0% (0.0%)
	Group 3	Group 4
Mean survival time (days)	1,078.10 (79.75)	2,097.45 (251.11)
900-day survival rate	72.1% (6.9%)	69.1% (10.0%)
1500-day survival rate	19.9% (6.3%)	69.1% (10.0%)
	Group 5	Group 6
Mean survival time (days)	1,445.11 (120.34)	1,657.53 (193.02)
900-day survival rate	81.5% (6.3%)	77.1% (5.3%)
1500-day survival rate	36.5% (8.0%)	43.2% (7.2%)
	Group 7	Group 8
Mean survival time (days)	1,791.66 (237.15)	1,807.58 (164.10)
900-day survival rate	57.3% (8.4%)	68.4% (6.3%)
1500-day survival rate	48.5% (9.2%)	56.9% (7.4%)

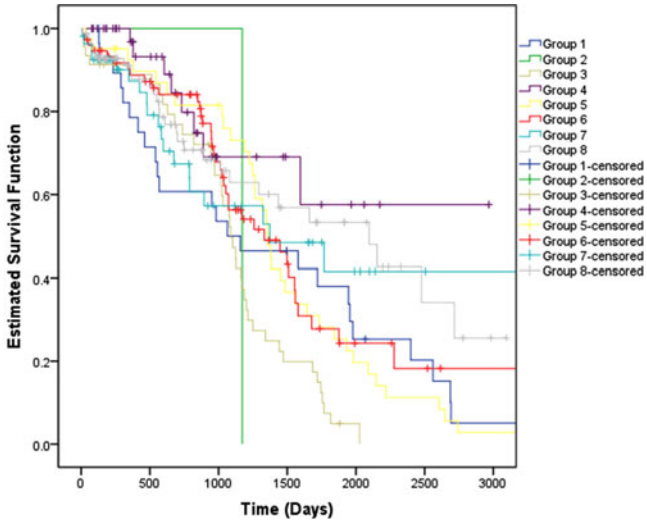

Fig. 10. The Kaplan-Meier plots for all ovarian cancer patients in eight different groups clustered by K-means.

Test results on two cancer datasets, as shown in Sections 3.2 and 3.3, demonstrate that our multimodal DBN framework provides a promising tool for integrative cancer data analysis. Our multimodal DBN model differs from traditional clustering approaches, such as K-means based methods [4] and Bayesian methods [1], in several aspects.

First, unlike the K-means based methods, which are normally sensitive to the choice of initial cluster centers, our multimodal DBN model is a probabilistic framework which remains stable under perturbation of initial states. Second, Bayesian approaches, generally require a prior distribution (e.g., normal distribution) about latent variables. In reality, the assumption of normality on latent variables for clustering is often ungrounded. In our multimodal DBN model, the distribution of hidden variables is generated automatically from its conditional distribution on visible variables. In practice, the states of hidden variables can be often interpreted based on our current knowledge about known biological processes or specific elements in the underlying cellular pathways. In fact, a similar functional representation of hidden variables has been discovered in the voice recognition application [26]. Third, most of the conventional clustering approaches [1], [2], [4], [26] are not particularly designed for dealing with cross-platform data, while our multimodal DBN is specially developed to capture both intrinsic statistical properties within each input modality and cross-modality correlations from multimodal input data. These advantages make our approach more suitable for analyzing multi-platform cancer data and identifying different subtypes of cancer, which will provide useful guidelines for personalized cancer therapy.

The multimodal DBN framework can be easily implemented and extended to address large-scale problems. In conventional K-means based approaches [4], it can be difficult to calculate eigenvectors and eigenvalues for a large amount of data. In addition, the underlying CD algorithm in multimodal DBN model has been proved more efficient than the Markov Chain Monte Carlo approaches [13], which have been widely used to learn those Bayesian methods for clustering. These facts indicate that multimodal DBN models can offer a more practical tool to handle big data challenges.

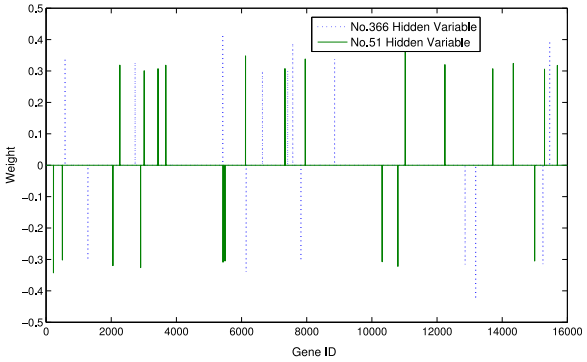

Fig. 11. The distribution of the weights that are more than 0.3 in absolute value and associated with the connections to No. 51 and No. 366 hidden variables in the bottom-most hidden layer.

In our analysis, the hidden variables in the top layer of our multimodal DBN model represent different cancer subtypes. Also, the hidden variables in other layers reveal certain biological information, such as the existence of aberrant proteins. When generating those hidden variables, our model constructed the linear combination of visible variables using weights as coefficients. The larger the absolute value of the weight, the higher impact its corresponding visible variable has on the hidden variable. Fig. 11 shows that in the GE modality of ovarian cancer data, the distributions of the weights that are more than 0.3 in absolute value are totally different between the connections to No. 51 and No. 366. A further genetic functional analysis [27], [28] has shown that two functional annotation clusters have been identified among those 22 genes which have large impacts on No. 366 hidden variable. One is related to olfactory reception and transduction (Enrichment Score: 2.77), and the other is related to transmembrane proteins (Enrichment Score: 0.89). These clusters demonstrate the functional annotations of certain hidden variables. The difference in the weight distributions reveals the distinction in functional annotation of hidden variables.

In this paper, we have mainly applied a multimodal DBN model to perform integrative clustering on multi-platform cancer data. In principle, our model can also be used to predict missing values based on other variables of a cancer patient after clustering the whole dataset. For example, our model can be used to predict drug use for each patient based on available genetic information. As our modal is a probabilistic framework, each new prediction can be associated with a confidence score. A similar strategy has been used in [29], [30] to predict missing drug-target interactions by exploiting the intrinsic correlations of previously known interactions.

Acknowledgments

Jianyang Zeng is the corresponding author. The authors thank the anonymous reviewers for their helpful comments and suggestions. The authors are grateful to Mr. Sai Zhang, Mr. Xin Zhang and Dr. Chao Cheng for their helpful discussions on our data analysis method. This work was supported in part by the National Basic Research Program of China Grant 2011CBA00300, 2011CBA00301, the National Natural Science Foundation of China Grant 61033001, 61361136003 and 61472205, and China's Youth 1000-Talent Program.

References

[1] R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, vol. 25, no. 22, pp. 2906–2912, 2009.

[2] S. Zhang, C.-C. Liu, W. Li, H. Shen, P. W. Laird, and X. J. Zhou, "Discovery of multi-dimensional modules by integrative analysis of cancer genomic data," *Nucleic Acids Res.*, vol. 40, no. 19, pp. 9379–9391, 2012.

[3] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *J. Royal Statist. Soc. Ser. C (Appl. Statist.)*, vol. 28, no. 1, pp. 100–108, 1979.

[4] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, p. 29.

[5] L. J. van't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[6] D. R. Rhodes, T. R. Barrette, M. A. Rubin, D. Ghosh, and A. M. Chinnaiyan, "Meta-analysis of microarrays: Interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer," *Cancer Res.*, vol. 62, no. 15, pp. 4427–4433, 2002.

[7] M. Kim, J. Nam, H. Lee, J. Ngiam, A. Khosla, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[8] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," *J. Mach. Learn. Res.*, vol. 15, pp. 2949–2980, 2014.

[9] G. E. Hinton, "A practical guide to training restricted boltzmann machines," Dept. of Computer Science, Univ. of Toronto, Tech. Rep. UTML TR 2010-003, 2010.

[10] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted boltzmann machines for collaborative filtering," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 791–798.

[11] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[12] M. Welling and G. Hinton, "A new learning algorithm for mean field boltzmann machines," in *Proc. Int. Conf. Artificial Neural Netw.*, 2002, pp. 351–357.

[13] M. A. Carreira-Perpinan and G. E. Hinton, "On contrastive divergence learning," in *Proc. 10th Int. Workshop Artif. Intell. Statist.*, 2005, pp. 59–66.

[14] T. Tieleman, "Training restricted boltzmann machines using approximations to the likelihood gradient," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1064–1071.

[15] M. J. Van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards, "A gene-expression signature as a predictor of survival in breast cancer," *New England J. Med.*, vol. 347, no. 25, pp. 1999–2009, 2002.

[16] K.-J. Wu, C. Grandori, M. Amacker, N. Simon-Vermot, A. Polack, J. Lingner, and R. Dalla-Favera, "Direct activation of tert transcription by c-myc," *Nat. Genetics*, vol. 21, no. 2, pp. 220–224, 1999.

[17] F. W. Huang, E. Hodis, M. J. Xu, G. V. Kryukov, L. Chin, and L. A. Garraway, "Highly recurrent tert promoter mutations in human melanoma," *Science*, vol. 339, no. 6122, pp. 957–959, 2013.

[18] A. Glukhov, L. Svinareva, S. Severin, and V. Shvets, "Telomerase inhibitors as novel antitumor drugs," *Appl. Biochem. Microbiol.*, vol. 47, no. 7, pp. 655–660, 2011.

[19] S.-Y. Park, J. H. Lee, M. Ha, J.-W. Nam, and V. N. Kim, "mir-29 mirnas activate p53 by targeting p85α and cdc42," *Nat. Struct. Molecular Biol.*, vol. 16, no. 1, pp. 23–29, 2008.

[20] H. J. Bae, J. H. Noh, J. K. Kim, J. W. Eun, K. H. Jung, M. G. Kim, Y. G. Chang, Q. Shen, S.-J. Kim, W. S. Park, J. Y. Lee, and S. W. Nam, "Microrna-29c functions as a tumor suppressor by direct targeting oncogenic sirt1 in hepatocellular carcinoma," *Oncogene*, vol. 33, no. 20, pp. 2557–2567, 2013.

[21] J.-J. Zhao, J. Lin, T. Lwin, H. Yang, J. Guo, W. Kong, S. Dessureault, L. C. Moscinski, D. Rezania, W. S. Dalton, E. Sotomayor, J. Tao, and J. Q. Cheng, "Microrna expression profile and identification of mir-29 as a prognostic marker and pathogenetic factor by targeting cdk6 in mantle cell lymphoma," *Blood*, vol. 115, no. 13, pp. 2630–2639, 2010.
[22] W. L. McGuire and C. K. Osborne, "The use of steroid hormone receptors in the treatment of human breast cancer: A review," *Bull Cancer*, vol. 66, no. 3, pp. 203–209, 1979.
[23] J. Underwood, "Oestrogen receptors in human breast cancer: Review of histopathological correlations and critique of histochemical methods," *Diagnostic Histopathol.*, vol. 6, no. 1, pp. 1–22, 1983.
[24] R. Salakhutdinov and G. E. Hinton, "Deep boltzmann machines," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2009, pp. 448–455.
[25] A. Mohamed, T. Sainath, G. Dahl, B. Ramabhadran, G. Hinton, and M. Picheny, "Deep belief networks using discriminative features for phone recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2011, pp. 5060–5063.
[26] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and T. N. Sainath, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
[27] B. T. Sherman, D. W. Huang, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using david bioinformatics resources," *Nat. Protocols*, vol. 4, no. 1, pp. 44–57, 2008.
[28] W. Huang da, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Res.*, vol. 37, no. 1, pp. 1–13, 2009.
[29] N. Le Roux and Y. Bengio, "Representational power of restricted boltzmann machines and deep belief networks," *Neural Comput.*, vol. 20, no. 6, pp. 1631–1649, 2008.
[30] Y. Wang and J. Zeng, "Predicting drug-target interactions using restricted boltzmann machines," *Bioinformatics*, vol. 29, no. 13, pp. i126–i134, 2013.

Muxuan Liang received the BS degree in mathematics and applied mathematics from Tsinghua University in 2014. He is currently a graduate student in the Department of Statistics and a research assistant in the Department of Biostatistics and Medical Informatics at the University of Wisconsin-Madison. His research interests include high-dimensional data analysis, statistical learning and machine learning, bioinformatics and statistical genomics.

Zhizhong Li received the BS degree in biology from Tsinghua University and the PhD degree in pharmacology and cancer biology from Duke University. He then finished the Presidential Postdoc fellowship at the Novartis Institutes for Biomedical Research. He is currently a principle investigator at the Genomics Institute of the Novartis Foundation (GNF), focusing on cancer genetics and genomics studies, in order to develop personalized cancer drugs. He has published multiple high impact research papers on cancer and stem cell biology with a total citation more than 2,500.

Ting Chen is a professor in computer science at Tsinghua University. His research focus is on the algorithmic design and statistical learning in bioinformatics. He received the Alfred Sloan Fellowship in 2004.

Jianyang Zeng received the PhD degree in computer science from Duke University in 2011. He is a tenure-track assistant professor in the Institute for Interdisciplinary Information Sciences (IIIS) at Tsinghua University. He was a postdoctoral associate in the Department of Computer Science at Duke University and the Duke University School of Medicine in 2011-2012. His research interests include computational biology, machine learning, and big data analysis. In 2012, he was chosen as a finalist for a lane fellowship in computational biology at Carnegie Mellon University (CMU). He received the Best Paper Award in the Sixth International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT) in 2005. He received the China's Youth 1000-Talent Program.