Trans-species learning of cellular signaling with bimodal deep belief networks

Lujia Chen, MS, Chunhui Cai, PHD, Vicky Chen, MS, Xinghua Lu, MD PHD, Department of Biomedical Informatics, University of Pittsburgh School of Medicine

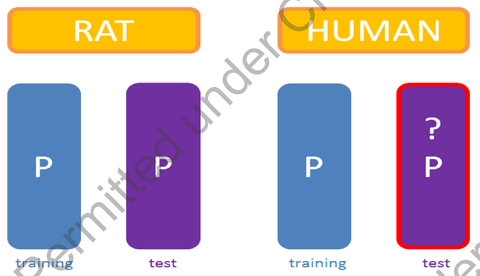
Introduction

Rodent studies have proved indispensable as models of human diseases and have undoubtedly helped to unravel molecular mechanisms. The biomedical field has generally worked under the assumption that biological processes in mice or rats are correspond to biological processes in human under analogous conditions. Yet few studies have addressed the limitation in which biological events observed in rodents can be translated to humans. This study provides some answers to this fundamental questions. As rodent models are still central tools in biomedical research, the answers will contribute a lot to the biomedical research community.

Are Rats and Humans Maybe Closer Than We Think?
https://www.sbvimprover.com/challenge-2/overview

Task: Inter-Species (RAT and HUMAN) Protein Phosphorylation Prediction

This study predicts human cells' proteomic responses to distinct stimuli based on the observed proteomic response to the same stimuli in rat cells. More specifically, **during the training phase**, participants were provided with data that measured the phosphorylation states of a common set of signaling proteins in primary cultured bronchial cells collected from rats and humans treated with distinct stimuli. **In the testing phase**, the proteomic data of rat cells treated with unknown stimuli were provided, and the task is to predict the proteomic responses of human cells treated with the same stimuli (https://www.sbvimprover.com/challenge-2/overview).



Methods and Results

Restricted Boltzmann Machine (RBM) and Deep Belief Network (DBN)

In this study, we investigated using the DBN model to represent the common encoding system of the signal transduction systems of human and rat bronchial cells. A DBN contains one visible layer and multiple hidden layers (Fig A). DBN is treated as a series of restricted Boltzmann machines (RBM) (Fig B) stacked on top of each other.

RBMs are probabilistic generative models that are able to automatically extract features of their input data using a completely unsupervised learning algorithm. RBMs consist of a layer of hidden and a layer of visible neurons with connection between hidden and visible neurons represented by an array of weights.

The RBM model defines the joint distribution of hidden and visible variables using a Boltzmann distribution as follows:

$$Pr(\mathbf{v},\mathbf{h};\theta) = \frac{1}{Z(\theta)} exp(-E(\mathbf{v},\mathbf{h};\theta))$$

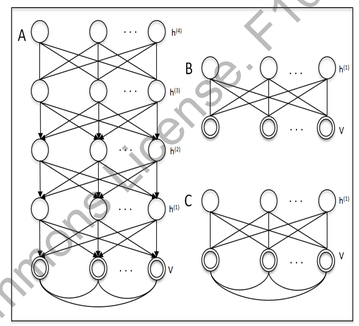
$$\mathbf{E}(\mathbf{v},\mathbf{h};\theta) = -\mathbf{a}^{\mathsf{T}}\mathbf{v} - \mathbf{b}^{\mathsf{T}}\mathbf{h} - \mathbf{v}^{\mathsf{T}}\mathbf{w}\mathbf{h} = -\sum_{i=1}^{D} a_i v_i - \sum_{j=1}^{F} b_j h_j - \sum_{i=1}^{D}\sum_{j=1}^{F} v_i h_j w_{ij}$$

The marginal distribution of visible variables is:

$$Pr(\mathbf{v};\theta) = \sum_{\mathbf{h}} Pr(\mathbf{v},\mathbf{h};\theta) = \frac{1}{Z(\theta)} \sum_{\mathbf{h}} exp(-E(\mathbf{v},\mathbf{h};\theta))$$

$$Z(\theta) = \sum_{\mathbf{v},\mathbf{h}} exp(-E(\mathbf{v},\mathbf{h};\theta))$$

To train an RBM, samples from a training set are used as input to the RBM through the visible neurons, and then the network alternatively samples back and forth between the visible and hidden neurons based on the following equations.



$$Pr(h_j = 1|\mathbf{v}) = \sigma(b_j + \sum_{i=1}^{n} W_{ij} v_i)$$

$$Pr(v_i = 1|\mathbf{h}) = \sigma(a_i + \sum_{j=1}^{m} W_{ij} p_j)$$

The update of parameters is given by:

$$\Delta W_{ij} = \epsilon(< v_i h_j >_{data} - < v_i h_j >_{model}) = \epsilon(< v_i h_j >_{Pr(h|v;w)} - < v_i h_j >_n)$$

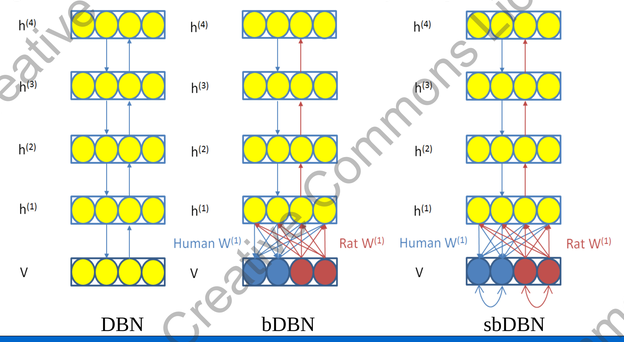
$$\Delta a_i = \epsilon(< v_i >_{data} - < v_i >_n)$$

$$\Delta b_j = \epsilon(< h_j >_{data} - < h_j >_n)$$

To train an DBN, 1) stack-wise learning of RBMs weight parameters and instantiation of hidden variables in the top layer 2) learning the generative weights across all layers which is achieved by a backpropagation algorithm as in training standard neural networks

Bimodal Deep Belief Network (bDBN):

We designed the bDBN to capture the joint distribution of rat and human proteomic data. When training the bDBN, we combined a rat training case with a human training case treated with a common stimulus into a joint input vector for the bDBN.



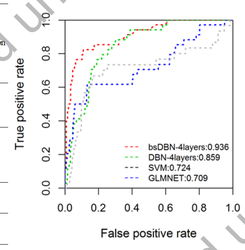
When using a trained bDBN to predict human cell response to a specific stimulus based on the observed rat cell response to the same stimulus, we only used the rat data to update the top hidden layer. Then, the bDBN propagated the information derived from rat data downwards using generative weights as in a feed forward neural network to predict the human data.

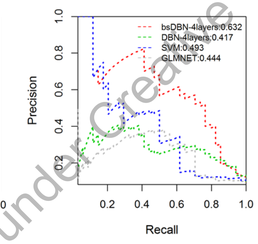
Semi-restricted Bimodal Deep Belief Network (sbDBN):

Since signaling proteins in a phosphorylation cascade have regulatory relationships among themselves, sbDBN inserts the edges between proteins from a common species to bDBN.

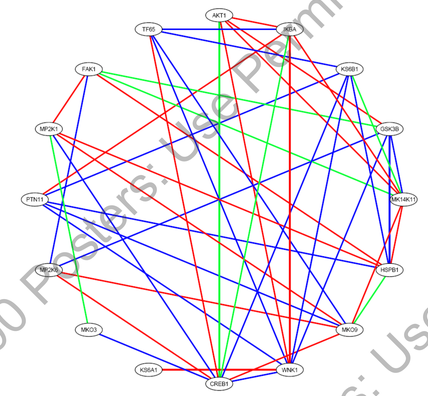
Result 1: Comparison among different models

	AUPRC	AUROC	Jaccard. Similarity	Matthews. Correlation. Coefficient	Speaman. Correlation	Pearson. Correlation
4-layered bDBN	0.400	0.861	0.750	0.373	0.323	0.235
4-layered sbDBN	**0.639**	**0.936**	**0.531**	**0.616**	**0.391**	**0.460**
SVM	0.471	0.721	0.692	0.411	0.231	0.392
GLMNET	0.442	0.686	0.717	0.374	0.194	0.282





Result 2: The correlation among the signaling proteins learnt from sbDBN



Conclusion

Our results indicate that, by learning to represent a common encoding system for both rat and human cells, the deep learning models outperform contemporary state-of-the-art classifiers in this trans-species learning task.

Acknowledgement

This study was partially supported by Grant Number R01LM 010144 from the National Library of Medicine, NIH, to XL.