Jiang *et al. BMC Bioinformatics* (2017) 18:447
DOI 10.1186/s12859-017-1859-6

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Identify Huntington's disease associated genes based on restricted Boltzmann machine with RNA-seq data

Xue Jiang[1,2], Han Zhang[1,2], Feng Duan[1,2] and Xiongwen Quan[1,2*]

Abstract

Background: Predicting disease-associated genes is helpful for understanding the molecular mechanisms during the disease progression. Since the pathological mechanisms of neurodegenerative diseases are very complex, traditional statistic-based methods are not suitable for identifying key genes related to the disease development. Recent studies have shown that the computational models with deep structure can learn automatically the features of biological data, which is useful for exploring the characteristics of gene expression during the disease progression.

Results: In this paper, we propose a deep learning approach based on the restricted Boltzmann machine to analyze the RNA-seq data of Huntington's disease, namely stacked restricted Boltzmann machine (SRBM). According to the SRBM, we also design a novel framework to screen the key genes during the Huntington's disease development. In this work, we assume that the effects of regulatory factors can be captured by the hierarchical structure and narrow hidden layers of the SRBM. First, we select disease-associated factors with different time period datasets according to the differentially activated neurons in hidden layers. Then, we select disease-associated genes according to the changes of the gene energy in SRBM at different time periods.

Conclusions: The experimental results demonstrate that SRBM can detect the important information for differential analysis of time series gene expression datasets. The identification accuracy of the disease-associated genes is improved to some extent using the novel framework. Moreover, the prediction precision of disease-associated genes for top ranking genes using SRBM is effectively improved compared with that of the state of the art methods.

Keywords: Restricted Boltzmann machine, Key genes associated to the disease progression, Huntington's disease, RNA-seq data

Background

Neurodegenerative disease is a type of chronic degenerative disease in the central nervous system with the degenerative changes of the neuronal cells in brain and spinal cord. The symptoms of the neurodegenerative disease deteriorate slowly and eventually lead to death [1, 2]. Thereinto, the Huntington's disease is due to a triplet (CAG) repeat elongation in the Huntington gene (IT15), which further affects numerous interactions between molecules. With the accumulation of the variant Htt protein in brain, a number of molecular pathways are affected in turn, resulting in neuronal malfunction and degeneration. Changes in Htt protein and the interactions between molecules are closely associated with the abnormalities of gene expression. It has been shown that there exist abnormalities of gene expression among the genes related to nerve conduction in the striatum tissue of Huntington's disease individuals [3, 4]. Since the complexity of chronic disease, the molecular pathogenesis of Huntington's disease is not entirely clear. Nevertheless, identifying the key genes associated with the disease deterioration can reveal useful insights into the disease pathogenesis.

The rapid development of high-throughput sequencing technologies, especially next-generation sequencing

*Correspondence: quanxw@nankai.edu.cn
[1]College of Computer and Control Engineering, Nankai University, Tongyan Road, 300350 Tianjin, China
[2]Tianjin Key Laboratory of Intelligent Robotics, Nankai University, Tongyan Road, 300350 Tianjin, China

methods, provides possibility for us to explore the molecular mechanisms of complex disease on a genome-wide scale. However, because of the complex etiology of chronic diseases [5], the traditional disease-associated gene prediction methods cannot effectively identify the genes affected during the disease development. Generally, the disease-associated prediction methods roughly fall into three categories: network-based methods [6, 7], statistic-based methods [8–10], and machine learning methods [11, 12]. At present, as a branch of machine learning methods, the deep learning methods have become the most advanced technology in the field of computer vision, speech recognition and natural language processing. Deep learning methods use the hierarchical structure of deep neural network to conduct the nonlinear transfer of the input data, which could learn automatically the internal features that represent the original data [13, 14]. Compared with methods that are of manual designed features, the deep learning methods could improve the prediction accuracy. Recently, the deep learning methods have been introduced into the field of bioinformatics. Liang et al. [15] designed a multimodal deep belief network to conduct the integrative data analysis on multi-platform genomic data including gene expression data, miRNA expression data, and DNA methylation data. They used the model to detect a unified representation of latent features, capture both intra- and cross-modality correlations, and to identify key genes that may play distinct roles in the pathogenesis between different cancer subtypes. Cheng et al. [16] designed a miRNA prediction algorithm based on convolutional neural network (CNN). The CNN automatically extracts essential information from the input data while the exact miRNA target mechanisms are not well known. Experimental results demonstrated that the algorithm significantly improved the prediction accuracy.

During neurodegenerative disease development, gene expression level is affected by many factors, e.g. the environment, impaired metabolic pathways, protein misfolding, etc [17–19]. Intuitively, identifying the key genes associated with the disease development is to screen the genes that are most seriously affected by these factors over with time. Consequently, the features that distinguish disease-related genes from non-disease-related genes could be represented by these factors. Extracting the deep hierarchical structure of the gene expression data and learning the important information represented by the decreased neurones in hidden layers are helpful to further understand the changes of gene expression during the disease development. In this paper, we designed a deep learning approach based on restricted Boltzmann machine to analyze the gene expression data [20], namely stacked restricted Boltzmann machine (SRBM). We used the unsupervised contrastive divergence algorithm (CD) to learn the parameters in each restricted Boltzmann machine [21, 22]. By maximizing the likelihood function, the probability distribution of the hidden layer variables fitted the probability distribution of the original data well. We trained the stacked restricted Boltzmann machine in a greedy layer-wise fashion [23]. Because the number of neurons in hidden layers is far smaller than that in the visible layer, we could reduce dimensions of the input data and capture useful high-level features of the input data at the same time. The gene expression level is manipulated by regulatory factors. In this work, we assume that the effects of regulatory factors can be captured by the hierarchical structure and narrow hidden layers of the SRBM. We used the model to rank the genes, aiming to make key genes that may play important roles in the pathogenesis of Huntington's disease with high rankings. First, according to the differentially activated hidden neurons obtained by gene expression datasets at different time periods, we selected disease-associated factors. Then, we selected disease-associated genes according to the changes of the gene energy in SRBM at different time periods. Experimental results demonstrated that SRBM can detect the important information for differential analysis of time series gene expression datasets. The identification accuracy of the disease-associated genes is improved to some extent. Moreover, the prediction precision of disease-associated genes for top ranking genes using SRBM is effectively improved compared with that of the state of the art methods.

The presented study is organized as follows: The deep learning approach proposed in this paper is presented in "Methods" section. Experiments that analyze the performance of the stacked restricted Boltzmann machine and the overall discussion of the experimental results are reported in "Results and discussion" section. Conclusions are presented in "Conclusions" section.

Methods

In this section, first, the stacked restricted Boltzmann machine model and the learning method are described. Next, we detailedly describe how the SRBM is used to extract the disease-associated genes with gene expression data at different disease stages. Finally, we present the parameter setting of the SRBM.

Stacked restricted Boltzmann machine

Model

RBM is a kind of undirected probabilistic graphical model containing a layer of observable variables and a single layer of hidden variables [24]. In the RBM model (Fig. 1), each visible variable connects to every hidden variable, but no connections are allowed between any two variables within the same layer.

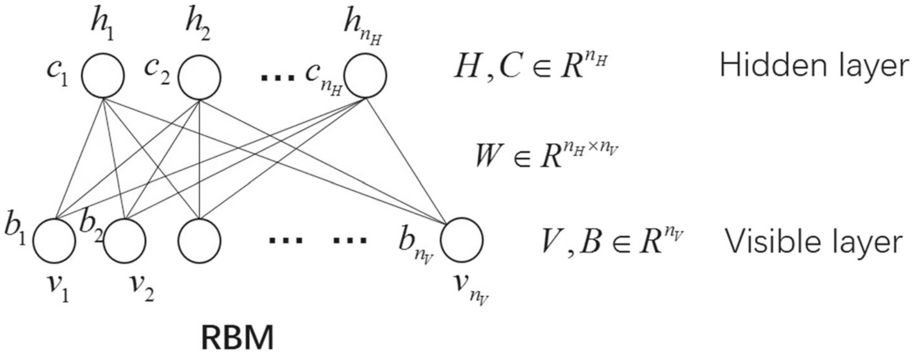

Fig. 1 Schematic illustration of RBM

In this study, we designed a stacked restricted Boltzmann machine to extract the hierarchical structures of gene expression dataset. The schematic illustration of SRBM is shown in Fig. 2. We add another RBM (denoted as RBM2 in Fig. 2) to the original RBM (denoted as RBM1 in Fig. 2). The input of visible layer in RBM2 is the output of hidden layer in RBM1. The dimension of gene expression data can be further reduced through the SRBM. As the gene expression data is real-valued data, we assume that the expression of genes obeys Gaussian distribution [15]. We use a Gaussian-Bernoulli RBM model for RBM1. However, variables in RBM2 are all binary numbers.

In the analysis of the gene expression dataset, the gene expression profile of a sample is $V = (v_1, v_2, \cdots, v_{n_V})$, where v_i represents the expression level of gene i and n_V is the number of genes. Here, v_i represents visible variable and V represents a layer of visible variables. $H = (h_1, h_2, \cdots, h_{n_H})$ denotes the layer of hidden variables, where h_j represents hidden variable and n_H is the number of hidden variables. The weight of the corresponding connection between hidden variable h_j and visible variable v_i is w_{ji}. The weight matrix $W = [w_{ji}]_{n_H \times n_V}$ represents the parameter setting of weights between the hidden layer and the visible layer. Let $B = (b_1, b_2, \cdots, b_{n_V})$ be the bias vector of visible layer, where b_i stands for the bias of visible variable v_i. Let $C = (c_1, c_2, \cdots, c_{n_H})$ be the bias vector of hidden layer, where c_j stands for the bias of hidden variable h_j.

In RBM1 (Gaussian-Bernoulli RBM), the conditional distribution over the visible variables is usually supposed to be a Gaussian distribution whose mean is a function of the hidden variables [25, 26]. The conditional probability of a visible variable is

$$p_\theta(v_i|H) = \mathcal{N}\left(\sum_{j=1}^{n_H} h_j w_{ji} + b_i, \sigma_i^2\right), \tag{1}$$

where $\theta = (W, B, C)$ represents the parameter setting of the model. Symbol σ_i is the standard deviation of Gaussian noise in visible variable v_i.

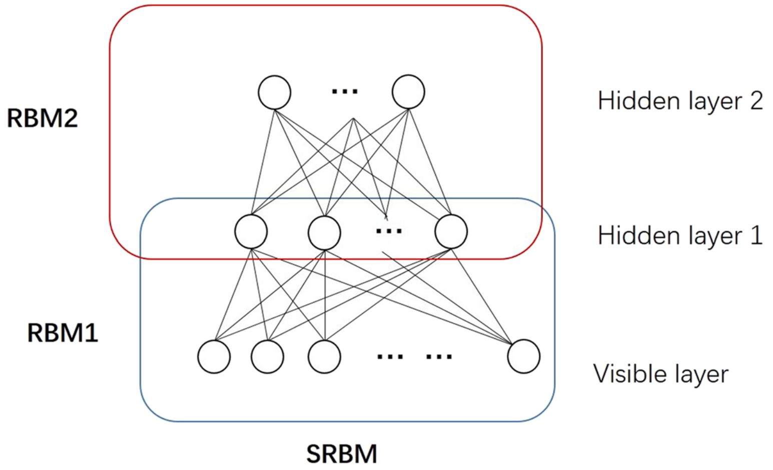

Fig. 2 Schematic illustration of SRBM

The energy function of the RBM1 with binary hidden variables and real-valued visible variables can be defined as

$$E_\theta(V,H) = \sum_{i=1}^{n_V} \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_{j=1}^{n_H} c_j h_j - \sum_{i=1}^{n_V} \sum_{j=1}^{n_H} \frac{v_i}{\sigma_i^2} h_j w_{ji}. \tag{2}$$

To simplify the parameter learning of the model, we standardized the input gene expression dataset, i.e., the average value of the visible variables v_i is equal to 0 and the variance of that is equal to 1 ($\sigma_i = 1$). In this way, the energy function in Eq. 2 can be rewritten as

$$E_\theta(V,H) = \sum_{i=1}^{n_V} \frac{(v_i - b_i)^2}{2} - \sum_{j=1}^{n_H} c_j h_j - \sum_{i=1}^{n_V} \sum_{j=1}^{n_H} v_i h_j w_{ji}. \tag{3}$$

The joint probability density function of (V,H) is given by

$$p_\theta(V,H) = \frac{1}{Z(\theta)} e^{-E_\theta(V,H)}, \tag{4}$$

where $Z(\theta)$ is a normalizing constant known as the partition function, $Z(\theta) = \sum_{V,H} e^{-E_\theta(V,H)}$. It is important to state that the variables are under independent identical distribution. We need to get the conditional probability distribution of the visible variables due to the unobservability of the hidden layer, thus to solve the model. The edge probability distribution of the visible variables is given by

$$p_\theta(V) = \sum_H p_\theta(V,H) = \frac{1}{Z(\theta)} \sum_H e^{-E_\theta(V,H)}. \tag{5}$$

Since the gene expression data are very noisy, we discretized the gene expression values into binary values during the Gibbs sampling process. And we used binary activations instead of the real-valued visible units sampled from a Gaussian distribution which are usually seen as their activations. Because a binary activation contains less information than a real-valued gene expression, using the binary activation to represent a gene expression is helpful to distinguish the genes. This is a straightforward way to reduce noise in the gene expression data. The conditional probability density distributions can be easily obtained according to Eqs. 4 and 5. (The detail derivation process is given in Additional file 1).

$$p\left(h_k = 1|V\right) = \frac{1}{1 + e^{-\left(c_k + \sum_{i=1}^{n_V} w_{ki} v_i\right)}}, \tag{6}$$

$$p(v_k = 1|H) = \frac{1}{1 + e^{-\left(-0.5 + b_k + \sum_{j=1}^{n_H} h_j w_{jk}\right)}}. \tag{7}$$

In RBM2, $v = (v_1, v_2, \cdots, v_{n_v})$ represents the input layer (hidden layer 1 in Fig. 2) and $h = (h_1, h_2, \cdots, h_{n_h})$ denotes the output layer (hidden layer 2 in Fig. 2). The weight of the corresponding connection between output variable h_j and input variable v_i is w_{ji}. The weight matrix $w = [w_{ji}]_{n_h \times n_v}$ represents the parameter setting of weights between the output layer and the input layer. Let $b = (b_1, b_2, \cdots, b_{n_v})$ be the bias vector of input layer, where b_i stands for the bias of variable v_i. Let $c = (c_1, c_2, \cdots, c_{n_h})$ be the bias vector of output layer, where c_j stands for the bias output variable h_j.

As the variables in RBM2 are all binary, the energy function of the RBM2 model is defined as

$$E_\theta(v,h) = -\sum_{i=1}^{n_v} b_i v_i - \sum_{j=1}^{n_h} c_j h_j - \sum_{i=1}^{n_v} \sum_{j=1}^{n_h} h_j w_{ji} v_i. \tag{8}$$

In the same way, we get the following conditional probability density distributions

$$p\left(h_k = 1|v\right) = \frac{1}{1 + e^{-\left(c_k + \sum_{i=1}^{n_v} w_{ki} v_i\right)}}, \tag{9}$$

$$p\left(v_k = 1|h\right) = \frac{1}{1 + e^{-\left(b_k + \sum_{j=1}^{n_h} h_j w_{jk}\right)}}. \tag{10}$$

Learning

Training the RBM model means to learn the parameters of the model, making sure that the probability density distribution of the hidden variables fit that of the variables in the visible layer well. Physically, the energy function of the system is minimized when the system reaches a steady state. Mathematically, the goal of RBM training is to maximize the logarithmic likelihood function. For such a type of optimization problem, we use gradient up method to learn the parameters of the model.

$$\theta := \theta + \eta \frac{\partial log p_\theta(V)}{\partial \theta}, \tag{11}$$

$$\frac{\partial log p_\theta(V)}{\partial \theta} = -\left\langle \frac{\partial E_\theta(V,H)}{\partial \theta} \right\rangle_{p_\theta(H|V)} + \left\langle \frac{\partial E_\theta(V,H)}{\partial \theta} \right\rangle_{p_\theta(V,H)}, \tag{12}$$

where η is learning rate, $\left\langle \frac{\partial E_\theta(V,H)}{\partial \theta} \right\rangle_{p_\theta(H|V)}$ is the expectation of energy gradient function $\frac{\partial E_\theta(V,H)}{\partial \theta}$ under the condition distribution $p_\theta(H|V)$, and $\left\langle \frac{\partial E_\theta(V,H)}{\partial \theta} \right\rangle_{p_\theta(V,H)}$ is the expectation of energy gradient function under the joint distribution $p_\theta(V,H)$. Since the hidden variables cannot be directly observed, we use CD-k algorithm to approximately estimate the probability $p_\theta(V)$ though Gibbs sampling in k steps [21, 22], thus to obtain the solution of $\left\langle \frac{\partial E_\theta(V,H)}{\partial \theta} \right\rangle_{p_\theta(V,H)}$. For sample V, the initial values of visible layer is $V^{(0)} = V$. We use $V^{(k)}$ to denote the sample obtained by CD-k.

The gradients for sample V in one iterative process are given by (The detail derivation process is given in Additional file 1).

$$\frac{\partial logp_\theta(V)}{\partial w_{ij}} = p\left(h_i = 1|V^{(0)}\right)v_j^{(0)} - p\left(h_i = 1|V^{(k)}\right)v_j^{(k)}, \tag{13}$$

$$\frac{\partial logp_\theta(V)}{\partial b_i} = v_i^{(0)} - v_i^{(k)}, \tag{14}$$

$$\frac{\partial logp_\theta(V)}{\partial c_i} = p\left(h_i = 1|V^{(0)}\right) - p\left(h_i = 1|V^{(k)}\right). \tag{15}$$

In this study, we use mini-batch strategy to learn parameters in the RBM. We use sample set $S = \{V^1, V^2, \cdots, V^n\}$ to train the model one batch. Here $n_{block} = n$ represents the size of mini-batch. The gradient calculation formula for one iteration is shown below

$$\frac{\partial logL_s}{\partial \theta} = \sum_{t=1}^{n} \frac{\partial \left(logp(V^t)\right)}{\partial \theta}, \tag{16}$$

where $L_s = p_\theta(S)$ is the likelihood function of product edge probability density distributions, V^t represents the t-th sample. The gradients for S in one iteration are given by

$$\frac{\partial logL_s}{\partial w_{ij}} = \sum_{t=1}^{n}\left[p\left(h_i = 1|V^{t(0)}v_j^{t(0)} - p\left(h_i = 1|V^{t(k)}\right)v_j^{t(k)}\right], \tag{17}$$

$$\frac{\partial logL_s}{\partial b_i} = \sum_{t=1}^{n}\left[v_i^{t(0)} - v_i^{t(k)}\right], \tag{18}$$

$$\frac{\partial logL_s}{\partial c_i} = p\left(h_i = 1|V^{t(0)}\right) - p\left(h_i = 1|V^{t(k)}\right). \tag{19}$$

In summary, the detail training process of the RBM is shown below.

Algorithm 1 Training for RBM

```
Input $k, J$, and sample sets $\{S_1, S_2, \cdots, S_m\}$
For $i = 1, 2, \cdots, m$
    For $iter = 1, 2, \cdots, J$
        $CD - k(k, S_i, n_V, n_H, RBM(W, B, C); \Delta W, \Delta B, \Delta C)$
        $W = W + \eta\left(\frac{1}{n_{block}}\Delta W\right), B = B + \eta\left(\frac{1}{n_{block}}\Delta B\right),$
        $C = C + \eta\left(\frac{1}{n_{block}}\Delta C\right)$
    End
End
```

We trained the stacked restricted Boltzmann machine in a greedy layer-wise fashion [23]. We first trained the RBM1 according to the above training process (see Algorithm 1), then trained RBM2 in the same way.

Identification of key genes

In our study, the regulatory factors are seen as high-level features which could be captured by the hierarchical structure and narrow hidden layers of the SRBM. On the one hand, the differentially activated hidden neurons are important for distinguishing different disease stage samples. On the other hand, the neurons differential activation indicates that the regulatory factors change greatly during the disease development. So, we select disease-related regulatory factors according to the differentially activated neurons in the hidden layers.

Biologically, the connections among neurons in one functional neural circuit are more strong. In fact, it has also been shown that the high-level hidden units in RBM tend to have strong positive weights to similar features in the visible layer [27]. In an SRBM model, the connections from a visible unit in the input layer to the high-level features (disease-related regulator factors) are seen as the connections in a functional neural circuit. And we use the energy of the neural circuit in the SRBM to measure the property of the input unit (represent a gene). Since the hidden units were activated very differently along with the disease progression, the energy of the neural circuit changed greatly. It suggests that the gene expression has been greatly affected during the disease development. Based on the above analysis, we rank the genes according to the energy changes at different time periods. The higher the ranking of gene it is, the more likely the disease-related gene it is.

Let x_i^s denote the activated frequency of neuron i in the first hidden layer, using the gene expression data of s time period samples. Symbol y_j^s denotes the activated frequency of neuron j in the second hidden layer, i.e., the output layer. Let E_g^s denote the energy of gene g at s time period. According to Eqs. 3 and 8, the energy of gene g is given by

$$E_g = \frac{\left(v_g - b_{1,g}\right)^2}{2} - \sum_{j=1}^{n_H} h_{1,j}w_{1,jg}v_g - \sum_{i=1}^{n_v} b_{2,i}v_{2,i} - \sum_{i=1}^{n_v}\sum_{j=1}^{n_h} h_{2,j}w_{2,ji}v_{2,i}, \tag{20}$$

where $b_{1,i}$, $h_{1,i}$, $w_{1,ji}$ represent the parameters in RBM1 and $b_{2,i}$, $v_{2,i}$, $h_{2,i}$, $w_{2,ji}$ represent the parameters in RBM2. Since the energy caused by the bias of the hidden layer in RBM1 is same for all genes, we omit the term in the calculation formula of gene energy.

The energy change of gene g at different time periods is computed by

$$C_g = \left| \frac{1}{|s_1|} \sum_{i=1}^{s_1} E_g^{s_1} - \frac{1}{|s_2|} \sum_{i=1}^{s_2} E_g^{s_2} \right|, \tag{21}$$

where s_i denotes the samples at i time period. The details for identifying key genes are shown below:

Step 1. Rank the two hidden layer neurons in descending order according to the difference of the activated frequency between different time periods, respectively. We select the top ranked neurons in the ranked lists as the differentially activated neurons, respectively. The neurons that are not differentially activated in the two hidden layers are all set to 0 in any case.

Step 2. Compute the energy changes of gene g at different time periods according to Eq. 21. Rank genes in descending order according to the energy changes of genes.

Parameter setting

Here, we initialize parameters in SRBM according to empirical studies in deep learning literature. The initialization weights obey Gaussian distribution $N(0, 0.01)$. The initialization bias variables are set to 0. The learning rate $\eta = 0.5$. The number of hidden neurons is usually about one tenth of visible neurons. In this study, the number of variables in the first hidden layer is 400 and that of the second hidden layer is 20. Moreover, the number of sampling steps in CD-k is set to be $k = 1$.

Results and discussion

We used the SRBM to analyze the gene expression data of Huntington's disease mice at different time periods. In this section, first, we briefly introduce the dataset used in this study. Second, we demonstrate the experimental results using SRBM. Then, we compare the performance of SRBM with other computational methods. Finally, we analyze and discuss the results of SRBM in detail.

Gene expression data

The gene expression dataset used in this study were downloaded from http://www.hdinhd.org, which were obtained from the striatum tissue of Huntington's disease mice by using RNA-seq technology. The genotype of Huntington's disease mice is ploy Q 111. There are 8 samples of 2-month-old mice and 8 samples of 6-month-old Huntington's disease mice. We conducted a preprocessing step to filter out noisy and redundant genes by selecting the genes with large mean value and variance of the 16 samples. Finally 4433 genes from the total 23,351 genes were left for further analysis. The data of modifier genes were from [28], which contained 520 genes, including 89 disease-related genes and 431 non-disease-related genes.

The results of SRBM

Figures 3 and 4 show the energy changes of RBM1 and RBM2 along with every iteration during the parameter training process. From Figs. 3 and 4, we can see that the changes become small with the increasing of iterations. In this study, since there are large amounts of parameters in RBM1, the iteration times of RBM1 are preset to be 50 to reduce computational time and avoid over-fit. The iteration times of RBM2 are preset to be 400 to avoid over-fit.

We statisticed the differentially activated frequency of neurons in the hidden layers using SRBM with gene expression datasets at different time periods. The results are shown in Table 1. Compared with the differentially activated frequency of neurons in the hidden layer 1, that in the hidden layer 2 is much larger. The number of neurons, whose differentially activated frequency in hidden layer 1 is 3, is too small to be used to distinguish samples at different time periods. It is better to use the neurons with largest differentially activated frequency in the hidden layer 2 to distinguish samples at different time periods, thus to identify the key genes that may be seriously affected during the disease progression.

Furthermore, we draw heatmaps of the weight matrices of RBM2 to investigate the deep structure difference between the gene expression data of Huntington's disease mice at different time periods. The weight matrices are obtained by using SRBM with gene expression datasets of Huntington's disease mice at different time periods (Figs. 5 and 6). The numbers in the left of the heatmap represent the corresponding neuron in the output layer. From Figs. 5 and 6, we can clearly see that there are significant difference between the two heatmaps. It suggests that the gene expression changes complicatedly during the disease progression.

Performance comparison between SRBM with other methods

To verify the performance of SRBM, we conducted other experiments using the original RBM method, t-test method [10], fold change rank-product method (FC-RP) [10], and joint non-negative matrix factorization meta-analysis method (jNMFMA) [11] with the gene expression data. We use true positive rate (TPR), false positive rate (FPR), precision, and recall to evaluate the prediction accuracy of disease-associated genes. TPR is defined as the ratio of correctly predicted disease genes to all disease genes. FPR is defined as the ratio of incorrectly predicted disease genes to all non-disease genes. Precision is defined as the ratio of correctly predicted disease genes to all the predicted disease genes. Recall is defined as the ratio of

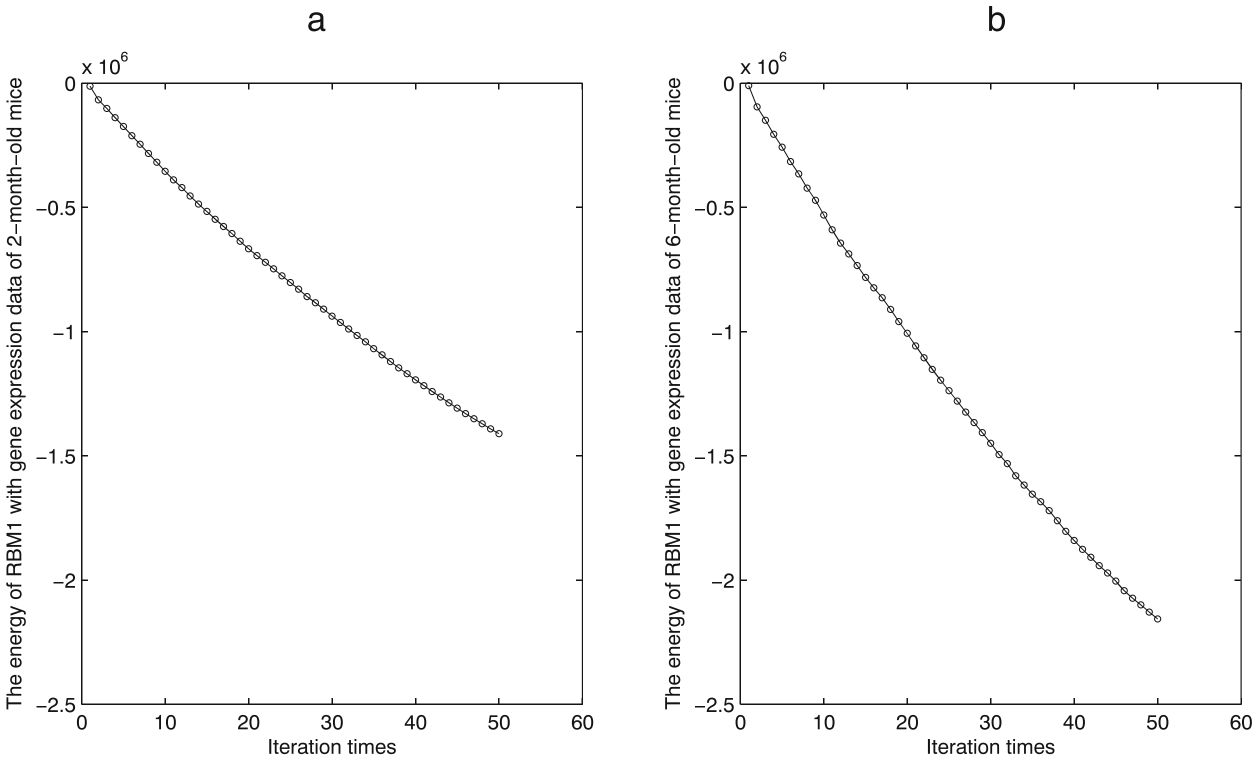

Fig. 3 The energy change of RBM1. **a** The energy change of RBM1 with gene expression data of 2-month-old Huntington's disease mice. **b** The energy change of RBM1 with gene expression data of 6-month-old Huntington's disease

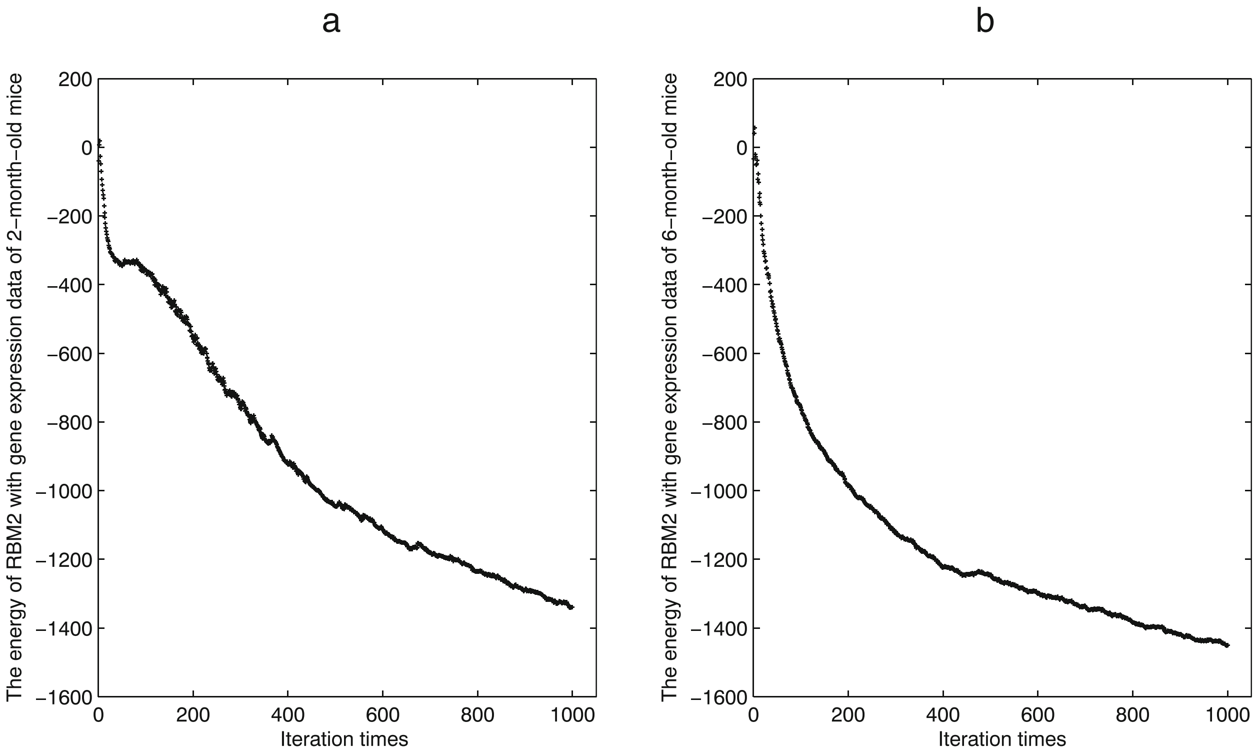

Fig. 4 The energy change of RBM2. **a** The energy change of RBM2 with gene expression data of 2-month-old Huntington's disease mice. **b** The energy change of RBM2 with gene expression data of 6-month-old Huntington's disease

Table 1 The number of neurons that are of the same differentially activated frequency using SRBM with different time period samples

Differentially activated frequency	Hidden layer 1	Hidden layer 2
5	0	5
4	0	2
3	4	3
2	57	3
1	199	4
0	140	3

correctly predicted disease genes to all disease genes. The receiver operating characteristic (ROC) curves were created by plotting TPR versus FPR. The precision-recall (PR) curves were created by plotting precision versus recall. The area under the ROC curve (AUC) and the area under the precision-recall curve (AUPR) were used as measures of the prediction accuracy [29].

To test the reasonability of the assumption in this study, we used all neurons in hidden layers to compute the gene energy while overlooking one third weak connections that from one neuron to all the neurons of the next layer. The corresponding experiments are denoted as SRBM-I . On the other hand, we selected differentially activated neurons at different time periods as factors that manipulate the expression of all genes during the disease progression, 61 neurons were selected in the first hidden layer with differentially activated frequency larger than 1, and 5 neurons were selected in the second hidden layer with differentially activated frequency larger than 5. Then, we computed the energy for each gene. The corresponding experiments are denoted as SRBM-II. Note that we use RBM-I and RBM-II to denote the experiments using the original RBM model.

From Fig. 7, we can see that the ROC cures of the seven methods are similar. The AUCs of these methods are around 0.5. It illustrates that these methods cannot separate the disease genes from non-disease genes in the modifier gene set. It also indicates that the expression of genes change complicatedly during the disease development. Nevertheless, the AUC of SRBM-II is mildly improved compared with that of the other six methods.

From Fig. 8, the PR curves of the seven methods are similar to some extent. However, the prediction precision for top ranked genes of the seven methods are clearly distinct. The prediction precision of SRBM-II is significantly

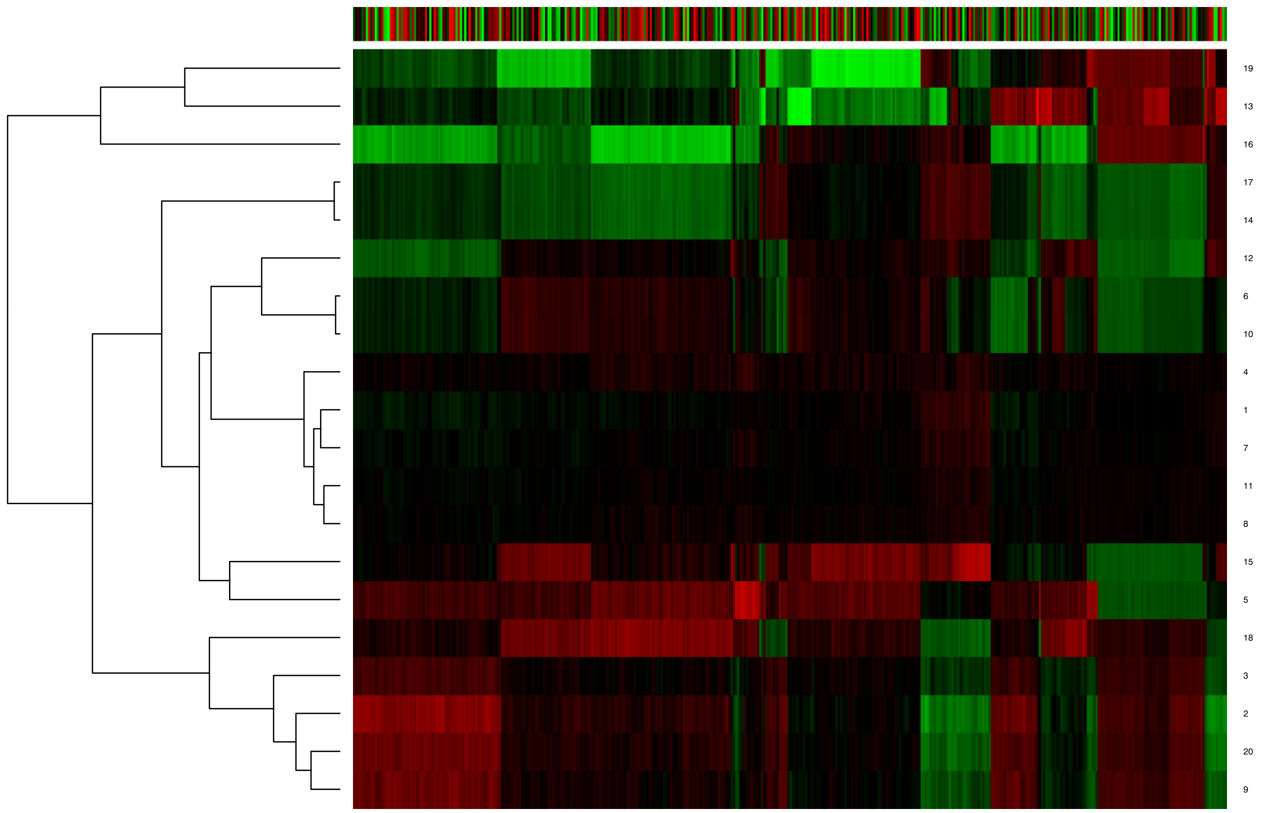

Fig. 5 Heatmap of weight matrix of RBM2 with 2-month-old gene expression data. The weight matrix is obtained using SRBM with gene expression data of 2-month-old Huntington's disease mice

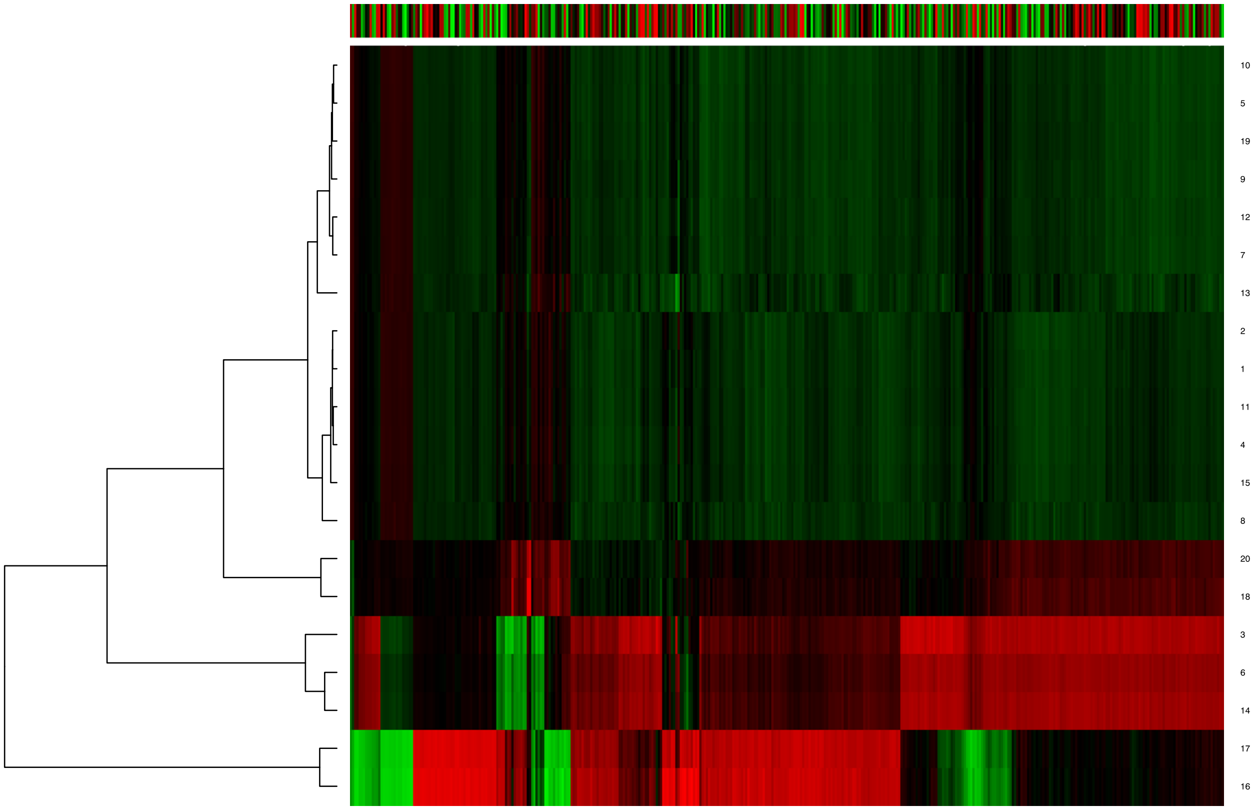

Fig. 6 Heatmap of weight matrix of RBM2 with 6-month-old gene expression data. The weight matrix is obtained using SRBM with gene expression data of 6-month-old Huntington's disease mice

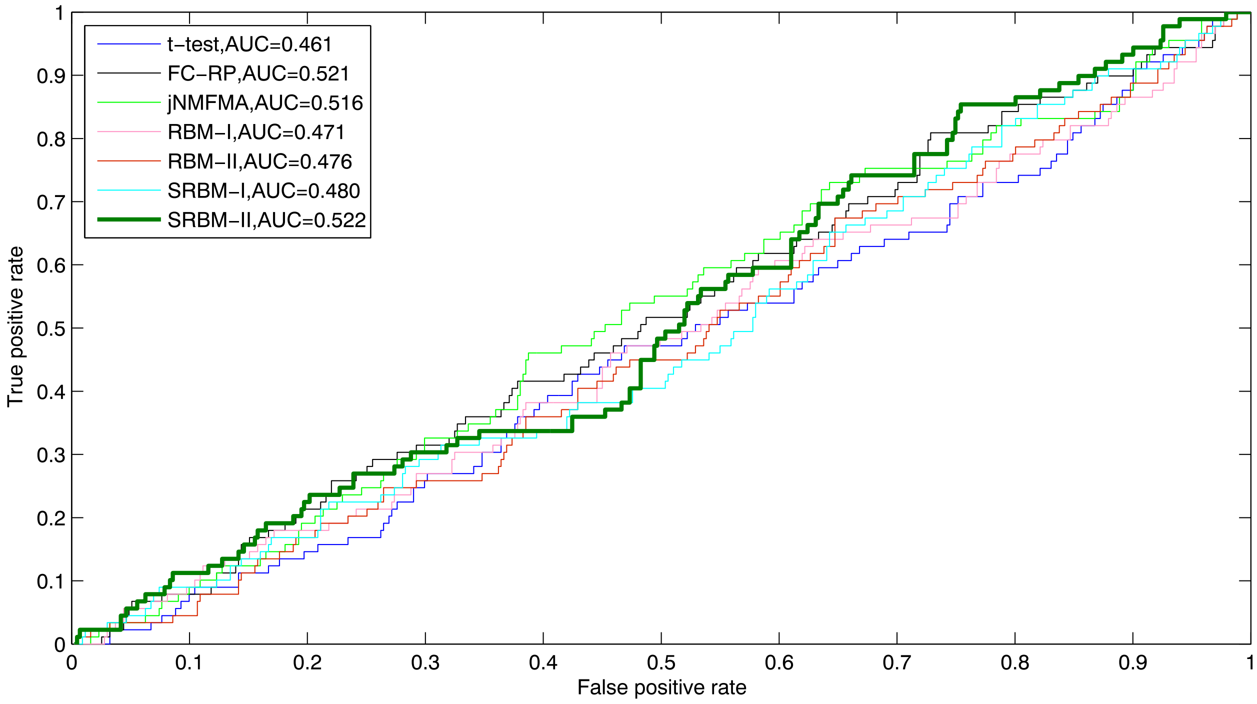

Fig. 7 ROC curves. The ROC curves of the prediction results using t-test, FC-RP, jNMFMA, RBM-I, RBM-II, SRBM-I and SRBM-II

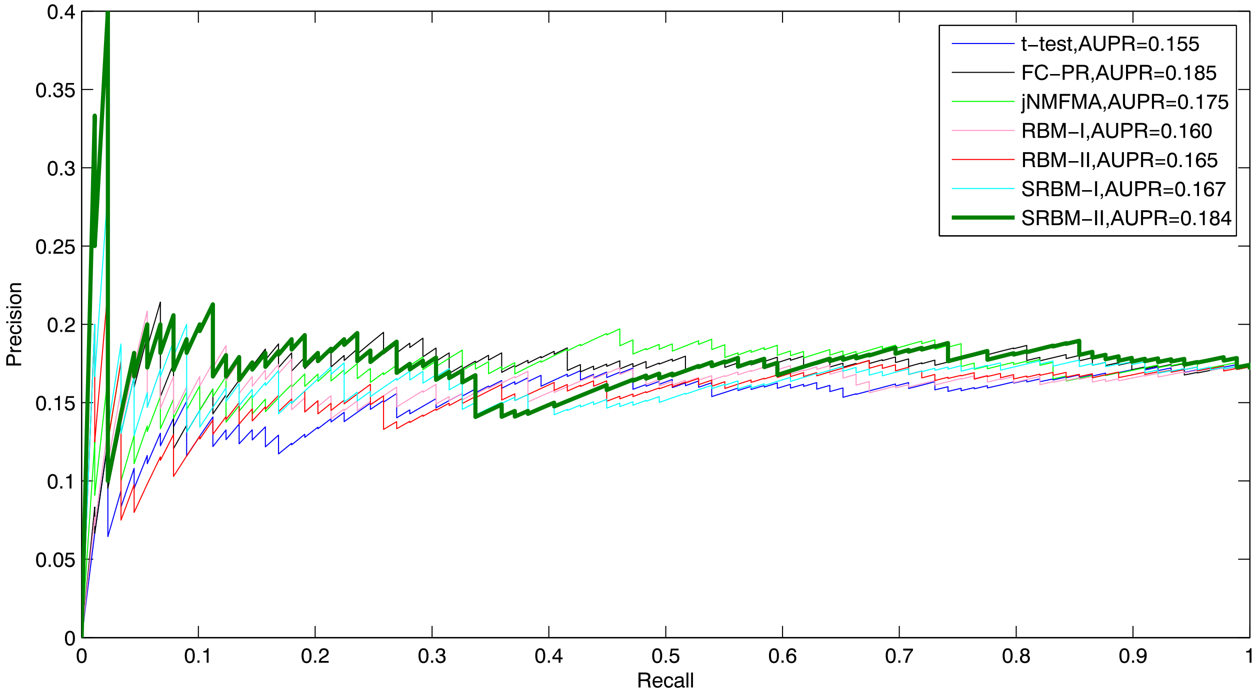

Fig. 8 Rank-product curves. The RP curves of the prediction results using t-test, FC-RP, jNMFMA, RBM-I, RBM-II, SRBM-I and SRBM-II

higher for top ranked genes compared with that of the other six methods.

We further investigate the distributions of the rankings of top ranked 10 disease genes in the ranked lists obtained by using the seven methods, respectively (Fig. 9). From Fig. 9, we can roughly know the rankings of the top ranked disease genes. Although the distributions obtained by these methods are similar, SRBM-II makes the disease genes get mild higher rankings compared with the other six methods.

In total, the performance of SRBM-II is moderately better than other methods. From Figs. 7, 8 and 9, we can know

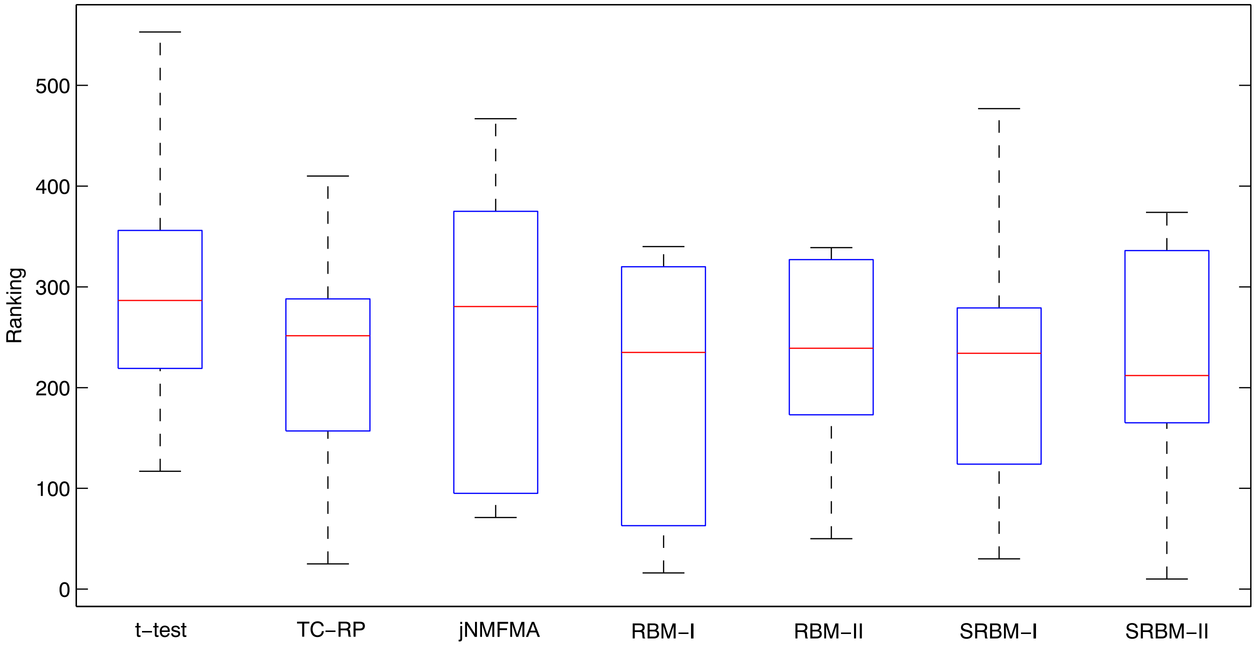

Fig. 9 Boxplots of the rankings of top ranked 10 disease genes. The rankings are obtained using different methods, including t-test, FC-RP, jNMFMA, RBM-I, RBM-II, SRBM-I and SRBM-II

Table 2 The number of overlapped genes (the degree of overlap) of top ranked 500 genes between any two ranked lists obtained using t-test, FC-RP, jNMFMA, RBM-I, RBM-II, SRBM-I, and SRBM-II

	FC-RP	jNMFMA	RBM-I	RBM-II	SRBM-I	SRBM-II
t-test	81 (16.2%)	36 (7.2%)	73 (14.6%)	74 (14.8%)	75 (15%)	73 (14.6%)
FC-RP		114 (22.4%)	28 (5.6%)	22 (4.4%)	38 (7.6%)	40 (8.0%)
jNMFMA			6 (1.2%)	8 (1.6%)	5 (1.0%)	9 (1.8%)
RBM-I				344 (68.8%)	252 (50.4%)	214 (42.8%)
RBM-II					245 (49.0%)	248 (49.6%)
SRBM-I						351 (70.2%)

that the performance of SRBM-II is better than SRBM-I. It suggests that we improved the prediction accuracy by selecting the differentially activated neurons, which are assumed to be disease-associated factors in our study. We can also know that the performance of SRBM methods are better than RBM methods. It verifies that we effectively separated some noisy factors from the gene expression dataset, using the deep structure of SRBM.

We also statisticed the overlapped degree of top ranked 500 genes between any two ranked lists, the results are shown in Table 2. It can be clearly seen that the overlapped degrees between any two ranked lists (except for that between SRBM-I and SRBM-II) are all small. However, the overlap degrees between jNMFMA and SRBM methods are smaller than that between others. The jNMFMA assumes that the gene expression is a weighted linear combination of metagenes. The jNMFMA selects disease-associated genes through differentially regulated metagenes. SRBM selects disease-associated genes according to the energy changes at different disease states. Since the basic assumptions of the two models are greatly different, the overlapped degrees of top ranked genes between the two ranked lists are smaller.

The top ranked 500 genes in different ranked lists share 4 common genes: Chmp1b, Poldip3, Lrrtm1 and Slc44a1. According to the annotation of Gene Ontology, the molecule function of Chmp1b is protein domain specific binding, that of Lrrtm1 is protein kinase inhibitor activity, that of Poldip3 is nudeotide binding, and that of Slc44a1 is choline transmembrane transporter activity. The functions of the four genes are all related to protein transportation. Those genes may be related to the

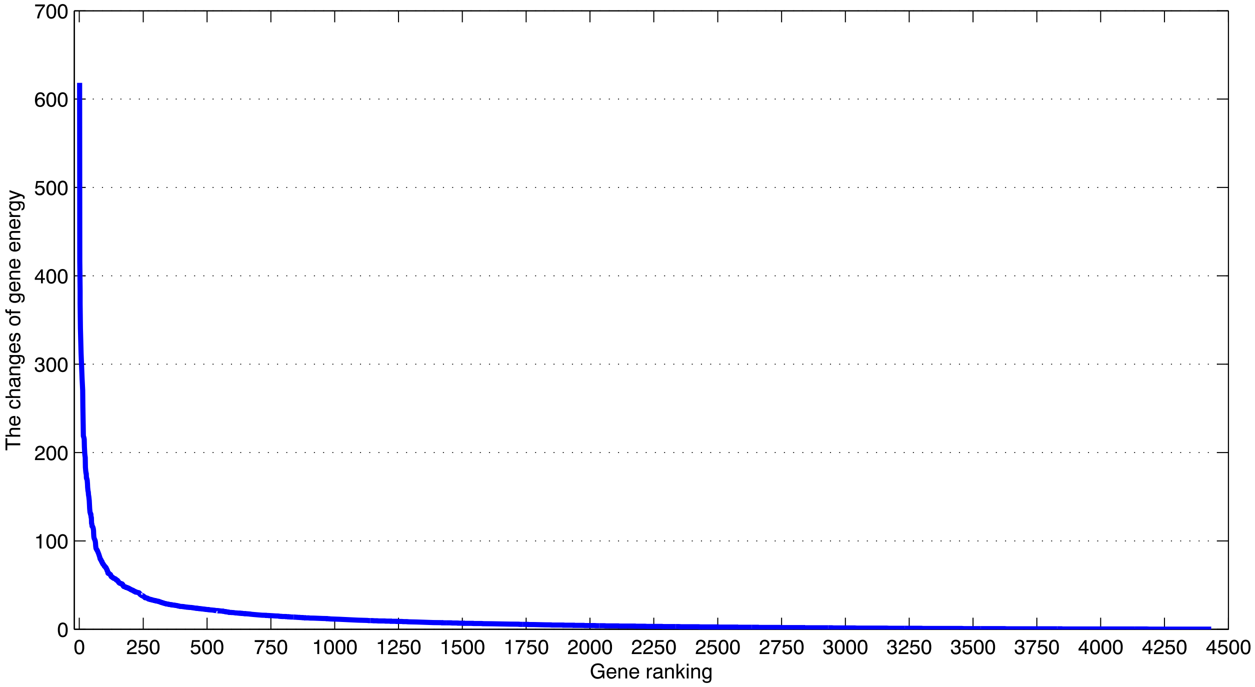

Fig. 10 The changes of gene energy. The gene ranking is obtained by using SRBM-II based on the changes of gene energy at different time periods

Table 3 The functional annotation clustering of the top ranked 100 genes in the ranked list obtained using SRBM-II

Annotation cluster	Annotation	Genes-included	*P*-value	Benjamini
Annotation cluster 1	Membrane	60	7.2E-8	7.3E-6
	Plasma membrance	42	5.0E-5	1.1E-3
Annotation cluster 2	Synapse	14	8.2E-7	3.3E-5
	Postsynaptic density	10	2.2E-6	7.3E-5
	Dendritic spine	7	7.8E-5	1.6E-3
	Cell junction	11	2.3E-3	2.5E-2
	Synaptic vesicle	4	2.3E-2	1.8E-1
	Postsynaptic membrane	4	9.0E-2	4.0E-1
Annotation cluster 3	Cell-cell adherens junction	8	8.0E-4	1.3E-2

disturbance of intracellular protein trafficking in Huntington's disease individuals [30].

Enrichment analysis

According to Fig. 10, it is obvious that the changes of gene energy for the top ranked 100 genes are significantly larger. Combined with Fig. 8, we known that the higher the ranking of gene it is, the more precise the prediction accuracy of disease-related gene it is. To avoid introducing too many false positives, we chose the top ranked 100 genes in the ranked list obtained by using SRBM-II to conduct enrichment analysis. We used the functional annotation clustering tool through DAVID [31] to annotate the functions of those genes, the result can be seen in Table 3. The annotations listed in the table are cellular component from GOTERM. From Table 3, we can see that those genes are related to membrane, synapse and cell junction. It suggests that the cellular form changes greatly during the Huntington's disease progression and deterioration. In fact, the connections between neurons get sparse, and the neurons finally died during the Huntington's disease deterioration [32, 33].

Conclusions

In this paper, we designed a stacked restricted Boltzmann machine to detect the hierarchical structures and to capture the important information for differential analyzing gene expression datasets of Huntington's disease mice at different time periods. We also proposed a new framework to identify the key genes that may be affected by the disease progression. Experimental results verify the feasibility of the assumption in this study. It also demonstrates that the performance of SRBM-II is mildly better than other traditional methods. Besides the exploratory analysis of the disease molecular mechanisms through enrichment analysis, we also conducted a integrated analysis on the ranked lists obtained by the seven methods. We found that four genes (Chmp1b, Poldip3, Lrrtm1 and Slc44a1) related to protein transportation are seriously affected during the disease progression.

Additional file

Additional file 1: Supplementary Material. The detail derivation process for solving the gradients of RBMs learning is given in the Supplementary Material. (PDF 321 kb)

Abbreviations

AUC: The area under the ROC curve; AUPR: The area under the PR curve; CD: Contrastive divergence; CNN: Convolutional neural network; FC-RP: Fold change rank-product; FPR: False positive rate; jNMFMA: Joint non-negative matrix factorization neta-analysis method; PR: Precision-recall; RBM: Restricted Boltzmann machine; ROC: Receiver operating characteristic; SRBM: Stacked restricted Boltzmann machine; TPR: True positive rate

Acknowledgements

The authors would like to thank the editor and the reviewers for their comments and suggestions, which helped improve the manuscript greatly.

Funding

This work has been supported by the Natural Science Foundation of Tianjin (15JCYBJC18900), the National Natural Science Foundation of China (31728013, 61673224, 61403213), and the Key Program of Science Foundation of Tianjin (14JCZDJC31800).

Availability of data and materials

The gene expression data used in this study were downloaded from http://www.hdinhd.org. To make the dataset available to public, we deposit it in publicly available repository, please download at https://figshare.com/s/071072960aa5132a6d2c. The modifier genes were from "Langfelder P, Cantle J P, Chatzopoulou D, et al. Integrated genomics and proteomics define huntingtin CAG length-dependent networks in mice[J]. Nature Neuroscience, 2016. PMID: 26900923 DOI: 10.1038/nn.4256". We also deposit it in publicly available repository, please download at https://figshare.com/s/13fdc5c17d736142dcd0. Moreover, we also deposit the ranked lists obtained by using SRBM-II in the publicly available repository for reference, please download at https://figshare.com/s/bf1ea56b278ef7b3f1ac.

Authors' contributions

HZ, FD and XQ conceived the research. XJ and FD designed the research. XJ performed the experiments and analyzed the data. XJ, HZ and FD wrote the manuscript. All authors reviewed the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 30 May 2017 Accepted: 2 October 2017
Published online: 11 October 2017

References

1. Appel SH, Smith RG, Le WD. Immune-mediated cell death in neurodegenerative disease. Adv Neurol. 1996;69(69):153–9.
2. Hardy J. Pathways to primary neurodegenerative disease. Ann N Y Acad Sci. 2000;924(1):29–34.
3. Luthicarter R, Apostol BL, Dunah AW, Dejohn MM, Farrell LA, Bates GP, Young AB, Standaert DG, Thompson LM, Cha JH. Complex alteration of

nmda receptors in transgenic Huntington's disease mouse brain: analysis of mrna and protein expression, plasma membrane association, interacting proteins, and phosphorylation. Neurobiol Dis. 2003;14(3): 624–36.
4. Luthi-Carter R, Strand A, Peters NL, Solano SM, Hollingsworth ZR, Menon AS, Frey AS, Spektor BS, Penney EB, Schilling G. Decreased expression of striatal signaling genes in a mouse model of Huntington's disease. Hum Mol Genet. 2000;9(9):1259–71.
5. Romanoski CE, Lee S, Kim MJ, Ingram-Drake L, Plaisier CL, Yordanova R, Tilford C, Guan B, He A, Gargalovic PS. Systems genetics analysis of gene-by-environment interactions in human cells. Am J Hum Genet. 2010;86(3):399–410.
6. Liu Y, Zeng X, He Z, Zou Q. Inferring microrna-disease associations by random walk on a heterogeneous network with multiple data sources. IEEE/ACM Trans Comput Biol Bioinforma. 2016;PP(99):1–11. doi:10.1109/TCBB.2016.2550432.
7. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4(1):1–37.
8. Hong F, Breitling R, Mcentee CW, Wittner BS, Nemhauser JL, Chory J. Rankprod: a bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics. 2006;22(22):2825–7.
9. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013; 14(1):1–18.
10. Hong F, Breitling R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. Bioinformatics. 2008;24(3):374–82.
11. Wang HQ, Zheng CH, Zhao XM. jNMFMA: a joint non-negative matrix factorization meta-analysis of transcriptomics data. Bioinformatics. 2015;31(4):572–80.
12. Saeys Y, Abeel T, Peer YVD. Robust feature selection using ensemble feature selection techniques. In: DBLP, Machine Learning and Knowledge Discovery in Databases, European Conference, vol. 5212. 2008. p. 313–25.
13. Lecun Y, Bengio Y, Hinton GE. Deep learning. Nature. 2015;521(7553): 436–44.
14. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2015;61:85–117.
15. Liang M, Li Z, Chen T, Zeng J. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. IEEE/ACM Trans Comput Biol Bioinforma. 2015;12(4):928–37.
16. Cheng S, Guo M, Wang C, Liu X, Liu Y, Wu X. Mirtdl: a deep learning approach for miRNA target prediction. IEEE/ACM Trans Comput Biol Bioinforma. 2015;13(6):1161–9.
17. Shynrye L, Hyung JK. Prion-like mechanism in amyotrophic lateral sclerosis: are protein aggregates the key? Exp Neurobiol. 2015;24(1):1–7.
18. Lim J, Yue Z. Neuronal aggregates: formation, clearance, and spreading. Dev Cell. 2015;32(4):491–501.
19. Wang X, Huang T, Bu G, Xu H. Dysregulation of protein trafficking in neurodegeneration. Mol Neurodegener. 2014;9(1):1–9.
20. Hinton GE. A practical guide to training restricted boltzmann machines. Springer Berlin Heidelberg. 2012;9(1):599–619.
21. Carreira-Perpinan MA, Hinton GE. On contrastive divergence learning. In: Proceedings of The 10th International Workshop on Artificial Intelligence and Statistics (AISTATS), vol. 10. 2005. p. 33–40.
22. Tieleman T. Training restricted Boltzmann machines using approximations to the likelihood gradient. In: The 25th International Conference on Machine Learning (ICML 2008). Helsinki; 2008. p. 1064–71. doi:10.1145/1390156.1390290.
23. Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. Neural Comput. 2006;18(7):1527–43.
24. Smolensky P. Information Processing in Dynamical Systems: Foundations of Harmony Theory. Massachusetts: MIT Press; 1986.
25. Hinton GE. A practical guide to training restricted Boltzmann machines. Springer Berlin Heidelberg. 2012;9(1):599–619.
26. Cho KH, Raiko T, Ilin A. Gaussian-Bernoulli deep Boltzmann machine. In: The 2013 International Joint Conference on Neural Networks (IJCNN); 2013. p. 1–7. doi:10.1109/IJCNN.2013.6706831.
27. Krizhevsky A. Learning multiple layers of features from tiny images. Computer Science Department. University of Toronto, Tech. Rep; 2009.
28. Langfelder P, Cantle JP, Chatzopoulou D, Wang N, Gao F, Alramahi I, Lu XH, Ramos EM, Elzein K, Zhao Y. Integrated genomics and proteomics define Huntingtin CAG length-dependent networks in mice. Nat Neurosci. 2016;19(4):623–38.
29. Bradley AP. The use of the area under the roc curve in the evaluation of machine learning algorithms. Elsevier Sci Inc. 1997;30(7):1145–59.
30. Ratovitski T, Nakamura M, D'Ambola J, Chighladze E, Liang Y, Wang W, Graham R, Hayden MR, Borchelt DR, Hirschhorn RR. N-terminal proteolysis of full-length mutant Huntingtin in an inducible pc12 cell model of Huntington's disease. Cell Cycle. 2007;6(23):2970–81.
31. Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA. The DAVID gene functional classification tool: a novel biologica module-centric algorithm to functionally analyze large gene lists. Genome Biol. 2007;8(9):183.
32. Waldvogel HJ, Kim EH, Thu DC, Tippett LJ, Faull RL. New perspectives on the neuropathology in Huntington's disease in the human brain and its relation to symptom variation. J Huntingtons Dis. 2012;1(2):143–53.
33. Difiglia M, Sapp E, Chase KO, Davies SW, Bates GP, Vonsattel JP, Aronin N. Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neurites in brain. Science. 1997;277(5334):1990–3.