Hao *et al. BMC Bioinformatics* (2018) 19:510
https://doi.org/10.1186/s12859-018-2500-z

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

PASNet: pathway-associated sparse deep neural network for prognosis prediction from high-throughput data

Jie Hao[1], Youngsoon Kim[2], Tae-Kyung Kim[3,4] and Mingon Kang[1,2*]

Abstract

Background: Predicting prognosis in patients from large-scale genomic data is a fundamentally challenging problem in genomic medicine. However, the prognosis still remains poor in many diseases. The poor prognosis may be caused by high complexity of biological systems, where multiple biological components and their hierarchical relationships are involved. Moreover, it is challenging to develop robust computational solutions with high-dimension, low-sample size data.

Results: In this study, we propose a Pathway-Associated Sparse Deep Neural Network (PASNet) that not only predicts patients' prognoses but also describes complex biological processes regarding biological pathways for prognosis. PASNet models a multilayered, hierarchical biological system of genes and pathways to predict clinical outcomes by leveraging deep learning. The sparse solution of PASNet provides the capability of model interpretability that most conventional fully-connected neural networks lack. We applied PASNet for long-term survival prediction in Glioblastoma multiforme (GBM), which is a primary brain cancer that shows poor prognostic performance. The predictive performance of PASNet was evaluated with multiple cross-validation experiments. PASNet showed a higher Area Under the Curve (AUC) and F1-score than previous long-term survival prediction classifiers, and the significance of PASNet's performance was assessed by Wilcoxon signed-rank test. Furthermore, the biological pathways, found in PASNet, were referred to as significant pathways in GBM in previous biology and medicine research.

Conclusions: PASNet can describe the different biological systems of clinical outcomes for prognostic prediction as well as predicting prognosis more accurately than the current state-of-the-art methods. PASNet is the first pathway-based deep neural network that represents hierarchical representations of genes and pathways and their nonlinear effects, to the best of our knowledge. Additionally, PASNet would be promising due to its flexible model representation and interpretability, embodying the strengths of deep learning. The open-source code of PASNet is available at https://github.com/DataX-JieHao/PASNet.

Keywords: Sparse deep neural network, Prognosis prediction, Long-term survival prediction, Pathway-based analysis, Glioblastoma multiforme, TCGA

Background

Predicting prognosis in patients from large-scale genomic data is a fundamentally challenging problem in genomic medicine [1–3]. Along with the rapid advances of high-throughput technologies and their effectivenesses, high-dimensional genomic data provides more accurate and richer biological descriptions of clinical phenotypes of interests than ever before. Therefore, translating large-scale genomic profiles to clinical outcomes not only improves predicting patient prognosis but also helps in identifying prognostic factors and biological processes.

The capabilities of high-level biological representation and interpretation of the prognosis are often more desired in biomedical research rather than merely improving predictive performance. Pathway-based analysis is an approach that a number of studies have been investigating

*Correspondence: mkang9@kennesaw.edu
[1]Kennesaw State University, Kennesaw, USA
[2]Kennesaw State University, Marietta, USA
Full list of author information is available at the end of the article

to improve both predictive performance and biological interpretability [4–6]. In pathway-based analyses, the incorporation of biological pathway databases in a model takes advantage of leveraging prior biological knowledge so that potential prognostic factors of well-known biological functionality can be identified. Pathway-based analyses identify biological links between pathways and clinical outcomes and enable the interpretation of biological processes where their corresponding genes and proteins are involved. Thus, pathway-based interpretation and visualization provide an intuitive and comprehensive understanding of functionally-related molecular mechanisms.

Moreover, pathway-based approaches have shown more reproducible analysis results than gene expression data analysis alone [4, 7–10]. High-level representations of gene co-expressions are considered in most pathway-based analyses; each of which represents a biological pathway while preserving the original information. Thus, pathway-based analyses remedy the limitations of gene expression data, which are intrinsically sensitive to stochastic fluctuations and are often caused by multiple potential sources, such as inherent stochasticity of biochemical processes, environmental differences, and genetic mutation [11]. Pathway-based markers were proposed for classifying breast cancer metastasis and ovarian cancer survival time [5]. Cancer subtypes were discovered with pathway-based markers via Restricted Boltzmann Machine (RBM) [8]. A group LASSO-based approach associated genes with pathways and characterized them based on biological pathways [10]. Higher-order functional representation of pathway-based metabolic features provided reproducible biomarkers for breast cancer diagnosis [9].

However, reliable and accurate prognosis still remains poor in many diseases due to the following challenges: high-dimension, low-sample size data and complex nonlinear effects between biological components.

Genomic data are highly dimensional relative to their sample sizes. High-dimension, low-sample size (HDLSS) data often make prediction models sensitive to noise and false positive associations, which consequently make predicting accurate prognoses difficult. LASSO-based approaches have been mainly considered to estimate the effects of a gene set that are associated with various types of clinical outcomes on HDLSS data. The LASSO-based approaches embed sparse coding schemes into linear or logistic regression models for selecting few but greatly informative features among the high-dimensional data. For instance, a logistic regression with sparse regularization was applied for the prognostic model of mortality after acute myocardial infarction [12]. Random LASSO was proposed to enhance the LASSO solution by applying multiple bootstrapping and was applied to predict patients' survival times with glioblastoma gene expression data [13]. LASSO-based regression models as a prediction model were validated with multiple imputed data in chronic obstructive pulmonary disease patients [14].

Pathway-based analysis also helps to reduce data dimensionality. The number of biological pathways is relatively smaller than the number of genes, and a set of genes in the same pathway can be represented by the pathway's effect. Thus, pathways can be used as summary variables for the input of the predictive model instead of including all genes, which consequently reduces the model complexity.

Most association studies between a gene set and various clinical outcomes have considered linear or logistic regression models for identifying prognostic factors as well as understanding a biological mechanism of the progression of disease. However, nonlinear effects of genes or pathways may fail to be identified by linear-based approaches. As a solution, kernel-based models have been proposed to capture nonlinear effects of complex pathways [15, 16]. Multiple kernel learning models were introduced to aggregate complex effects from multiple pathways [17, 18]. Kernel Principle Component Analysis (KPCA) was applied to reduce the dimensionality of the feature space by using the correlation structure of the pathways [18].

Recently, several attempts to capture hierarchical effects of genes and pathways have been made. Inferences of multilayered hierarchical gene regulatory networks have been considered to understand how pathways regulate each other hierarchically. A bottom-up graphic Gaussian model [19] and a recursive random forest algorithm [20] were proposed to construct multilayered hierarchical gene regulatory networks. Moreover, complex biological networks were modeled by inferring the multiple hierarchical models (1) between gene expression and pathways and (2) within pathways [21]. However, complex hierarchical relationships between pathways have not been considered for prognostic studies yet, to the best of our knowledge, although hierarchical effects of pathways are prevalent in biological systems [22].

In this paper, we propose a Pathway-Associated Sparse Deep Neural Network (PASNet) to achieve the goals: (1) to predict prognosis in patients accurately by incorporating biological pathways, (2) to provide a solution for hierarchical interpretation of nonlinear relationships between biological pathways of disease systematically, and (3) to handle computational problems on HDLSS data with unbalanced classes. An innovative aspect of our model is biological interpretability; we achieved this with sparse coding and by constructing hidden layers with biological pathways, which oppose the *black box* nature of deep learning. Our new sparse deep learning architecture represents multiple molecular biological layers, such as a gene layer and a pathway layer, along

with their hierarchical relationships, which use sparse regularization.

Results

Pathway-Associated Sparse Deep Neural Network (PASNet) identifies a subset of genes and pathways involved in a disease as prognostic biomarkers, as well as their interactions. PASNet models a multilayered, hierarchical biological system of genes and pathways on a disease, while leveraging the strengths of deep learning for competitive predictive performance. The sparsity of PASNet allows one to interpret the model, which is what conventional fully-connected networks lack. The architecture of PASNet and the strategies for training a sparse neural network model with HDLSS and imbalanced data are described in "Methods" section.

We conducted experiments to evaluate PASNet's predictive performance for long-term survival prediction in Glioblastoma multiforme (GBM). The capability of the prediction was assessed by comparing our model with the classifiers that have been used for long-term survival prediction. Furthermore, we will describe how PASNet can represent the biological system of GBM in the following section.

Data

GBM is a primary brain cancer that shows poor prognosis performance due to the above challenges. Comprising more than half of all brain tumors, GBM is the most prevailing and aggressive malignant type of primary astrocytomas [23]. Patients with GBM have a median survival time of approximately 15 months with intensive treatments [24]. Furthermore, long-term survival patients with GBM are rare as more than 90% of patients are deceased within three years of diagnosis. Although treatments in neurosurgery, chemotherapy, and radiotherapy have improved, the prognosis of GBM remains poor [25]. Hence, the advancement in understanding molecular mechanisms and related biological pathways of GBM is significant to accelerating the progress for new treatments [24].

We used the gene expression data of GBM patients, which is available at The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov). The dataset includes the gene expression data of 522 samples and 12,042 genes and provides survival time and status. We considered patients who survived past 24 months (regardless of survival status) as long-term survivals (LTS) and patients that deceased in less than 24 months as short-term survivals (non-LTS). Living patients with a survival time of less than 24 months were excluded in the experiments and considered censored data. Finally, we obtained 99 LTS and 376 non-LTS samples, where around 20% of the samples were LTS patients.

For pathway-based analysis, we utilized a biological pathway database from the Molecular Signatures Database (MSigDB) [26]. In MSigDB, we extracted the biological pathways of Reactome. Then, we excluded the pathways that include less than ten genes, because small pathways are often redundant with larger pathways. As the input features, we considered the genes that belong to at least one pathway, since pathway annotations of genes are essential to construct the mask matrix $\mathbf{M}$ between the gene layer and the pathway layer. Finally, we considered 574 pathways and 4359 genes in the experiments. The gene expression data were standardized to a mean of zero and a standard deviation of one.

Experimental setting

We followed a typical design of conventional deep neural networks for PASNet. A sigmoid function and cross-entropy were considered for the activation and the cost function, respectively. A softmax function was used in the output layer so that the probabilities of output nodes add up to one. For the optimal tuning of PASNet's training, we empirically determined the hyper-parameters by random search before cross-validation experiments. The learning rate (η) was set to 1e−4, and L^2 regularization (λ) was set to 3e−4. Adaptive Moment Estimation (Adam) was performed as the stochastic optimizer [27]. The dropouts for two intermediate layers were also applied with a dropping probability of 0.8 and 0.7, respectively. PASNet was implemented by PyTorch, and the source code is available at https://github.com/DataX-JieHao/PASNet.

Comparison

We evaluated PASNet by comparing the performance with classifiers that have been used for prognosis prediction: Support Vector Machine (SVM), Random LASSO [13], LASSO Logistic Regression (LLR) [1], and neural network with dropout (Dropout NN).

Specifically, we used a SVM with a radial basis function (RBF) kernel ($\gamma = 2^{-16}$ and $C = 2^{3.9}$ by two-step grid search [28]). Random LASSO was trained so that every feature could be selected 20 times on average by bootstrapping, and the L^1 regularization parameter was determined by 10-fold cross-validation. The LASSO parameter for LLR was also selected by 10-fold cross-validation. The fully-connected Dropout NN was designed with the same numbers of intermediate layers and neurons as the proposed PASNet as well as the dropout probabilities. The learning rate was 0.01 and the L^2 regularization was 0.005. Note that PASNet has less number of weights to be trained in each epoch because of sparse coding, compared to Dropout NN. Hence, the optimal hyper-parameters of L^2 regularization and learning rate should be different between PASNet and Dropout NN.

We empirically searched the optimal hyper-parameters for PASNet and Dropout NN separately through multiple experiments. Dropout NN was implemented by PyTorch (https://pytorch.org/).

The experiments were carried out by stratified 5-fold cross-validation for maintaining the same proportions of the imbalanced samples in the classes. The cross-validation experiments were repeated ten times for performance reproducibility. Data preprocessing, such as data normalization, was separately applied on each fold. The testing data on each fold was scaled with the mean and standard deviation of the training data of the same fold.

The predictive performances of the five models were evaluated with two metrics: Area Under the Curve (AUC) and F1-scores. The Receiver Operating Characteristic (ROC) curve (see Fig. 1) was traced over the thresholds of scores to examine the trade-off between True Positive Rate ($TPR = TP/(TP + FN)$) and False Positive Rate ($FPR = FP/(FP + TN)$), where LTS was considered positive. An AUC was computed by the area under the ROC curve. An F1-score, an average of Positive Predicted Value ($PPV = TP/(TP + FP)$) and TPR, is calculated by $2(PPV \times TPR)/(PPV + TPR)$. The F1-score was computed for the LTS class.

The average AUC and the average F1-score of the five methods on the test datasets are shown in Table 1. PASNet outperformed others as both AUC and F1-score are relatively high. PASNet produced AUC of 0.6622±0.013 (mean±std) and F1-score of 0.3978±0.016. Following PASNet, Dropout NN produced AUC of 0.6408±0.014, and SVM produced AUC of 0.6337±0.015.

To statistically assess the performance of PASNet (AUC) as compared to others, we conducted the Wilcoxon signed-rank test: a non-parametric paired, two sided test for the null hypothesis that states the median difference in paired samples is zero. Specifically, the null hypothesis is that the benchmark classifier has equal or better performance than our proposed algorithm. Table 2 shows the performance of PASNet is significantly better than others, where the null hypotheses are rejected at the 5% significance level (p-value <0.05). Hence, the outperformance of PASNet was statistically significant compared to the benchmark classifiers.

SVM and Dropout NN showed a higher AUC than LASSO logistic regression and Random LASSO, probably because of their capability of capturing nonlinear effects of genes. Compared to Dropout NN, PASNet is a relatively thin network, where the connections between layers are very sparse. However, PASNet interestingly produced higher performance than Dropout NN. It shows that PASNet builds a robust network model, which is simplified to represent the biological processes for prognosis prediction by incorporating biological prior knowledge.

Discussion

Although PASNet yielded competitive predictive performance in the experiments, a more promising contribution of PASNet is in the model's interpretability. In this section, we demonstrate a plausible biological mechanism inferred by PASNet for long-term survival prediction in GBM. The graphical representations of the PASNet model are illustrated in Figs. 2, 3 and 4 in the top-down order. The

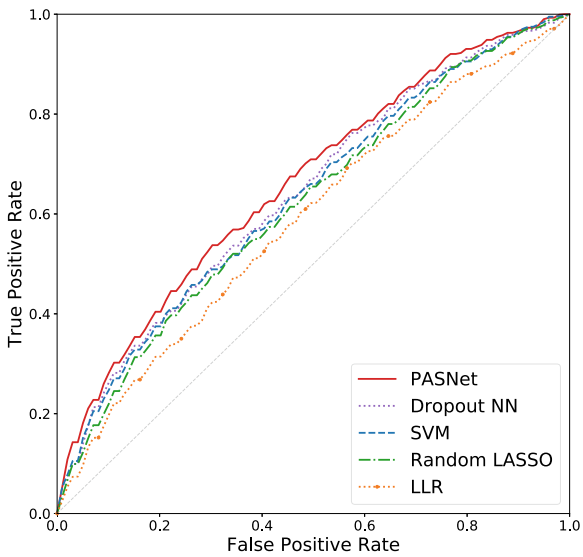

Fig. 1 ROC Curves. PASNet produces the highest AUC of 0.6622 while the AUC of Dropout NN, SVM, random LASSO, and LLR is 0.6408, 0.6337, 0.6209, and 0.5899, respectively

Table 1 Comparison of AUC and F1-score in over ten stratified 5-fold cross-validations

Model	AUC	F1-Score
Logistic LASSO	0.5899±0.020	0.3347±0.025
Random LASSO	0.6209±0.020	0.3370±0.020
SVM	0.6337±0.015	0.3446±0.015
Dropout NN	0.6408±0.014	0.2957±0.025
PASNet	0.6622±0.013	0.3978±0.016

heatmaps were generated by sorting the weights and node values of LTS, and positive and negative weight values are colored in red and blue, respectively.

First, Fig. 2 manifests the posterior probability of the samples in the clinical outcomes. The dark block on the top shows the output node values ($-log_2$(node value)) of the LTS samples, while the remaining ones are non-LTS samples. The weight values of the connections from hidden nodes to the output nodes are depicted in Fig. 3a, where dropped connections are colored in white. The figure reveals distinct patterns of weights (opposite signs) to the two output neurons. Note that there are hidden nodes disconnected to the neurons in the output layer (colored in white) by sparse coding, which shows that the hidden nodes are insignificant.

The hidden node values of the samples are shown in Fig. 3b. The values of the hidden nodes indicate the intensity of the group effects on the pathways, which are connected to the hidden nodes. For instance, the first 16 hidden nodes in Fig. 3b show distinguishable intensities on LTS and non-LTS patients. The LTS patients present significant intensities of the group effects of the 16 pathways while non-LTS patients show significant lower values.

The weights between the pathway nodes and the hidden nodes are exhibited in Fig. 3c, and the top-10 ranked pathways among them are zoomed in Fig. 4a. It appears that a small number of pathways mainly contribute to the hidden nodes simultaneously, which implies that the cohort of the pathways may be candidates of prognostic biomarkers in long-term survival of GBM. The top-10 ranked pathways include signaling by GPCR, GPCR downstream signaling, innate immune system, adaptive immune system, metabolism of carbohydrates, transmembrane transport of small molecules, developmental biology, metabolism of proteins, class A/1 (rhodopsin-like receptors), and axon guidance. Most of the pathways are referred to as significant pathways in GBM in biological literature. The pathways and the references are listed in Table 3. Since the top-10 ranked pathways are all large (gene numbers > 200), we further explored small pathways as well. Class B/2 (Secretin family receptors) pathway which includes 88 genes is ranked 14th. One of the subgroups in Class B/2 family is categorized as brain-specific angiogenesis inhibitors that are growth suppressors of glioblastoma cells [29]. Hence, Class B/2 pathway may play an important role in inhibition of GBM.

Table 2 The Wilcoxon signed-rank tests for comparing PASNet with the Benchmark Classifiers

	W Statistic	*P*-value
PASNet vs. Dropout NN	146.5	2.13e-06
PASNet vs. RBF-SVM	137.0	1.35e-06
PASNet vs. Random LASSO	45.0	1.06e-08
PASNet vs. Logistic LASSO	43.0	9.52e-09

The genes of the pathways are illustrated by the weight values in Fig. 4b. Since the connections between the gene layer and the pathway layer are given by pathway databases, e.g., Reactome, they are very sparse. It also shows that multiple pathways share genes in common. The genes, which are most frequently shown in the ten pathways, include CDC42, PRKCQ, RAC1, AKT1, AKT2, AKT3, C3, CREB1, GRB2, HRAS, KRAS, NRAS, PRKACA, PRKACB, PRKACG, RAF1, and YWHAB, where CDC42, PRKCQ, and RAC1 are shown in six pathways and others are in five pathways. Among them, several genes have been reported as biomarkers in GBM. For instance, AKT1, AKT2, and AKT3, belonging to the five pathways of signaling by GPCR, GPCR downstream signaling, innate immune system, adaptive immune system, and developmental biology, are three isoforms of AKT in PI3K/AKT pathway, which is an important drug target in many cancers including GBM [30]. In particular, AKT2 is a well-known proto-oncogene that promotes the growth of tumors and reduces the survival of patients in GBM [31, 32].

Finally, we demonstrate a hierarchical representation of genes and pathways in PASNet. In Fig. 5a, PASNet is partially visualized, where positive and negative weights are colored in red and blue respectively. The pathways are represented by the corresponding genes in the pathway layer, and then the nonlinear effects of the pathways are described in the hidden layer. The hierarchical representations can be captured in the output layer, which produces a posterior probability for prognosis prediction. Although we considered a single hidden layer to simplify the model with HDLSS data in this study, multiple hidden layers may be able to capture the biological processes and their effects more accurately if a sufficient number of samples are available. Figure 5b–c illustrate distinctive representations of LTS and non-LTS samples in PASNet. The color of nodes in the figures shows the values computed with LTS/non-LTS samples in average. Note that node values between the pathway layer and the output layer are between zero and one. The node with a high value may be

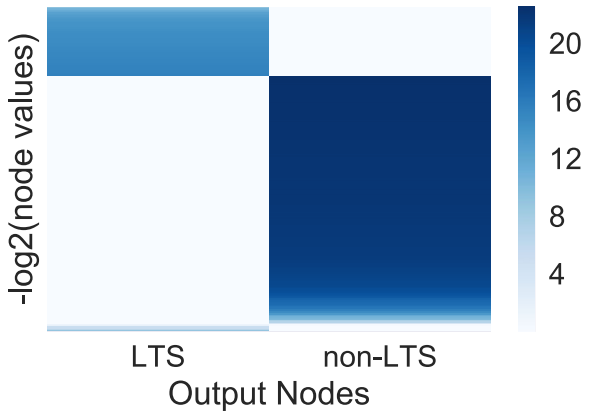

Fig. 2 Graphical representation of the output node values over the samples by PASNet. LTS samples obtain higher node values in LTS node than non-LTS samples. Similarly, non-LTS samples obtain higher node values in non-LTS node than LTS samples

a potential prognostic biomarker in the group. Figure 5b shows that pathways including aquaporin-mediated transport, signaling by BMP, and cytokine signaling in immune system are activated with LTS samples. The second node in the hidden layer is triggered by the active pathways, and the hidden node activates the LTS node in the output layer. On the other hand, Fig. 5c shows that additional pathways of signaling by GPCR and innate immune system are also activated for non-LTS samples. The other two hidden nodes take the active pathways into account, and they activate the non-LTS node in the output layer. Hence, the two pathways of signaling by GPCR and innate immune system may be potential prognostic biomarkers for predicting LTS/non-LTS. Pathway of signaling by GPCR has been investigated as a potential therapeutic target to inhibit the progression of glioblastomas. [33]. Activating the innate immune system, i.e. immunotherapy, is a promising strategy for the treatment of GBM [34].

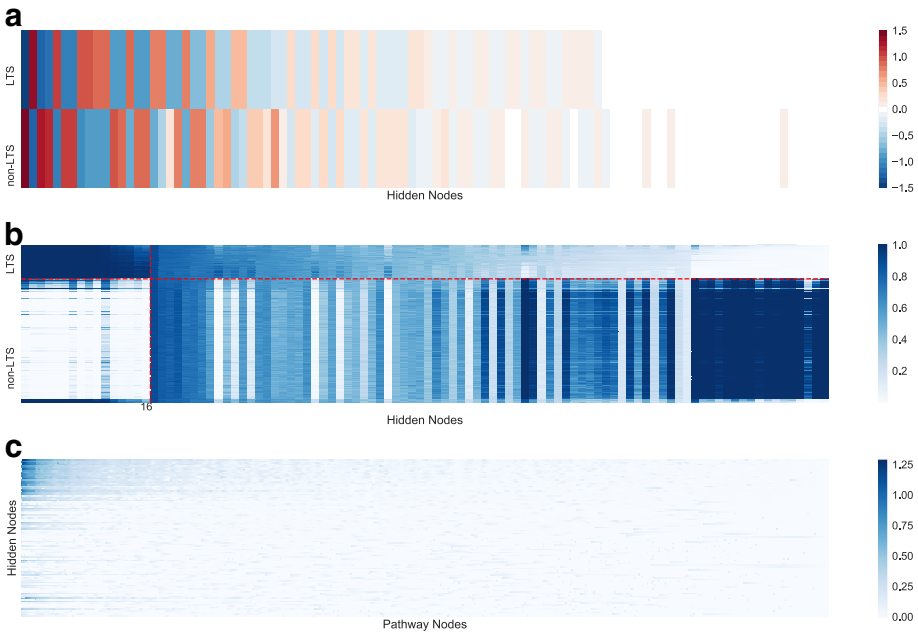

Fig. 3 Graphical representation among the output layer, hidden layer, and pathway layer in PASNet. (**a**) The weights between the hidden layer and the output layer. Hidden nodes are sorted in a descending order. (**b**) The node values in the hidden layer. The horizontal dotted lines indicates LTS/non-LTS samples. The vertical dotted lines indicates LTS/non-LTS samples are significantly distinguished by top 16 pathways. (**c**) The absolute weights between the pathway layer and the hidden layer

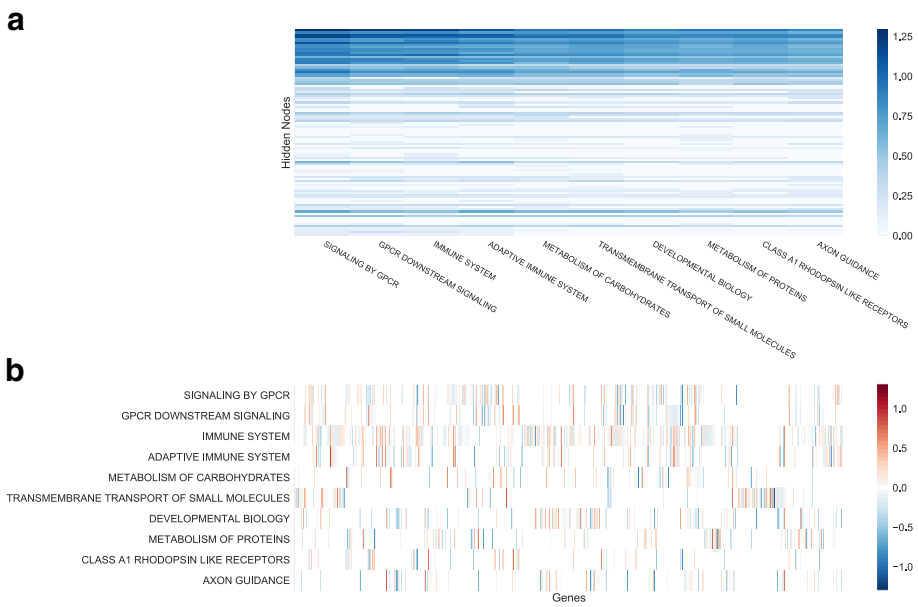

Fig. 4 Graphical representation of the 10 top-ranked pathways by PASNet (**a**) The absolute weights between the 10 top-ranked pathway nodes and the hidden layer. It is a zoom-in view of Fig. 3c. (**b**) Weights between the gene layer and the 10 top-ranked pathway nodes. The connections are determined by Reactome database

Vascular endothelial growth factor (VEGF), a modulator of the innate immune system, is reported crucial for the tumor progression [35]. Moreover, aquaporin-mediated transport, signaling by BMP, and cytokine signaling in immune system may play an important role in GBM, since they are shown in common as active in both LTS and non-LTS. Note that the activation/inactivation of a node in PASNet does not directly represent biological activation in the system, whereas it indicates different states of the biological components in the groups.

Conclusions

In this paper, we proposed pathway-associated sparse deep neural network for prognosis predictions (long-term survivals in GBM in this study). PASNet builds a network model by leveraging prior biological knowledge of pathway databases and by taking hierarchical nonlinear relationships of biological processes into account. To improve the model interpretability, PASNet introduces sparse coding. Moreover, we developed a training strategy to avoid the overfitting problem with HDLSS data and the imbalanced problem.

Table 3 Top-10 ranked pathways for survival prediction in GBM by PASNet

Pathway name	Pathway size	Reference	Top-5 ranked genes[a]
Signaling by GPCR	920	[33]	SHH, PTGFR, GNG5, CHRM5, LHB
GPCR downstream signaling	805	[50]	PTGFR, OR7C2, GNG5, OR10H3, MLNR
Innate immune system	933	[35]	CD79B, INPPL1, SRC, NUP85, DNM2
Adaptive immune system	539	[51]	CD79B, ASB6, PTEN, NCF4, FBXO2
Metabolism of carbohydrates	247	-	HS3ST3B1, NUP85, PFKFB3, LUM, SLC2A4
Transmembrane transport of small molecules	413	[52]	SLC9A7, ABCA7, GNG5, AQP8, HK3
Developmental biology	396	-	NRP2, FES, WNT10B, MYOD1, SLC2A4
Metabolism of proteins	518	-	EIF3G, CCT2, TIMM22, RPL3L, GMPPA
Class A/1 (rhodopsin-like receptors)	305	[53]	PTGFR, OPRD1, CHRM5, NPFF, NTSR2
Axon guidance	251	[54]	NRP2, NRTN, AGRN, FES, RPS6KA4

[a] The genes were ranked by absolute weights in the pathways

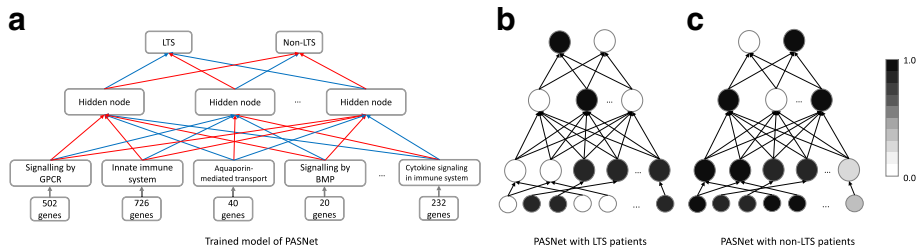

Fig. 5 Hierarchical representation of pathways in PASNet. (**a**) PASNet is partially visualized showing the five pathways. Distinct neural network activations between LTS (**b**) and non-LTS (**c**) are shown via PASNet. The nodes of the neural network of (b) and (c) correspond to (a). For instance, the nodes in the pathway layer of (b) and (c) represent signaling by GPCR, innate immune system, aquaporin-mediated transport, signaling by BMP, and Cytokine signaling in immune system. The pathways of signaling by GPCR and innate immune system are inactive with LTS patients, whereas the both pathways are active with non-LTS patients

To investigate the performance of PASNet, we used gene expression data of GBM patients in TCGA. PASNet was assessed by comparing the predictive performance with support vector machine, random LASSO, LASSO logistic Regression, and neural network with dropout that have been widely used for prognosis prediction. PASNet outperformed them with respect to both AUC and F1-score in the multiple stratified 5-fold cross-validation experiments. Furthermore, we discussed how PASNet can describe the biological system of GBM.

PASNet is the first deep neural network-based model that represents hierarchical representations of genes and pathways and their nonlinear effects, to the best of our knowledge. Additionally, PASNet would be promising due to its flexible model representation and interpretability, embodying the strengths of deep learning.

Methods

The architecture of PASNet

PASNet incorporates biological pathways and the concept of sparse modeling based on Deep Neural Network (DNN). The neural network architecture of PASNet consists of a gene layer (an input layer), a pathway layer that represents the biological pathways linked with input genes, a hidden layer that represents hierarchical relationships among biological pathways, and an output layer that corresponds with clinical outcomes, e.g. a binary class that has long-term survival and short-term survival, stages of cancer (see Fig. 6).

In PASNet, sparse coding is considered on the connections between layers for model interpretability. Sparse coding provides a solution to capture significant components of a biological mechanism in the

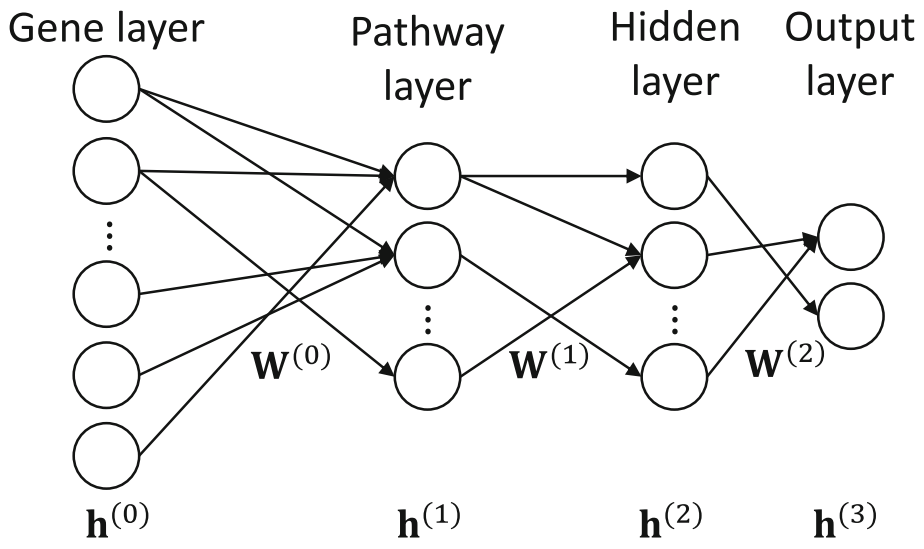

Fig. 6 Architecture of PASNet. The structure of PASNet is constructed by a gene layer (an input layer), a pathway layer that represents the biological pathways linked with input genes, a hidden layer that represents hierarchical relationships among biological pathways, and an output layer that corresponds with clinical outcomes, e.g. a binary class that has long-term survival and short-term survival, stages of cancer

model, since biological processes may involve only a few biological components. On the other hand, conventional fully-connected networks lack to represent biological mechanisms.

Gene layer

The gene layer (as an input layer) corresponds to gene expression data. A patient sample of m gene expressions is formed as a column vector, which is denoted by $\mathbf{x} = \{x_1, x_2, ..., x_m\}$. Each input node represents one gene feature.

Pathway layer

The pathway layer represents biological pathways, where each node indicates an individual pathway. The connections between the gene layer and the pathway layer are established by well-known pathway databases (e.g., Reactome and KEGG). Pathway databases contain associations between pathways and genes; each of which provides a set of gene components. Therefore, the pathway layer makes it possible to interpret the model as a pathway-based analysis.

To begin with initializing the connections between the gene layer and the pathway layer, we consider a binary biadjacency matrix ($\mathbf{A}$) from biological pathway databases. The biadjacency matrix can be defined as $\mathbf{A} \in \mathbb{B}^{n \times m}$, where n is number of pathways and m is number of genes. Then, an element of $\mathbf{A}$, i.e., a_{ij}, is set to one if gene j belongs to pathway i; otherwise, zero. Sparse coding is applied based on the matrix $\mathbf{A}$ to represent the relationships between genes and pathways in the model.

Hidden layer

Biological components may cooperate with others instead of functioning alone. A biological system involves multiple pathways which have interactions together, whereas a node in the pathway layer indicates a biological pathway. The associative interactions between pathways can be represented in the hidden layer. In PASNet, the hidden layer represents biological nonlinear associations between the pathways to outputs.

Sparse coding between the pathway and the hidden layers enables one to interpret these relationships. Although we consider only a single hidden layer in this study for simplicity's sake, multiple hidden layers can be used for deeper hierarchical representations of pathways. For example, if there are two hidden layers, the second hidden layer will represent deeper hierarchical associations of the nodes of the first hidden layer, which are association effects of pathways.

Output layer

The output layer shows clinical outcomes for which nodes compute the posterior probabilities. In this layer, sparse coding allows to distinguish hierarchical groups of pathways (which are detected from hidden layers) to predict clinical outcomes. In PASNet, more than two clinical outcomes can be easily represented with multiple nodes in the output layer.

Consequently, PASNet can dissect distinguishable biological processes of hierarchical nonlinear relationships and associations of genes and pathways to predict clinical outcomes. Furthermore, this *generative* model-based approach would be often useful to predict prognosis accurately with complex data of HDLSS. When data is highly complex and only small sample sizes are available, model optimization may be easily biased to the training data rather than providing a general solution. On the other hand, the integration of the biological structures and prior knowledge to the model would produce a robust solution.

Overall description of PASNet training

The main challenge in training PASNet is to reduce both risk of overfitting and computational complexity of training on HDLSS data. The related works that have handled the HDLSS data problem are discussed in "Related works in deep learning" section. To unravel the problems, PASNet optimizes a small sub-network, which involves feasible nodes and parameters to train instead of the whole network and then makes the sub-network sparse. Figure 7 illustrates the overall training flow of PASNet.

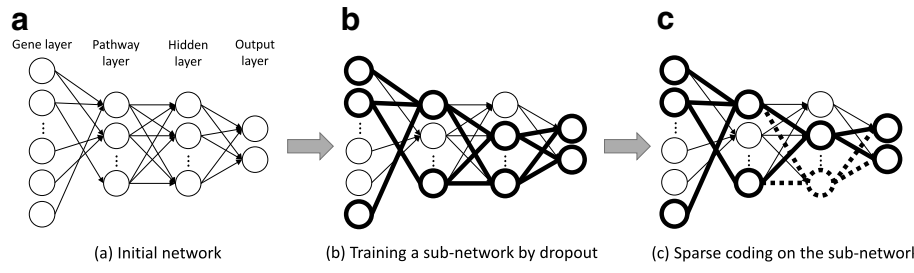

Fig. 7 Training of PASNet. (**a**) Weights and biases are randomly initialized. Connections between the gene layer and the pathway layer are determined by biological pathway databases, and the remaining layers are considered as fully-connected in this step. (**b**) A sub-network is randomly selected using a dropout technique and trained. (**c**) Sparse coding optimizes the sparsity of connections in the sub-network

First, we initialize the connections between the gene layer and the pathway layer with prior biological knowledge of pathways (see Fig. 7a). Active/inactive connections are determined by the biadjacency matrix, **A**. The weights of active connections and biases are randomly initialized from standard normal distribution, while the weights of inactive connections are set to zero. The sparsity of the connections between the gene layer and the pathway layer is invariant over the entire training. The remaining layers are fully interconnected as the initial.

In the training phase, we repeat training sub-networks and applying sparse coding on the sub-networks until convergence (Fig. 7b–c). A sub-network is selected by a dropout technique, where neurons are randomly dropped in the intermediate layers. In Fig. 7b, a small sub-network is shown with bold solid circles and lines. Then, the small sub-network is trained by feed-forward and backpropagation. Note that only weights and biases of the sub-network are trained. Upon the completion of the sub-network's training, sparse coding is applied to the sub-network by trimming the connections that do not contribute or worsen to minimize the loss. In Fig. 7c, the dropped connections and nodes are marked as bold, dashed lines. The details of the training are elucidated in the following sections.

Sparse coding

Once a small sub-network is completed to train with the HDLSS data, the sub-network is imposed to be sparse for the model interpretation. The sparsity of the sub-network is determined by the mask matrix **M** on each layer as:

$$\mathbf{h}^{(\ell+1)} = a\left(\left(\mathbf{W}^{(\ell)} \star \mathbf{M}^{(\ell)}\right)\mathbf{h}^{(\ell)} + \mathbf{b}^{(\ell)}\right), \tag{1}$$

where $\star$ denotes element-wise multiplication, and $a(\cdot)$ is an activation function. $\mathbf{h}^{(\ell)}$ denotes an output vector on the ℓ-th layer, and $\mathbf{W}^{(\ell)}$ and $\mathbf{b}^{(\ell)}$ are a weight matrix and a bias vector, respectively. An element value of **M** is either one or zero, which determines whether the associated weights are dropped in the current epoch.

The mask matrix **M** is generated with respect to a sparsity level (S) that indicates the proportion of weights to be dropped in a single layer. S is a value between 0 to 100, where zero creates a fully-connected layer while 100 causes no connection. The optimal S^* is approximated on each layer individually in the sub-network, while most related methods consider a single hyper-parameter for the sparsity of all layers [36, 37]. The individual setting of the sparsity on each layer shows different levels of biological associations on the genes and pathways.

We obtain the optimal sparsity level S^* that minimizes the cost score. For efficient computation, the cost scores are computed with a small number of finite sparsity levels. Then, the optimal sparsity level is estimated by applying a cubic-spline interpolation to the cost scores with the assumption that the cost function, with respect to the sparsity level, is continuous.

In particular, an element of **M** is set to one if the absolute value of the corresponding weight is greater than threshold Q; otherwise, the element is zero, where Q is an S-th percentile of absolute values of **W**. Note that the mask between the gene layer and the pathway layer, i.e. $\mathbf{M}^{(0)}$, is determined by the biadjacency matrix **A** of biological pathways. Thus, the mask matrices are formulated as

$$\mathbf{M}^{(\ell)} = \begin{cases} \mathbb{1}\left(|\mathbf{W}^{(\ell)}| \geq Q^{(\ell)}\right), & \text{if } \ell \neq 0 \\ \mathbf{A}, & \text{if } \ell = 0 \end{cases} \tag{2}$$

where $Q^{(\ell)}$ is the S-th percentile of $|\mathbf{W}^{(\ell)}|$ if $\ell \neq 0$.

Cost-sensitive learning for imbalanced data

We refine the cost function and the backpropagation for cost-sensitive learning, since imbalanced data causes bias of the predictions towards the majority class. We adapt the Mean False Error (MFE) method [38], which penalizes the errors of the majority class.

Let K be the number of clinical outcomes. The normalized cost is computed separately for each class by:

$$\mathcal{L} = \sum_{k=1}^{K} \mathcal{C}_k + \frac{1}{2}\lambda\|\mathbf{W}\|_2, \tag{3}$$

$$\mathcal{C}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} c\left(\mathbf{y}_i, \tilde{\mathbf{y}}_i\right), \tag{4}$$

where $\mathcal{C}_k$ denotes mean error on the class k, and n_k is the number of samples in the class k. $\mathbf{y}_i$ is a vectorized ground truth class label of the i-th sample, and $\tilde{\mathbf{y}}_i$ is its vectorized prediction. $c(\cdot)$ denotes a cost function (e.g., cross-entropy loss), and $\mathcal{L}$ is the total cost. $\|\mathbf{W}\|_2$ denotes a L^2-norm of **W**, and $\lambda > 0$ is a regularization hyperparameter.

In the backpropagation phrase, the gradient is also computed separately for each class. Hence, the weights and biases on the ℓ-th layer are updated by:

$$\mathbf{W}^{(\ell)} \leftarrow (1-\eta\lambda)\mathbf{W}^{(\ell)} - \eta\sum_{k=1}^{K}\frac{\partial\mathcal{C}_k}{\partial\mathbf{W}^{(\ell)}}, \tag{5}$$

$$\mathbf{b}^{(\ell)} \leftarrow \mathbf{b}^{(\ell)} - \eta\sum_{k=1}^{K}\frac{\partial\mathcal{C}_k}{\partial\mathbf{b}^{(\ell)}}, \tag{6}$$

where η is a learning rate. The algorithm of PASNet is briefly described in Algorithm 1.

Algorithm 1 Training of PASNet

Initialize weights $\mathbf{W}^{(\ell)}$ and biases $\mathbf{b}^{(\ell)}$
$\mathbf{W}^{(0)} \leftarrow \mathbf{W}^{(0)} \star \mathbf{M}^{(0)}$
repeat
  Select a small sub-network via dropout
  Train the sub-network by Eqs. (5) and (6)
  Sparse coding with the optimal $\mathbf{M}^{(\ell)}$ by Eq. (2)
  $\mathbf{W}^{(\ell)} \leftarrow \mathbf{W}^{(\ell)} \star \mathbf{M}^{(\ell)}$
until convergence

Related works in deep learning

In recent years, deep learning has been spotlighted as the most active research field in various machine learning communities, such as image analysis, speech recognition, and natural language processing as its promising potential is being actively discussed in bioinformatics and biomedicine [39]. Most deep learning-based approaches have been developed for classification and association studies in bioinformatics. For instance, D-GEX infers the expression of target genes from landmark genes, capturing the nonlinear relationships by combining gene expression, DNA methylation, and miRNA expression data [40]. A convolutional neural network (CNN) was adapted to predict DNA-protein binding sites with Chromatin Immunoprecipitation sequencing (ChIP-seq) data [41]. Additionally, CNN-based DeepBind was proposed to predict whether a specific DNA/RNA binding protein will bind to a specific DNA sequence [42]. The functionality of non-coding variants was predicted by DeepSEA by employing a CNN model [43].

Although only a small subset of deep learning research has been reported in bioinformatics due to the difficulty of structure definition and interpretation, the future of deep learning in biology and medicine is promising [44]. First, since a neural network is inspired by the neurons in the human brain, a neuron network architecture is applicable to modeling a mechanism for a complex biological system. Specifically, deep learning approaches take advantage of flexible representation of hierarchical structures from inputs to outputs. The representation of nonlinear effects of neurons in multiple layers in neural networks may be able to model hierarchical biological signals. DCell constructs a multilayer neural network based on extensive prior biological knowledge to simulate the growth of a eukaryotic cell [45]. However, DCell's network architecture is entirely based on well-known prior biological knowledge, so the model was applied to relatively simple biological system of yeast. Moreover, deep learning captures nonlinear effects of variables with high-level feature representation, which allows deep learning to outperform other state-of-the-art methods.

However, training deep neural networks with HDLSS data poses a computational problem. A large number of parameters are involved in deep neural networks, and it often makes the training infeasible or causes a model overfit on HDLSS data. Particularly, backpropagation gradients in neural networks are of high variance on HDLSS data, which consequently causes the model overfit [46]. In order to tackle the HDLSS problem, the leave-one-out approach was used to avoid the overfitting problem in backpropagation [47]. Regarding backpropagation, the risk of overfitting was examined with validation data by the leave-one-out approach and terminates the training early when overfitting occurs. For an alternative solution, an attempt to reduce the dimensionality of the input space to a feasible size has been made [48]. Dimension reduction techniques, such as subsampled randomized Hadamard transform (SRHT) and Count Sketch-base construction, were utilized to reduce the dimensional size of the input data. Then, the projected data into the lower space were introduced to a neural network for training.

For HDLSS data, feature selection is one of the conventional approaches. Deep Feature Selection (DFS) was developed to select a discriminative feature subset in a deep learning model [49]. Although DFS is not the optimal solution to low-sample size data, DFS shows that deep learning can detect informative and discriminative features of nonlinearity effects through multiple layers with high-dimensional data. Then, Deep Neural Pursuit (DNP) improved the solution of the feature selection in deep learning, taking the HDLSS data problem into account [46]. DNP iteratively augments features in the input layer by performing multiple dropouts. The multiple dropouts grant the ability to train a small-sized sub-network at a time and to compute gradients with low variance for alleviating the overfitting problem.

Abbreviations

Adam: Adaptive moment estimation; AUC: Area under the curve; ChIP-seq: Chromatin Immunoprecipitation sequencing; CNN: Convolutional neural network; DFS: Deep feature selection; DNN: Deep neural network; DNP: Deep neural pursuit; Dropout NN: Neural network with dropout; GBM: Glioblastoma multiforme; HDLSS: High-dimension, low-sample size; KPCA: Kernel principle component analysis; LLR: LASSO logistic regression; MFE: Mean false error; MSigDB: Molecular signatures database; PASNet:Pathway-associated sparse deep neural network; RBF: Radial basis function; RBM: Restricted Boltzmann machine; ROC: Receiver operating characteristic; SRHT: Subsampled randomized hadamard transform; SVM: Support vector machine; TCGA: The cancer genome atlas

Acknowledgements

We would like to thank Dr. Jung Hun Oh for his help and advice in this study.

Funding

Not applicable.

Availability of data and materials

The datasets are publicly available and accessible at http://cancergenome.nih.gov.

Authors' contributions
MK designed and supervised the research project. JH developed and implemented the algorithms, and performed the data analyses. JH and MK wrote the manuscript. YK helped the implementation and performed the experiments. TK discussed and verified the biological interpretation of PASNet. All authors have read and approved the final version of the 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.

Author details
[1]Kennesaw State University, Kennesaw, USA. [2]Kennesaw State University, Marietta, USA. [3]University of Texas Southwestern Medical Center, Dallas, USA. [4]Department of Life Sciences, Pohang Institute of Science and Technology (POSTECH), Dallas, USA.

Received: 29 June 2018 Accepted: 16 November 2018
Published online: 17 December 2018

References

1. Lu J, Cowperthwaite MC, Burnett MG, Shpak M. Molecular Predictors of Long-Term Survival in Glioblastoma Multiforme Patients. PloS ONE. 2016;11(4):0154313. https://doi.org/10.1371/journal.pone.0154313.
2. Onaitis MW, et al. Prediction of Long-Term Survival After Lung Cancer Surgery for Elderly Patients in The Society of Thoracic Surgeons General Thoracic Surgery Database. Ann Thorac Surg. 2018;105(1):309–16. https://doi.org/10.1016/j.athoracsur.2017.06.071.
3. Cao Y, et al. Prediction of long-term survival rates in patients undergoing curative resection for solitary hepatocellular carcinoma. Oncol Letters. 2018;15(2):2574–82. https://doi.org/10.3892/ol.2017.7612.
4. Jin L, et al. Pathway-based Analysis Tools for Complex Diseases: A Review. Genomics Proteomics Bioinforma. 2014;12(5):210–20. https://doi.org/10.1016/j.gpb.2014.10.002.
5. Kim S, Kon M, DeLisi C. Pathway-based classification of cancer subtypes. Biol Direct. 2012;7:21. https://doi.org/10.1186/1745-6150-7-21.
6. Cirillo E, Parnell LD, Evelo CT. A review of pathway-based analysis tools that visualize genetic variants. Front Genet. 2017;8(174):174. https://doi.org/10.3389/fgene.2017.00174.
7. Drier Y, Sheffer M, Domany E. Pathway-based personalized analysis of cancer. Proc Natl Acad Sci U S A. 2013;110(16):6388–93. https://doi.org/10.1073/pnas.1219651110.
8. Mallavarapu T, Kim Y, Oh JH, Kang M. R-pathcluster: Identifying cancer subtype of glioblastoma multiforme using pathway-based restricted boltzmann machine. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2017. p. 1183–8. https://doi.org/10.1109/BIBM.2017.8217825.
9. Huang S, et al. Novel personalized pathway-based metabolomics models reveal key metabolic pathways for breast cancer diagnosis. Genome Med. 2016;8(1):34. https://doi.org/10.1186/s13073-016-0289-9.
10. Li Y, Nan B, Zhu J. Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. Biometrics. 2015;71(2):354–63. https://doi.org/10.1111/biom.12292. http://arxiv.org/abs/15334406.
11. Raser JM, O'Shea EK. Noise in Gene Expression: Orgins, Consequences, and Control. Science. 2005;309(5743):2010–3. https://doi.org/10.1126/science.1105891. http://arxiv.org/abs/NIHMS150003.
12. Steyerberg EW, Eijkemans MJC, Habbema JDF. Application of Shrinkage Techniques in Logistic Regression Analysis: A Case Study. Statistica Neerlandica. 2001;55(1):76–88. https://doi.org/10.1111/1467-9574.00157.
13. Wang S, Nan B, Rosset S, Zhu J. Random lasso. Ann Appl Stat. 2011;5(1):468–85. https://doi.org/10.1214/10-AOAS377. http://arxiv.org/abs/1104.3398.
14. Musoro JZ, Zwinderman AH, Puhan MA, Ter Riet G, Geskus RB. Validation of prediction models based on lasso regression with multiply imputed data. BMC Med Res Methodol. 2014;14(1). https://doi.org/10.1186/1471-2288-14-116.
15. Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. Biometrics. 2007;63(4):1079–88. https://doi.org/10.1111/j.1541-0420.2007.00799.x.
16. Liu D, Ghosh D, Lin X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics. 2008;9. https://doi.org/10.1186/1471-2105-9-292.
17. Bach FR, Lanckriet GRG, Jordan MI. Multiple kernel learning, conic duality, and the SMO algorithm. In: Twenty-first International Conference on Machine Learning - ICML '04. 2004. p. 6. https://doi.org/10.1145/1015330.1015424. http://portal.acm.org/citation.cfm?doid=1015330.1015424.
18. Sinnott JA, Cai T. Pathway aggregation for survival prediction via multiple kernel learning. Stat Med. 2018;0(0). https://doi.org/10.1002/sim.7681. http://arxiv.org/abs/https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.7681.
19. Kumari S, et al. Bottom-up GGM algorithm for constructing multilayered hierarchical gene regulatory networks that govern biological pathways or processes. BMC Bioinformatics. 2016;17(1). https://doi.org/10.1186/s12859-016-0981-1.
20. Deng W, Zhang K, Busov V, Wei H. Recursive random forest algorithm for constructing multilayered hierarchical gene regulatory networks that govern biological pathways. PLoS ONE. 2017;12(2). https://doi.org/10.1371/journal.pone.0171532.
21. Pham LM, Carvalho L, Schaus S, Kolaczyk ED. Perturbation Detection Through Modeling of Gene Expression on a Latent Biological Pathway Network: A Bayesian Hierarchical Approach. J Am Stat Assoc. 2016;111(513):73–92. https://doi.org/10.1080/01621459.2015.1110523. http://arxiv.org/abs/1409.0503.
22. Kher S, Peng J, Wurtele ES, Dickerson J. In: Pérez-Sánchez H, editor. Hierarchical Biological Pathway Data Integration and Mining, Bioinformatics: IntechOpen; 2012. https://doi.org/10.5772/49974. Available from: https://www.intechopen.com/books/bioinformatics/hierarchical-biological-pathway-data-integration-and-mining.
23. Hanif F, Muzaffar K, Perveen K, Malhi SM, Simjee SU. Glioblastoma Multiforme: A Review of its Epidemiology and Pathogenesis through Clinical Presentation and Treatment. Asian Pac J Cancer Prev. 2017;18(1):3–9. https://doi.org/10.22034/APJCP.2017.18.1.3.
24. Davis ME. Glioblastoma: Overview of Disease and Treatment. Clin J Oncol Nurs. 2016;20(5):1–14. https://doi.org/10.1188/16.CJON.S1.2-8.
25. Walid MS. Prognostic factors for long-term survival after glioblastoma. Permanente J. 2008;12(4):45–8. https://doi.org/10.7812/TPP/08-027.
26. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov J, Tamayo P. The molecular signatures database hallmark gene set collection. Cell Syst. 2015;1(6):417–25. https://doi.org/10.1016/j.cels.2015.12.004.
27. Kingma DP, Ba J. Adam: A method for stochastic optimization. CoRR. 2014;abs/1412.6980. http://arxiv.org/abs/1412.6980.
28. Hsu C-W, Chang C-C, Lin C-J. A Practical Guide to Support Vector Classification. Available from: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf. Accessed 15 June 2008.
29. Harmar AJ. Family-B G-protein-coupled receptors. Genome Biol. 2001;2(12):3013–1301310. https://doi.org/10.1186/gb-2001-2-12-reviews3013.
30. Joy A, et al. The role of AKT isoforms in glioblastoma: AKT3 delays tumor progression. J Neuro-Oncol. 2016;130(1):43–52. https://doi.org/10.1007/s11060-016-2220-z.
31. Hu B, et al. Astrocyte elevated gene-1 interacts with Akt isoform 2 to control glioma growth, survival, and pathogenesis. Cancer Res. 2014;74(24):7321–32. https://doi.org/10.1158/0008-5472.CAN-13-2978.
32. Hinske LC, et al. Intronic mirna-641 controls its host gene's pathway pi3k/akt and this relationship is dysfunctional in glioblastoma multiforme. Biochem Biophys Res Commun. 2017;489(4):477–83. https://doi.org/10.1016/j.bbrc.2017.05.175.

33. Cherry AE, Stella N. G protein-coupled receptors as oncogenic signals in glioma: Emerging therapeutic avenues. Neuroscience. 2014;278(1): 222–36. https://doi.org/10.1016/j.neuroscience.2014.08.015.
34. Lim M, Xia Y, Bettegowda C, Weller M. Current state of immunotherapy for glioblastoma. Nat Rev Clin Oncol. 2018;15(7):422–42. https://doi.org/10.1038/s41571-018-0003-5.
35. Turkowski K, et al. VEGF as a modulator of the innate immune response in glioblastoma. GLIA. 2018;66(1):161–74. https://doi.org/10.1002/glia.23234.
36. Han S, et al. DSD: Dense-Sparse-Dense Training for Deep Neural Networks. Int Conf Learn Represent. 2017. http://arxiv.org/abs/1607.04381.
37. Wang B, Klabjan D. Regularization for Unsupervised Deep Neural Nets. CoRR. 2016;1:1–7. http://arxiv.org/abs/1608.04426.
38. Wang S, Liu W, Wu J, Cao L, Meng Q, Kennedy PJ. Training deep neural networks on imbalanced data sets. 2016 Int Jt Conf Neural Netw. 2016;4368–4374. https://doi.org/10.1109/IJCNN.2016.7727770.
39. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinforms. 2017;18(5):851–69. https://doi.org/10.1093/bib/bbw068. http://arxiv.org/abs/1603.06430.
40. 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. https://doi.org/10.1109/TCBB.2014.2377729.
41. Zeng H, Edwards MD, Liu G, Gifford DK. Convolutional neural network architectures for predicting DNA-protein binding. Bioinformatics. 2016;32(12):121–7. https://doi.org/10.1093/bioinformatics/btw255.
42. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33(8):831–8. https://doi.org/10.1038/nbt.3300.
43. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12(10):931–4. https://doi.org/10.1038/nmeth.3547. http://arxiv.org/abs/15334406.
44. Ching T, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15(141). https://doi.org/10.1098/rsif.2017.0387. http://arxiv.org/abs/http://rsif.royalsocietypublishing.org/content/15/141/20170387.full.pdf.
45. Ma J, Yu MK, Fong S, Ono K, Sage E, Demchak B, Sharan R, Ideker T. Using deep learning to model the hierarchical structure and function of a cell. Nat Methods. 2018;15(4):290–8. https://doi.org/10.1038/nmeth.4627.
46. Liu B, Wei Y, Zhang Y, Yang Q. Deep Neural Networks for High Dimension, Low Sample Size Data. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17; 2017. p. 2287–93. https://doi.org/10.24963/ijcai.2017/318.
47. Pasini A. Artificial neural networks for small dataset analysis. J Thorac Dis. 2015;7(5):953–60. https://doi.org/10.3978/j.issn.2072-1439.2015.04.61.
48. Wójcik PI, Kurdziel M. Training neural networks on high-dimensional data using random projection. Pattern Anal Applic. 2018. https://doi.org/10.1007/s10044-018-0697-0.
49. Li Y, Chen C-Y, Wasserman WW. Deep feature selection: Theory and application to identify enhancers and promoters. J Comput Biol. 2016;23(5):322–36. https://doi.org/10.1089/cmb.2015.0189. PMID: 26799292.
50. Zhang J, Feng H, Xu S, Feng P. Hijacking GPCRs by viral pathogens and tumor. 2016. https://doi.org/10.1016/j.bcp.2016.03.021.
51. Feng L, et al. Heterogeneity of tumor-infiltrating lymphocytes ascribed to local immune status rather than neoantigens by multi-omics analysis of glioblastoma multiforme. Sci Reports. 2017;1(7). https://doi.org/10.1038/s41598-017-05538-z.
52. Zhou C, et al. Analysis of the gene-protein interaction network in glioma. Genet Mol Res. 2015;14(4):14196–206. https://doi.org/10.4238/2015.November.13.3.
53. Choi HY, et al. G protein-coupled receptors in stem cell maintenance and somatic reprogramming to pluripotent or cancer stem cells. BMB Rep. 2015;48(2):68–80. https://doi.org/10.5483/BMBRep.2015.48.2.250.
54. Chédotal A, Kerjan G, Moreau-Fauvarque C. The brain within the tumor: New roles for axon guidance molecules in cancers. 2005. https://doi.org/10.1038/sj.cdd.4401707.