Boosted Categorical Restricted Boltzmann Machine for Computational Prediction of Splice Junctions

Taehoon Lee Sungroh Yoon

Advanced Computing Laboratory
Electrical and Computer Engineering
Seoul National University

ICML Lille
International Conference on Machine Learning

Outline

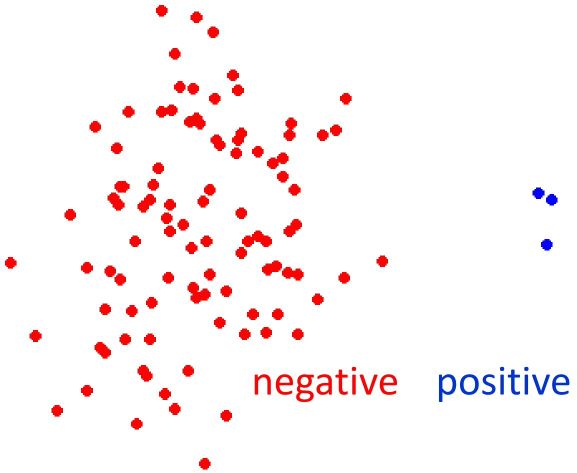
- Motivation
- Preliminary
- Boosted contrastive divergence
- Categorical restricted Boltzmann machine
- Experiment results
- Conclusion

Motivation

- Deep Neural Networks (DNN) show human level performance on many recognition tasks.
- We focus on **class-imbalanced prediction**.
    - Insufficient samples to represent the true distribution of a class.

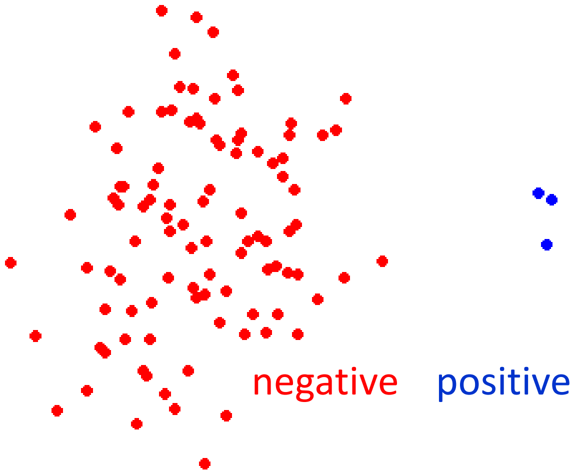
Motivation

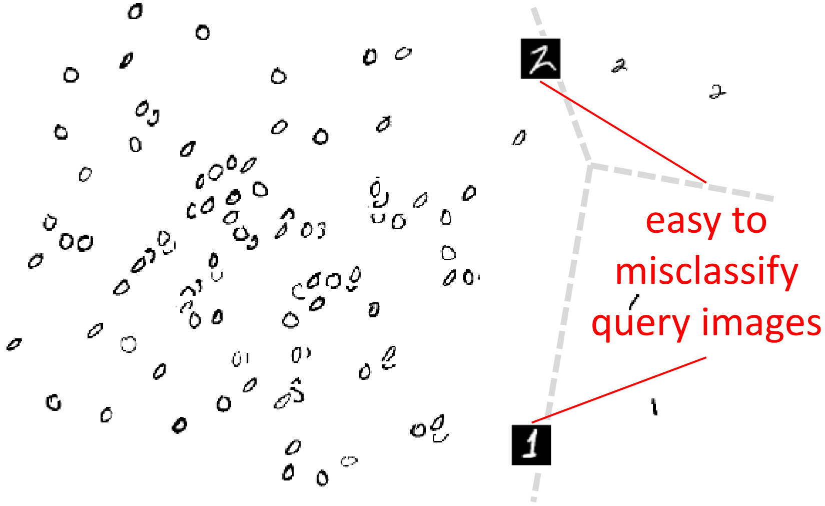
- Deep Neural Networks (DNN) show human level performance on many recognition tasks.
- We focus on **class-imbalanced prediction**.
  - Insufficient samples to represent the true distribution of a class.



Motivation

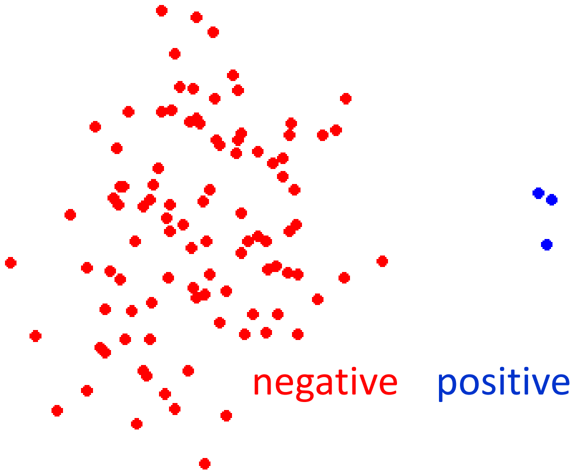
- Deep Neural Networks (DNN) show human level performance on many recognition tasks.
- We focus on **class-imbalanced prediction**.
  - Insufficient samples to represent the true distribution of a class.

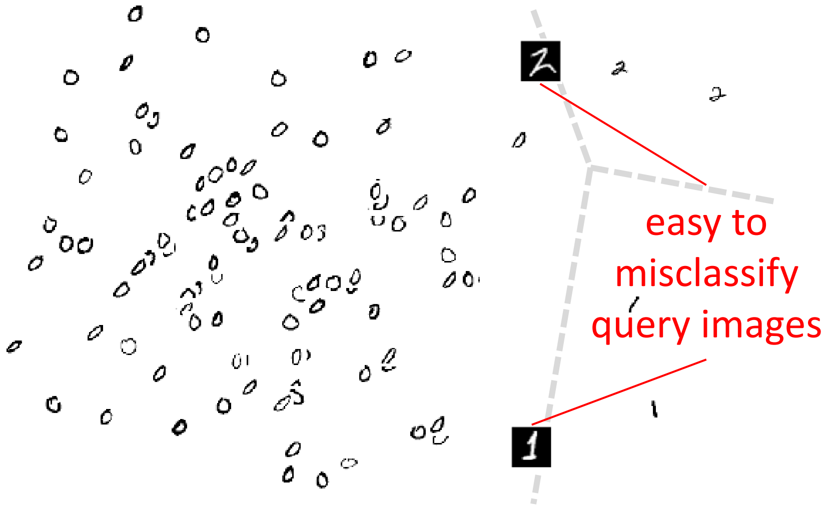



easy to misclassify query images

Motivation

- Deep Neural Networks (DNN) show human level performance on many recognition tasks.
- We focus on **class-imbalanced prediction**.
  - Insufficient samples to represent the true distribution of a class.

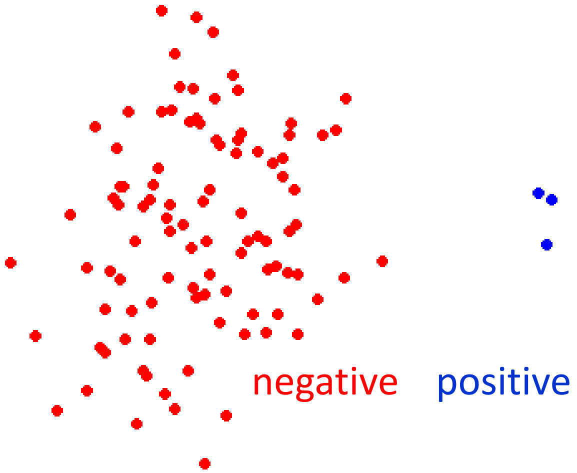


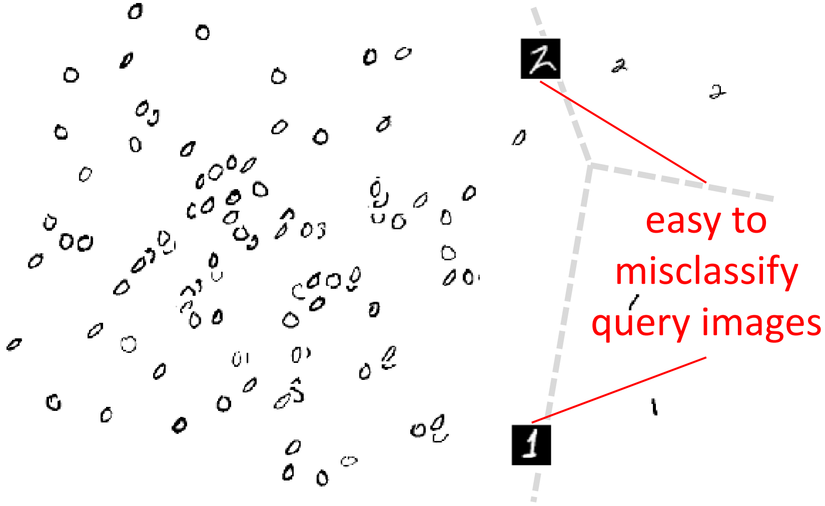


- Q. How can we learn **minor but important features** using neural networks?

Motivation

- Deep Neural Networks (DNN) show human level performance on many recognition tasks.
- We focus on **class-imbalanced prediction**.
    - Insufficient samples to represent the true distribution of a class.

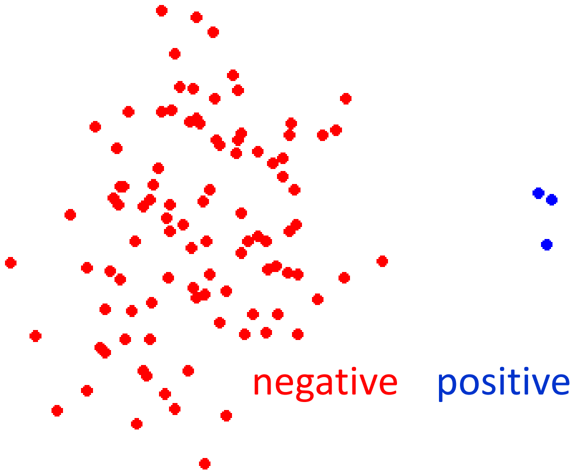


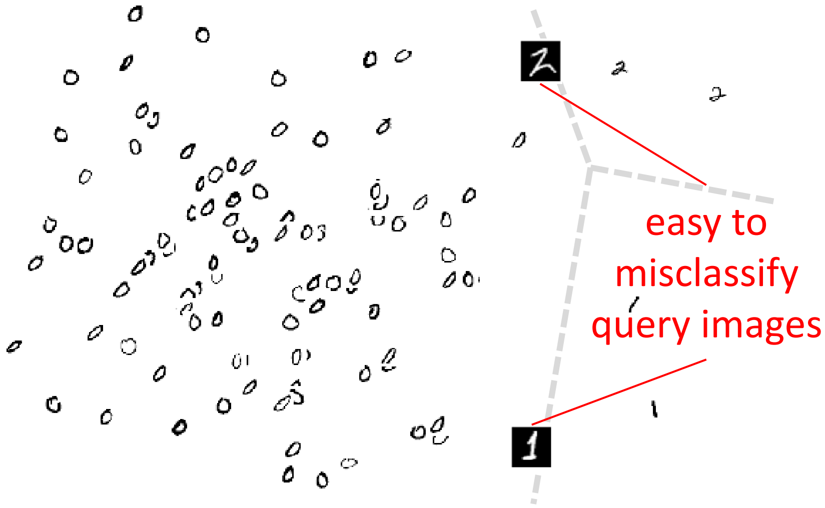


- Q. How can we learn **minor but important features** using neural networks?
    - We propose a **new RBM training method** called ***boosted CD***.

Motivation

- Deep Neural Networks (DNN) show human level performance on many recognition tasks.
- We focus on **class-imbalanced prediction**.
    - Insufficient samples to represent the true distribution of a class.





- Q. How can we learn **minor but important features** using neural networks?
    - We propose a **new RBM training method** called ***boosted CD***.
    - We also devise a **regularization term** for sparsity of DNA sequences.

(Splice) Junction Prediction: Extremely Class-Imbalanced Problem

- Genetic information flows through the **gene expression** process.
- **DNA: a sequence** of four types of nucleotides (A, G, T, C).
- **Gene: a segment of DNA** (the basic unit of heredity).

(Splice) Junction Prediction: Extremely Class-Imbalanced Problem

- Genetic information flows through the **gene expression** process.
- **DNA: a sequence** of four types of nucleotides (A, G, T, C).
- **Gene: a segment of DNA** (the basic unit of heredity).

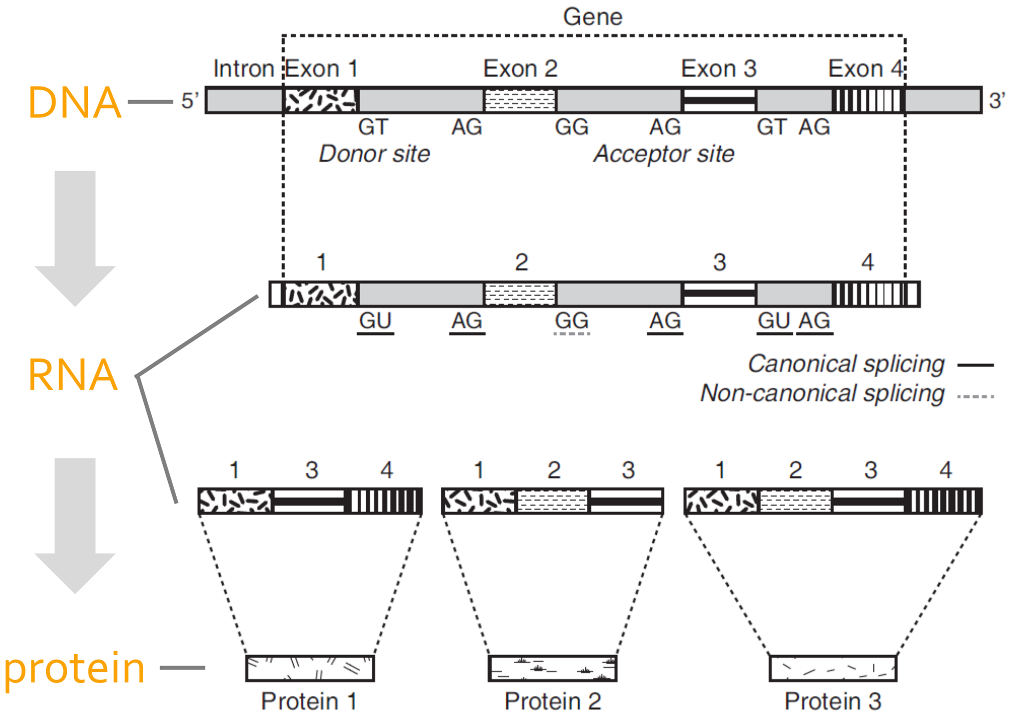
DNA

RNA

protein

(Splice) Junction Prediction: Extremely Class-Imbalanced Problem

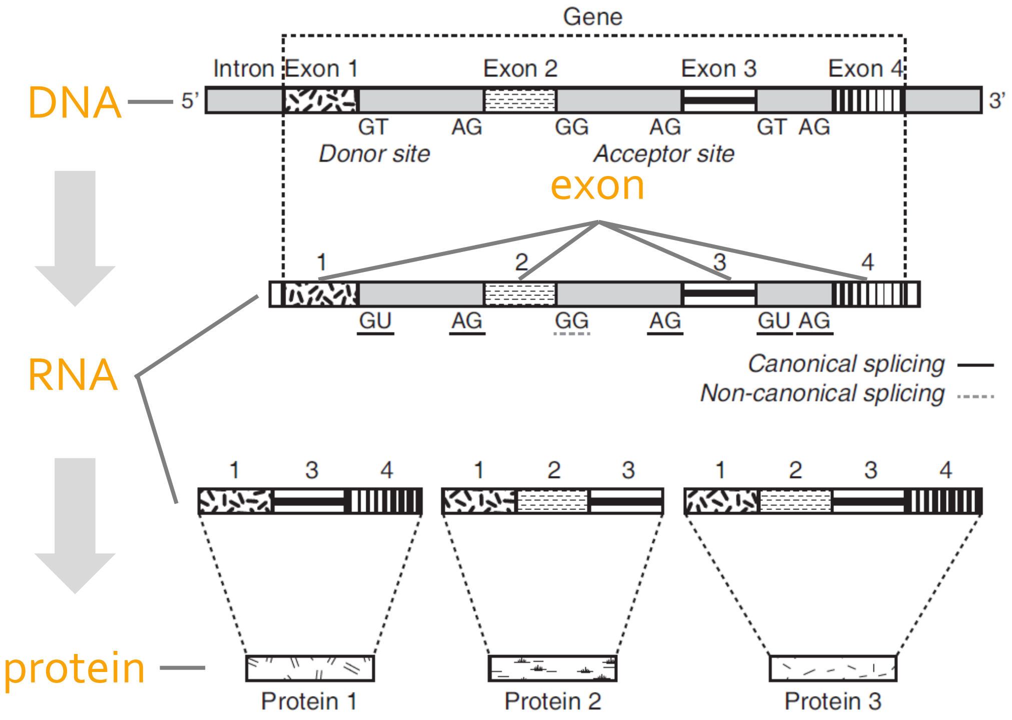
- Genetic information flows through the **gene expression** process.
- **DNA: a sequence** of four types of nucleotides (A, G, T, C).
- **Gene: a segment of DNA** (the basic unit of heredity).



gene expression

(Splice) Junction Prediction: Extremely Class-Imbalanced Problem

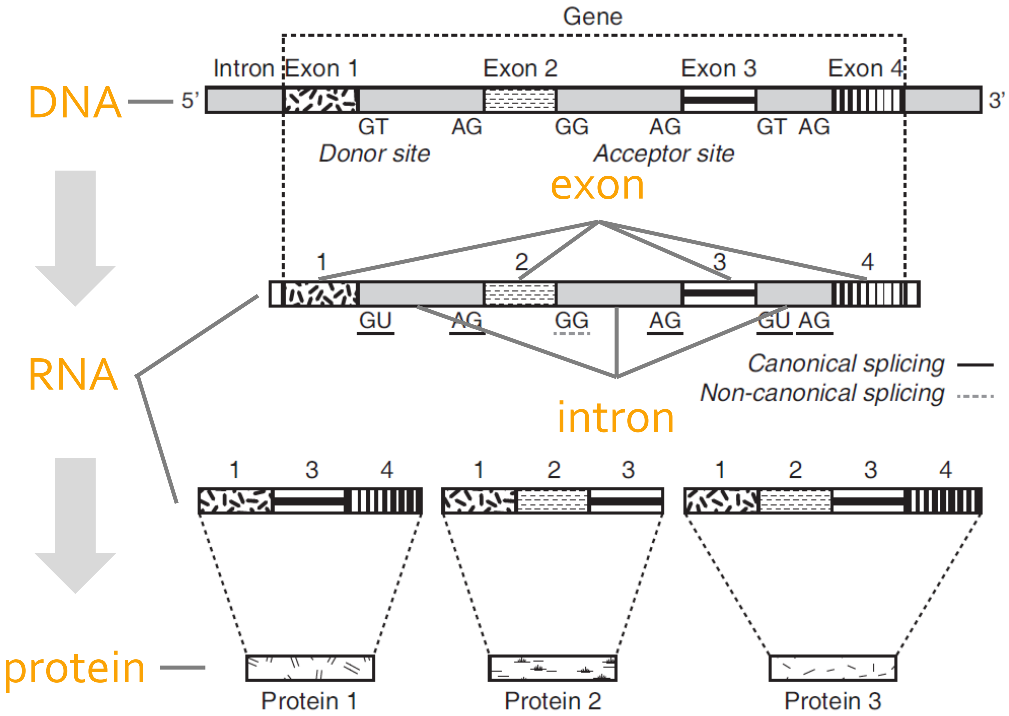
- Genetic information flows through the **gene expression** process.
- **DNA: a sequence** of four types of nucleotides (A, G, T, C).
- **Gene: a segment of DNA** (the basic unit of heredity).



gene expression

(Splice) Junction Prediction: Extremely Class-Imbalanced Problem

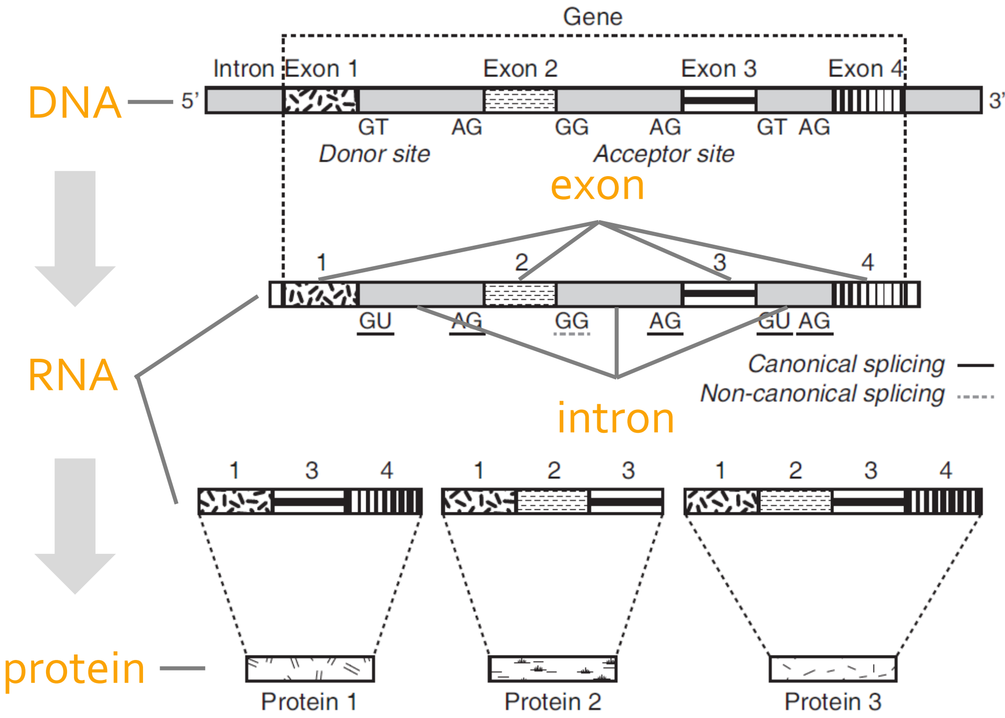
- Genetic information flows through the **gene expression** process.
- **DNA: a sequence** of four types of nucleotides (A, G, T, C).
- **Gene: a segment of DNA** (the basic unit of heredity).



gene expression

(Splice) Junction Prediction: Extremely Class-Imbalanced Problem

- Genetic information flows through the **gene expression** process.
- **DNA: a sequence** of four types of nucleotides (A, G, T, C).
- **Gene: a segment of DNA** (the basic unit of heredity).



gene expression

exon

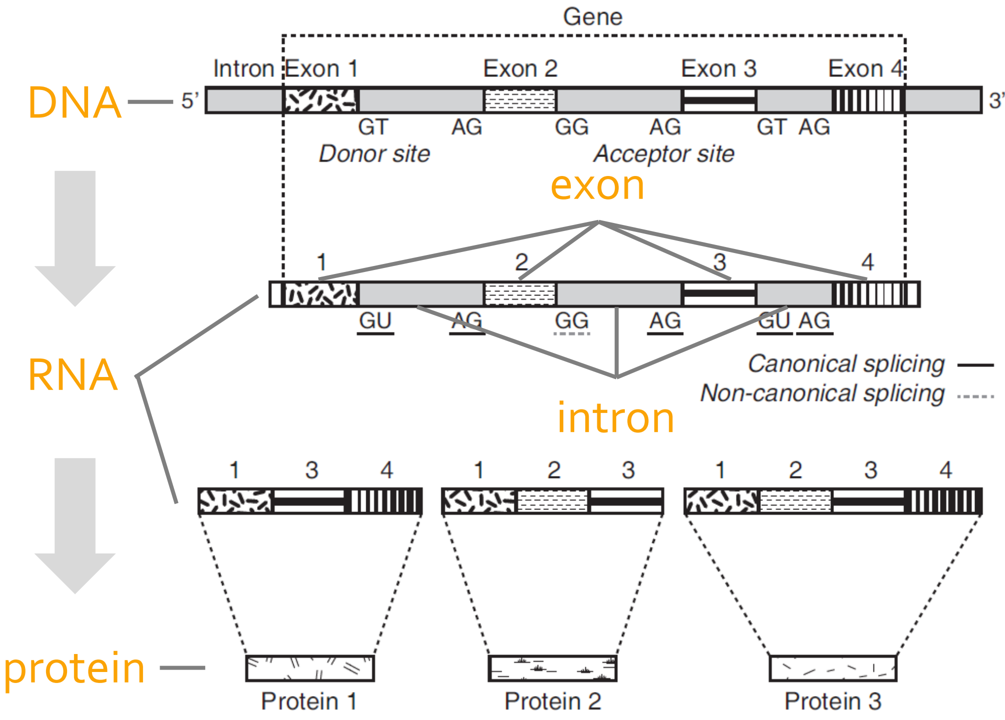
ACGTCGACTGCTACGTAGCAGCGA
TACGTACCGATCATCACTATCATC
GAGGTACGATCGATCGATCGATCA
GTCGATCGTCGTTCAGTCAGTCGA
TATCAGTCATATGCACATCTCAGT

GT: false boundary

GT: true boundary

(Splice) Junction Prediction: Extremely Class-Imbalanced Problem

- Genetic information flows through the **gene expression** process.
- **DNA: a sequence** of four types of nucleotides (A, G, T, C).
- **Gene: a segment of DNA** (the basic unit of heredity).



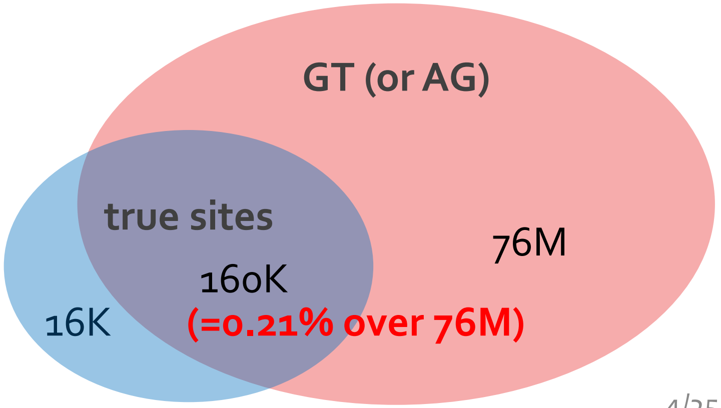
gene expression

exon

ACGTCGACTGCTACGTAGCAGCGA
TACGTACCGATCATCACTATCATC
GAGGTACGATCGATCGATCGATCA
GTCGATCGTCGTTCAGTCAGTCGA
TATCAGTCATATGCACATCTCAGT

GT: false boundary

GT: true boundary



Previous Work on Junction Prediction

- Two approaches:

1. Machine learning-based:
   - ANN (Stormo et al., 1982; Noordewier et al., 1990; Brunak et al., 1991),
   - SVM (Degroeve et al., 2005; Huang et al., 2006; Sonnenburg et al., 2007),
   - HMM (Reese et al., 1997; Pertea et al., 2001; Baten et al., 2006).
2. Sequence alignment-based:
   - TopHat (Trapnell et al., 2010), MapSplice (Wang et al., 2010), RUM (Grant et al., 2011).

Previous Work on Junction Prediction

- Two approaches:

1. Machine learning-based:
    - ANN (Stormo et al., 1982; Noordewier et al., 1990; Brunak et al., 1991),
    - SVM (Degroeve et al., 2005; Huang et al., 2006; Sonnenburg et al., 2007),
    - HMM (Reese et al., 1997; Pertea et al., 2001; Baten et al., 2006).
2. Sequence alignment-based:
    - TopHat (Trapnell et al., 2010), MapSplice (Wang et al., 2010), RUM (Grant et al., 2011).

1

We want to construct a **learning model** which can boost prediction performance in a **complementary** way to **alignment-based** method.

2

Previous Work on Junction Prediction

- Two approaches:

1 Machine learning-based:

- ANN (Stormo et al., 1982; Noordewier et al., 1990; Brunak et al., 1991),
- SVM (Degroeve et al., 2005; Huang et al., 2006; Sonnenburg et al., 2007),
- HMM (Reese et al., 1997; Pertea et al., 2001; Baten et al., 2006).

2 Sequence alignment-based:

- TopHat (Trapnell et al., 2010), MapSplice (Wang et al., 2010), RUM (Grant et al., 2011).

We want to construct a **learning model** [1] which can boost prediction performance in a **complementary** way to **alignment-based** [2] method.

We propose a **learning model** based on (multilayer) RBMs and its **training scheme**.

Related Methodologies

- Training methods of RBM

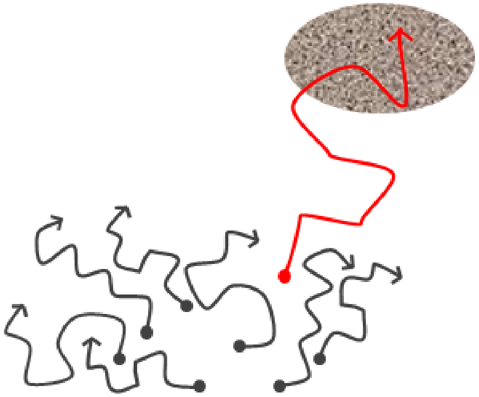
	Description	Training cost	Noise handling	Class-imbalance handling
CD (Hinton, Neural Comp. 2002)	Standard and widely used	-	-	-
Persistent CD (Tieleman, ICML 2008)	Use of a single Markov chain	-	☺	-
Parallel tempering (Cho et al., IJCNN 2010)	Simultaneous Markov chains generation	☹	☺	☺

- RBM for categorical values
  - Softmax input units (Salakhutdinov et al., ICML 2007).

- Class-imbalance problems
  - Refer to a review by Galar et al. (IEEE T SMC 2012).

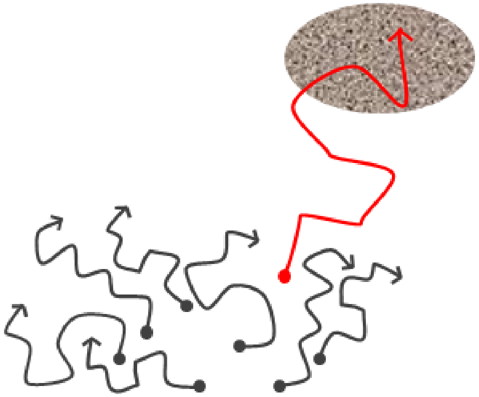
Main Contributions

Main Contributions



New RBM training methods called *boosted CD*

Main Contributions

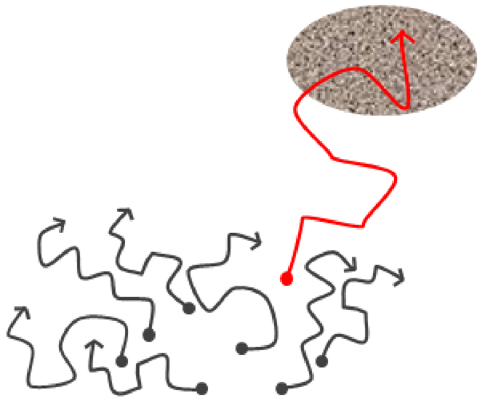


$$\phi(\mathbf{v}) = \frac{1}{2}\sum_{i=1}^{m}(\sum_{j=1}^{n_c} v^{(k)}_{n_c(i-1)+j} - 1)^2$$

New RBM training methods called *boosted CD*

New penalty term to handle sparsity of DNA sequences

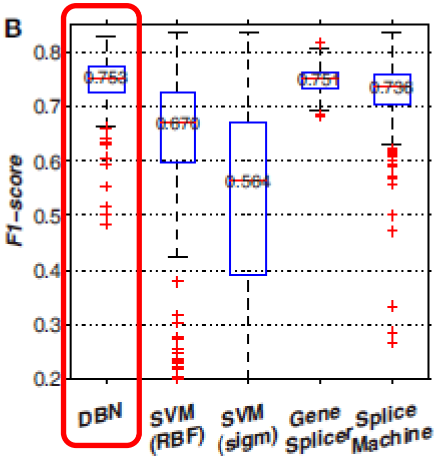
Main Contributions



$$\phi(\mathbf{v}) = \frac{1}{2}\sum_{i=1}^{m}(\sum_{j=1}^{n_c} v^{(k)}_{n_c(i-1)+j} - 1)^2$$

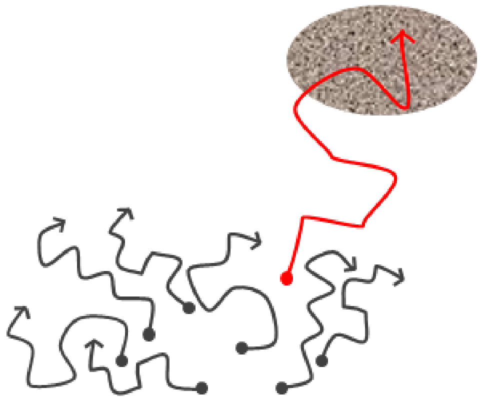
New RBM training methods called *boosted CD*

New penalty term to handle sparsity of DNA sequences



Significant boosts in splicing prediction performance

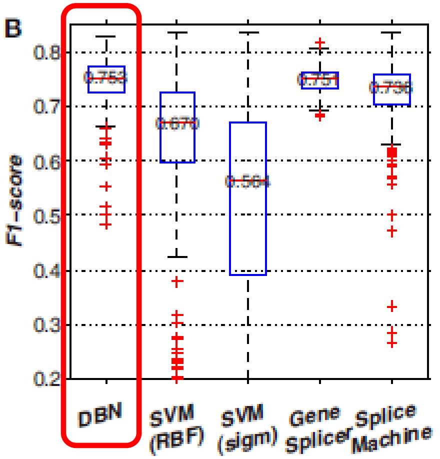
Main Contributions

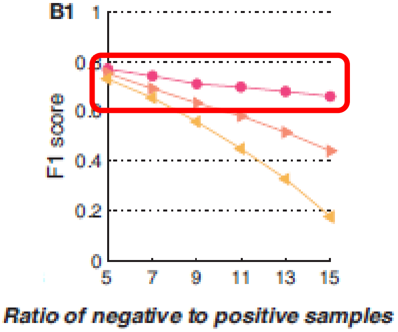


$$\phi(\mathbf{v}) = \frac{1}{2}\sum_{i=1}^{m}(\sum_{j=1}^{n_c} v^{(k)}_{n_c(i-1)+j} - 1)^2$$

New RBM training methods called *boosted CD*

New penalty term to handle sparsity of DNA sequences

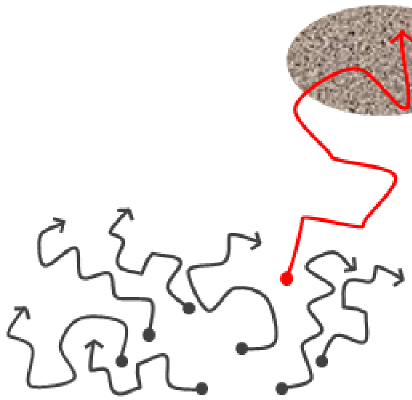




Significant boosts in splicing prediction performance

Robustness to high-dimensional class-imbalanced data

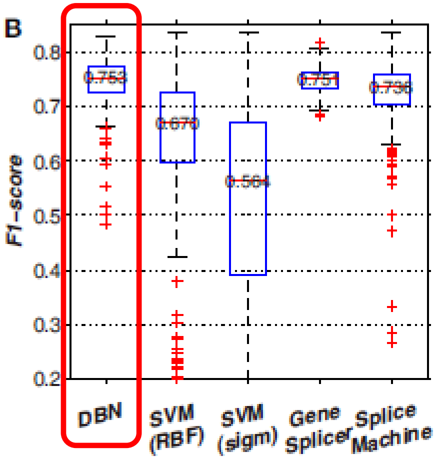
Main Contributions

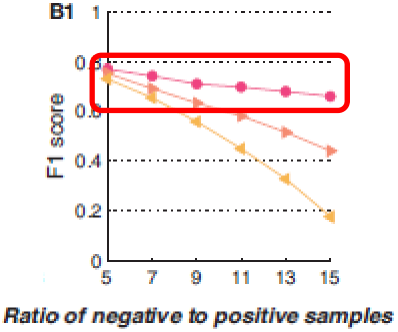


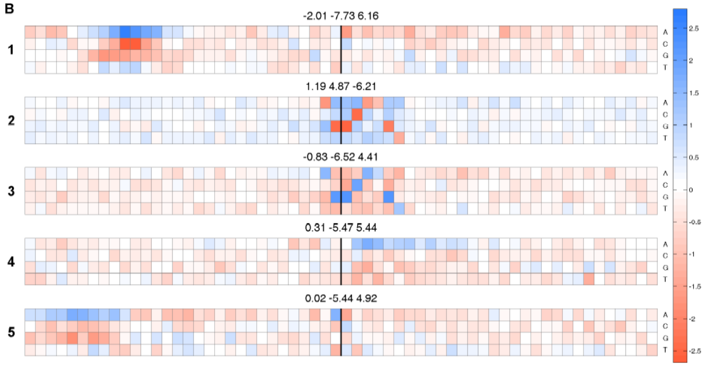
$$\phi(\mathbf{v}) = \frac{1}{2}\sum_{i=1}^{m}(\sum_{j=1}^{n_c} v^{(k)}_{n_c(i-1)+j} - 1)^2$$

New RBM training methods called *boosted CD*

New penalty term to handle sparsity of DNA sequences







Significant boosts in splicing prediction performance

Robustness to high-dimensional class-imbalanced data

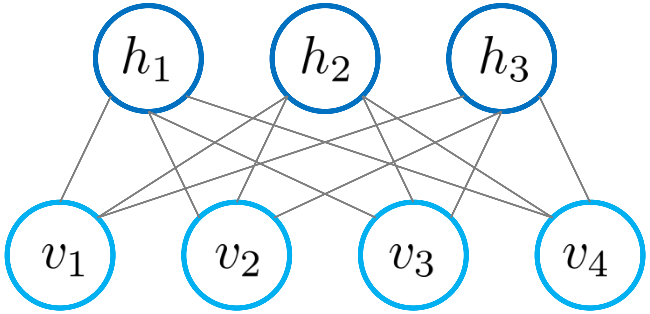
The ability to detect subtle non-canonical splicing signals

Outline

- Motivation
- Preliminary
- Boosted contrastive divergence
- Categorical restricted Boltzmann machine
- Experiment results
- Conclusion

Restricted Boltzmann Machines

- RBM is a type of logistic belief network whose structure is a bipartite graph.
  - Nodes:
    - Input layer: $\mathbf{v} = \{v_1, ..., v_{n_v}\}$
    - Hidden layer: $\mathbf{h} = \{h_1, ..., h_{n_h}\}$



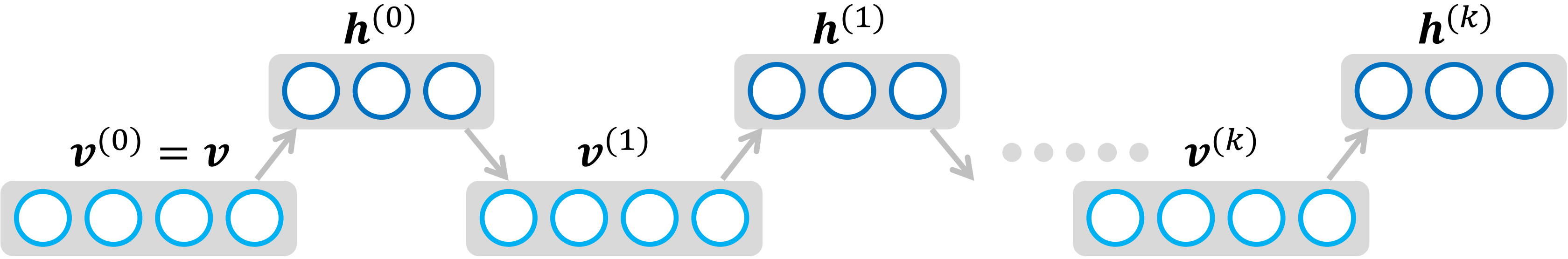
  - Probability of a configuration $(\mathbf{v}, \mathbf{h})$:
    - $P(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{v},\mathbf{h})} \quad Z = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}$
    - $$E(\mathbf{v}, \mathbf{h}) = -\sum_{i=1}^{n_v} \sum_{j=1}^{n_h} v_i w_{ij} h_j - \sum_{i=1}^{n_v} b_i v_i - \sum_{j=1}^{n_h} c_j h_j.$$
  - Each node is a stochastic binary unit:
    - $P(v_i = 1 | \mathbf{h}) = \mathrm{sigm}(b_i + \sum_{j=1}^{n_h} w_{ij} h_j)$
    - $P(h_j = 1 | \mathbf{v}) = \mathrm{sigm}(c_j + \sum_{i=1}^{n_v} v_i w_{ij})$ can be used as a feature.

Contrastive Divergence (CD) for Training RBMs

- Training weights to minimize **negative log-likelihood** of data.

$$W^*, \mathbf{b}^*, \mathbf{c}^* = \arg\min_{W,\mathbf{b},\mathbf{c}} \underbrace{\mathbf{E}[-\sum_{n=1}^{N} \log P(\mathbf{v}_n)]}_{L(W,\mathbf{b},\mathbf{c};\mathbf{v}_1,\dots,\mathbf{v}_N)} .$$

  - Run the MCMC chain $\boldsymbol{v}^{(0)}, \boldsymbol{v}^{(1)}, \dots, \boldsymbol{v}^{(k)}$ for k steps.



  - The **CD-k updates** after seeing example $\boldsymbol{v}$:

$$\frac{\partial L}{\partial w_{ij}} = \frac{\partial}{\partial w_{ij}} \mathbf{E}\left[-\log \left(\frac{\sum_{\mathbf{h}} e^{-E(\mathbf{v}_n,\mathbf{h})}}{\sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}} \right) \right]$$
$$= \mathbf{E}_{data}[v_i h_j] - \mathbf{E}_{model}[v_i h_j].$$

$$\frac{\partial L}{\partial W} \approx \frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{v}_n^{(0)} \mathbf{h}_n^{(0)^T} - \mathbf{v}_n^{(k)} \mathbf{h}_n^{(k)^T} \right), \qquad (2)$$

$$\frac{\partial L}{\partial \mathbf{b}} \approx \frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{v}_n^{(0)} - \mathbf{v}_n^{(k)} \right),$$

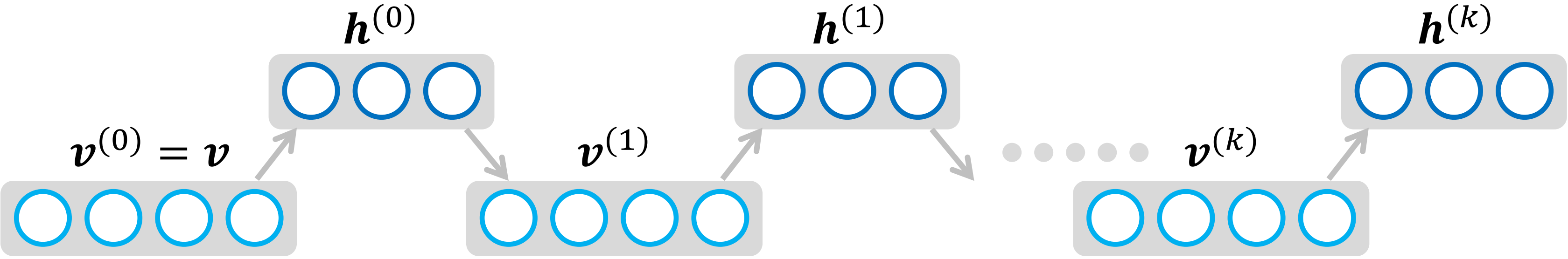
$$\frac{\partial L}{\partial \mathbf{c}} \approx \frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{h}_n^{(0)} - \mathbf{h}_n^{(k)} \right).$$

Contrastive Divergence (CD) for Training RBMs

- Training weights to minimize **negative log-likelihood** of data.

$$W^*, \mathbf{b}^*, \mathbf{c}^* = \arg\min_{W,\mathbf{b},\mathbf{c}} \underbrace{\mathbf{E}[-\sum_{n=1}^{N} \log P(\mathbf{v}_n)]}_{L(W,\mathbf{b},\mathbf{c};\mathbf{v}_1,\ldots,\mathbf{v}_N)}.$$

  - Run the MCMC chain $\boldsymbol{v}^{(0)}, \boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(k)}$ for k steps.



  - The **CD-k updates** after seeing example $\boldsymbol{v}$:

$$\frac{\partial L}{\partial w_{ij}} = \frac{\partial}{\partial w_{ij}} \mathbf{E}\left[-\log \left(\frac{\sum_{\mathbf{h}} e^{-E(\mathbf{v}_n,\mathbf{h})}}{\sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}} \right) \right]$$
$$= \mathbf{E}_{data}[v_i h_j] - \mathbf{E}_{model}[v_i h_j].$$

approximated by
k-step Markov chain

$$\frac{\partial L}{\partial W} \approx \frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{v}_n^{(0)} \mathbf{h}_n^{(0)^T} - \mathbf{v}_n^{(k)} \mathbf{h}_n^{(k)^T} \right), \qquad (2)$$

$$\frac{\partial L}{\partial \mathbf{b}} \approx \frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{v}_n^{(0)} - \mathbf{v}_n^{(k)} \right),$$

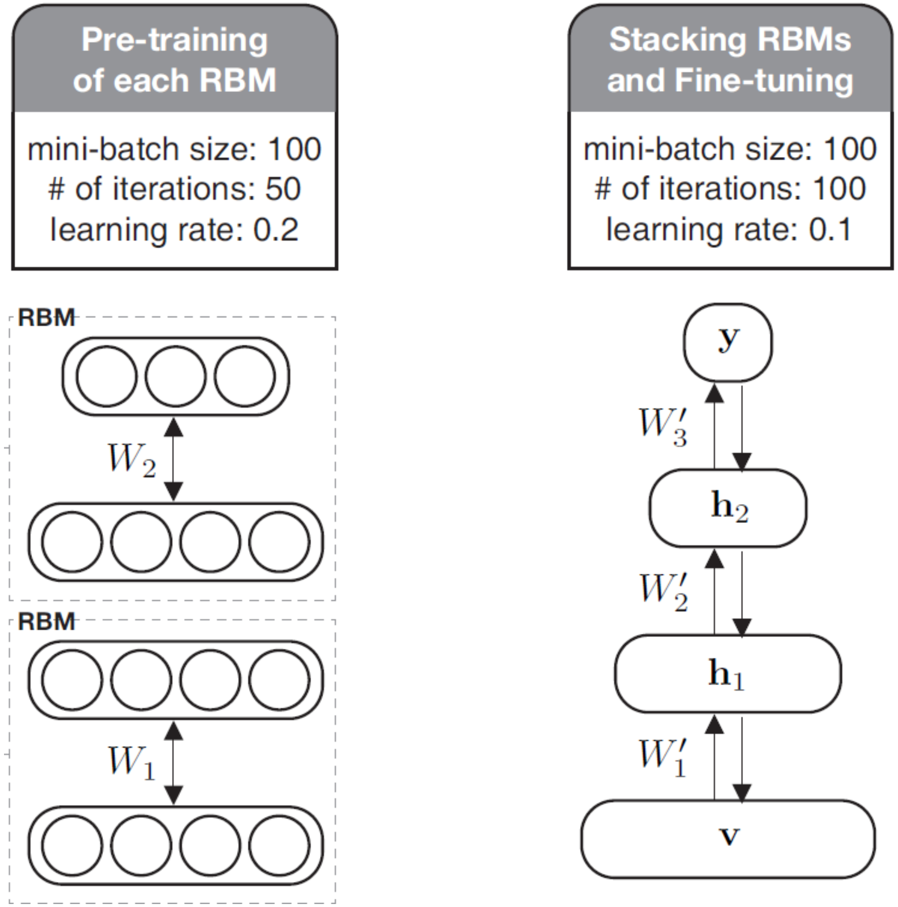
$$\frac{\partial L}{\partial \mathbf{c}} \approx \frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{h}_n^{(0)} - \mathbf{h}_n^{(k)} \right).$$

Outline

- Motivation
- Preliminary
- Boosted contrastive divergence
- Categorical restricted Boltzmann machine
- Experiment results
- Conclusion

Overview of Proposed Methodology

Overview of Proposed Methodology



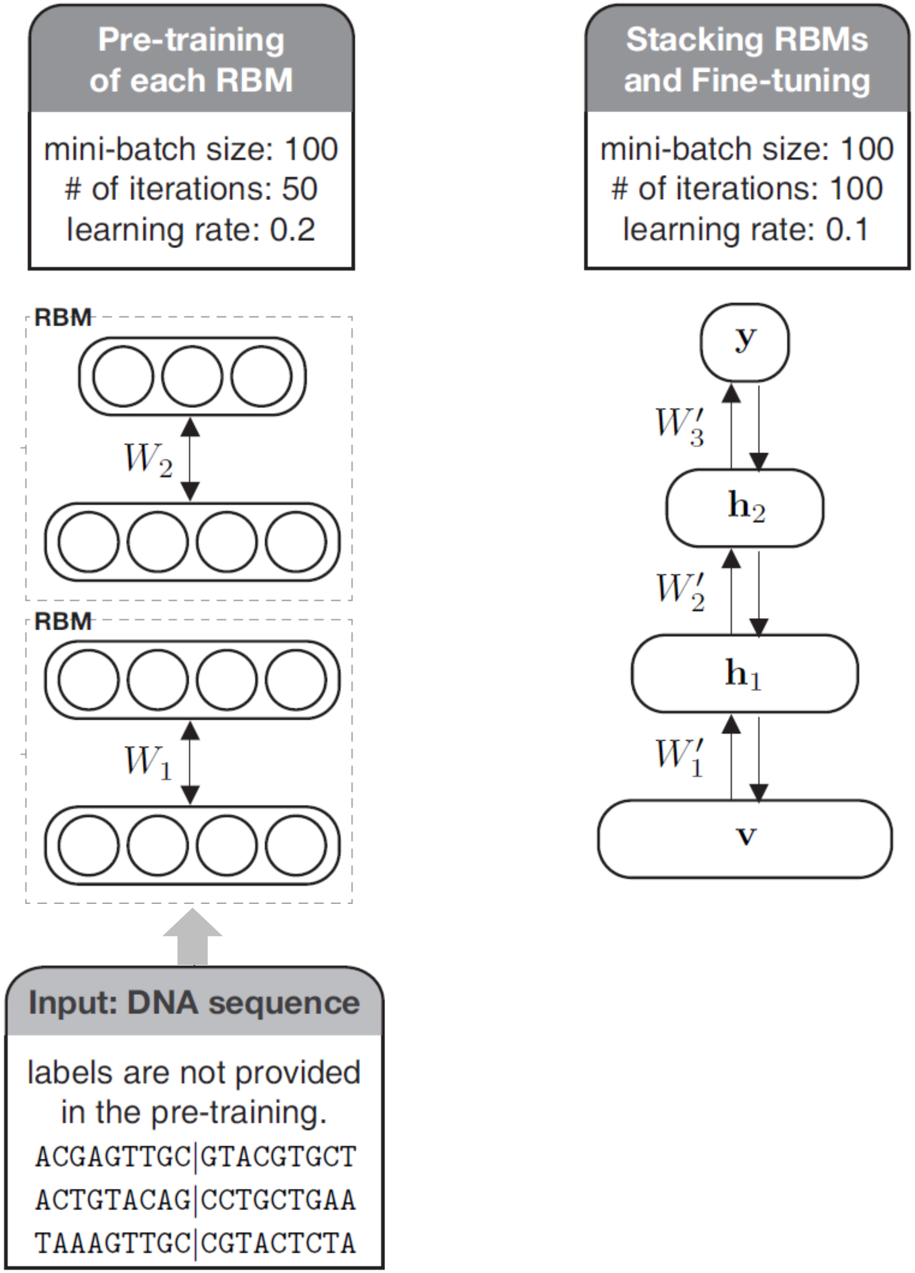
Pre-training of each RBM

mini-batch size: 100
# of iterations: 50
learning rate: 0.2

Stacking RBMs and Fine-tuning

mini-batch size: 100
# of iterations: 100
learning rate: 0.1

Overview of Proposed Methodology



Pre-training of each RBM

mini-batch size: 100
# of iterations: 50
learning rate: 0.2

Stacking RBMs and Fine-tuning

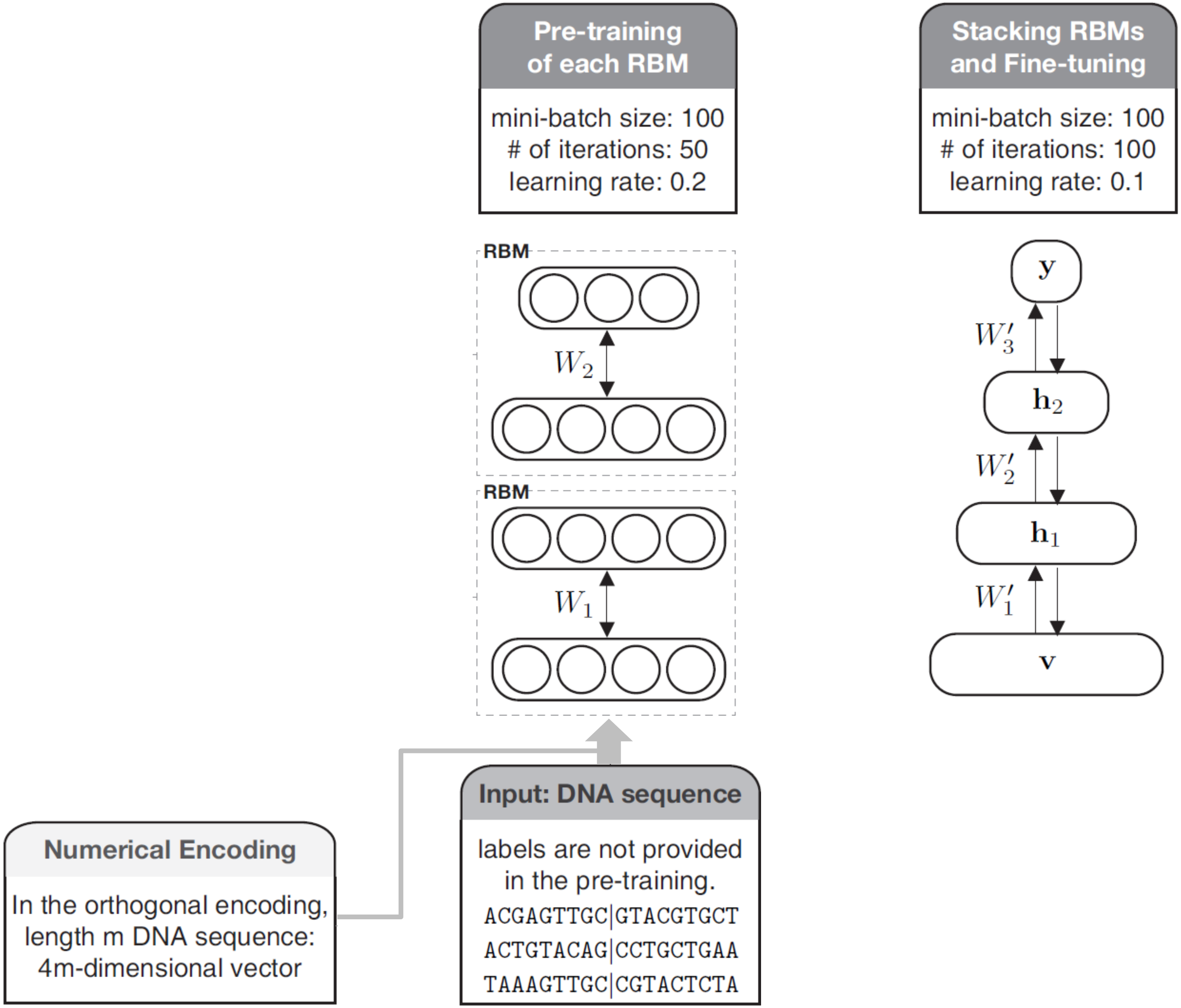
mini-batch size: 100
# of iterations: 100
learning rate: 0.1

Input: DNA sequence

labels are not provided in the pre-training.

```
ACGAGTTGC|GTACGTGCT
ACTGTACAG|CCTGCTGAA
TAAAGTTGC|CGTACTCTA
```

Overview of Proposed Methodology



Pre-training of each RBM

mini-batch size: 100
# of iterations: 50
learning rate: 0.2

Stacking RBMs and Fine-tuning

mini-batch size: 100
# of iterations: 100
learning rate: 0.1

Input: DNA sequence

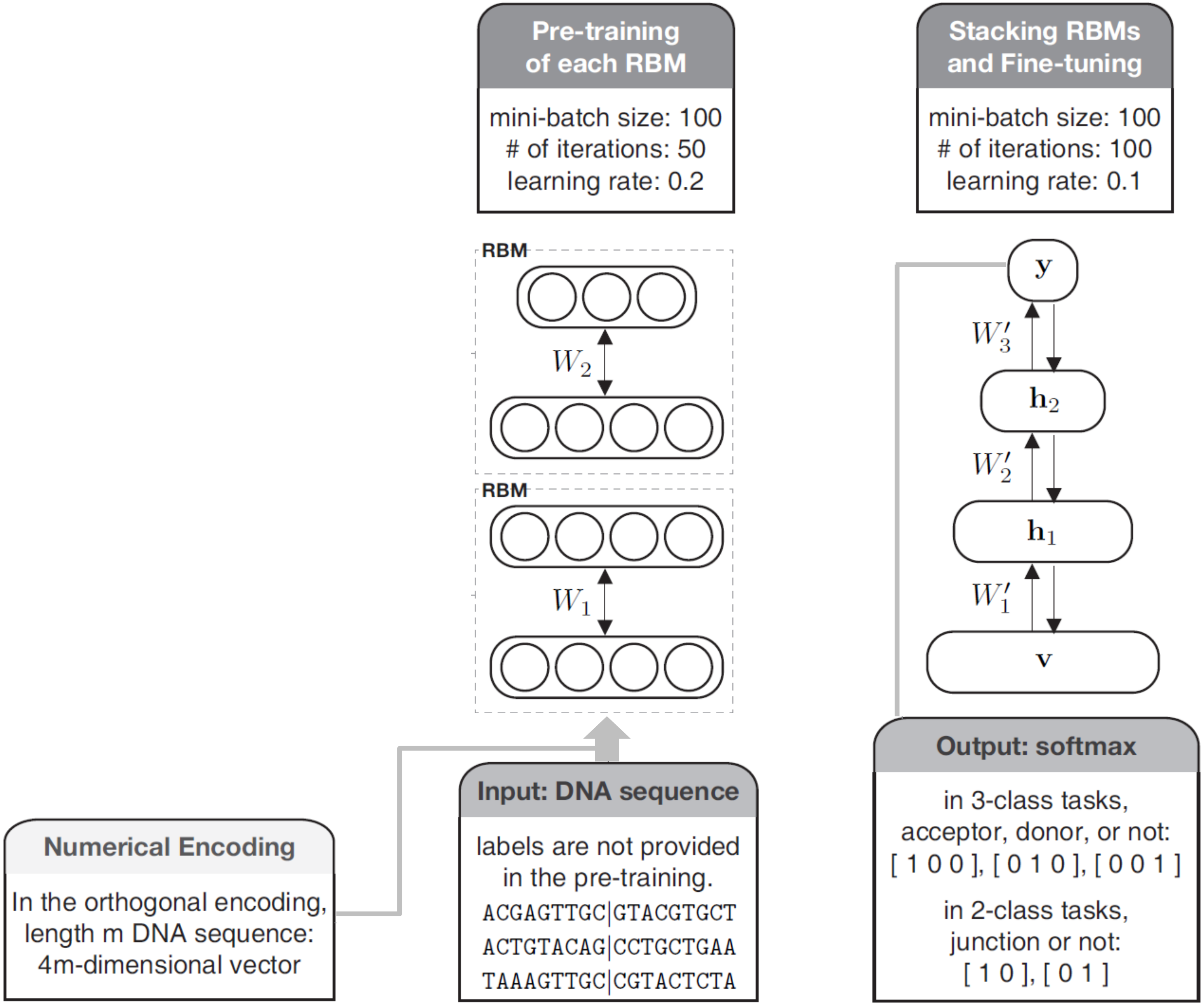
labels are not provided in the pre-training.

```
ACGAGTTGC|GTACGTGCT
ACTGTACAG|CCTGCTGAA
TAAAGTTGC|CGTACTCTA
```

Numerical Encoding

In the orthogonal encoding, length m DNA sequence: 4m-dimensional vector

Overview of Proposed Methodology



Pre-training of each RBM

mini-batch size: 100
# of iterations: 50
learning rate: 0.2

Stacking RBMs and Fine-tuning

mini-batch size: 100
# of iterations: 100
learning rate: 0.1

Numerical Encoding

In the orthogonal encoding, length m DNA sequence: 4m-dimensional vector

Input: DNA sequence

labels are not provided in the pre-training.

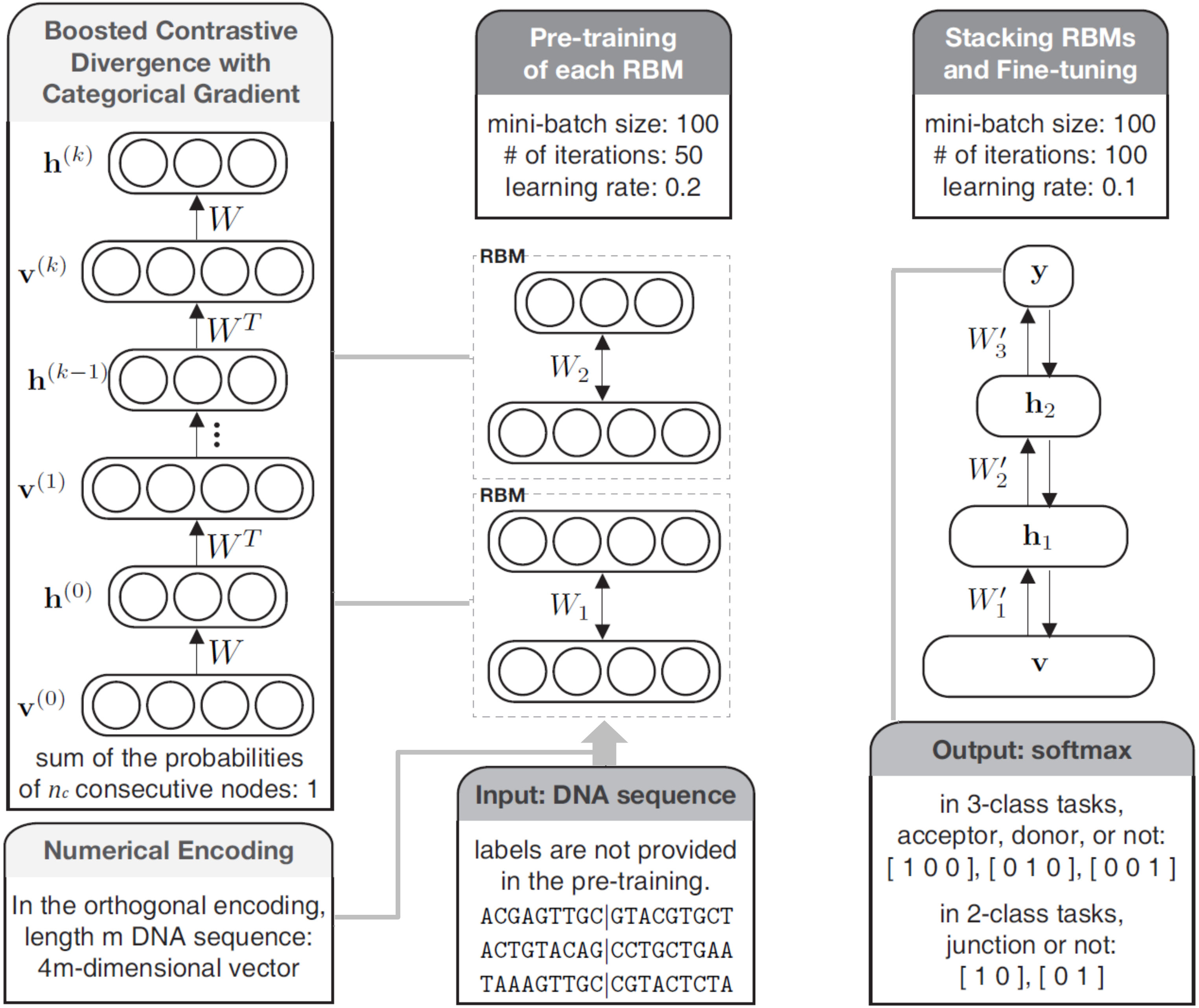
```
ACGAGTTGC|GTACGTGCT
ACTGTACAG|CCTGCTGAA
TAAAGTTGC|CGTACTCTA
```

Output: softmax

in 3-class tasks, acceptor, donor, or not:
[1 0 0], [0 1 0], [0 0 1]

in 2-class tasks, junction or not:
[1 0], [0 1]

Overview of Proposed Methodology



Boosted Contrastive Divergence with Categorical Gradient

sum of the probabilities of n_c consecutive nodes: 1

Numerical Encoding

In the orthogonal encoding, length m DNA sequence: 4m-dimensional vector

Pre-training of each RBM

mini-batch size: 100
# of iterations: 50
learning rate: 0.2

Input: DNA sequence

labels are not provided in the pre-training.

ACGAGTTGC|GTACGTGCT
ACTGTACAG|CCTGCTGAA
TAAAGTTGC|CGTACTCTA

Stacking RBMs and Fine-tuning

mini-batch size: 100
# of iterations: 100
learning rate: 0.1

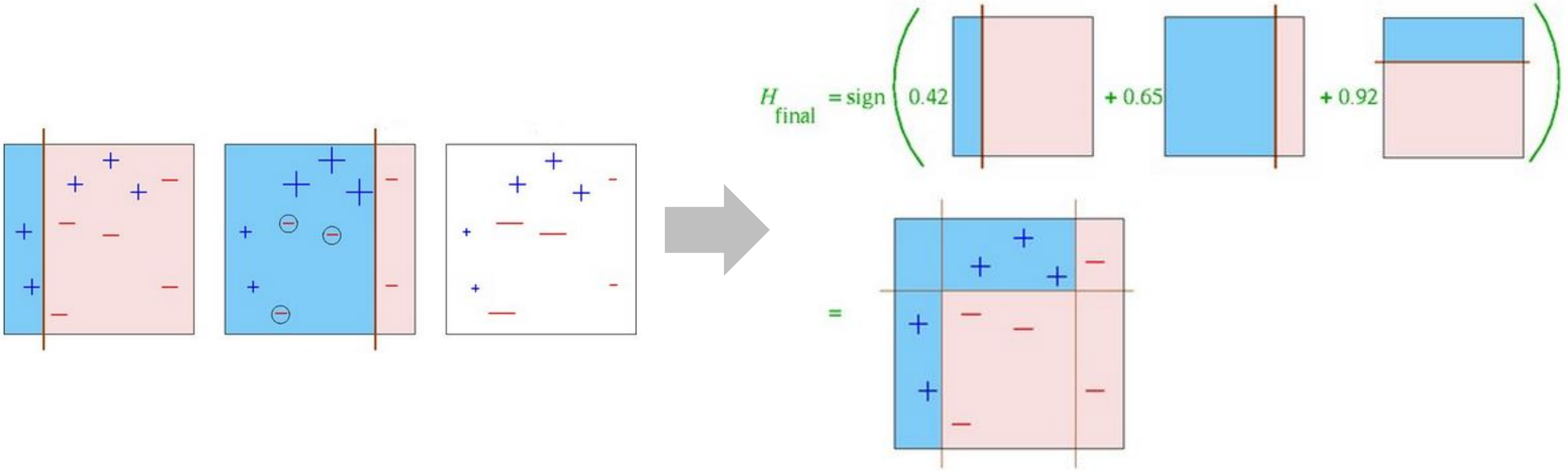
Output: softmax

in 3-class tasks, acceptor, donor, or not: [1 0 0], [0 1 0], [0 0 1]

in 2-class tasks, junction or not: [1 0], [0 1]

What Boosting Is

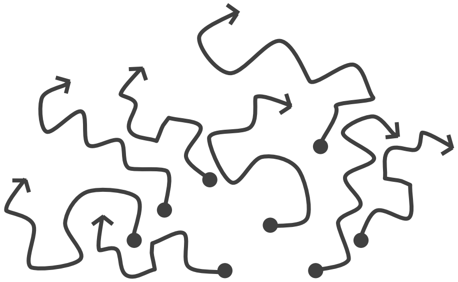
- Boosting is a meta-algorithm which converts weak learners to strong ones.
- Most boosting algorithms consist of **iteratively learning weak classifiers** with respect to a distribution and adding them to a final strong classifier.
- The main variation between many boosting algorithms:
  - The method of **weighting training data points** and hypotheses.
  - AdaBoost, LPBoost, TotalBoost, …



from lecture notes @ UCIrvine CS 271 Fall 2007

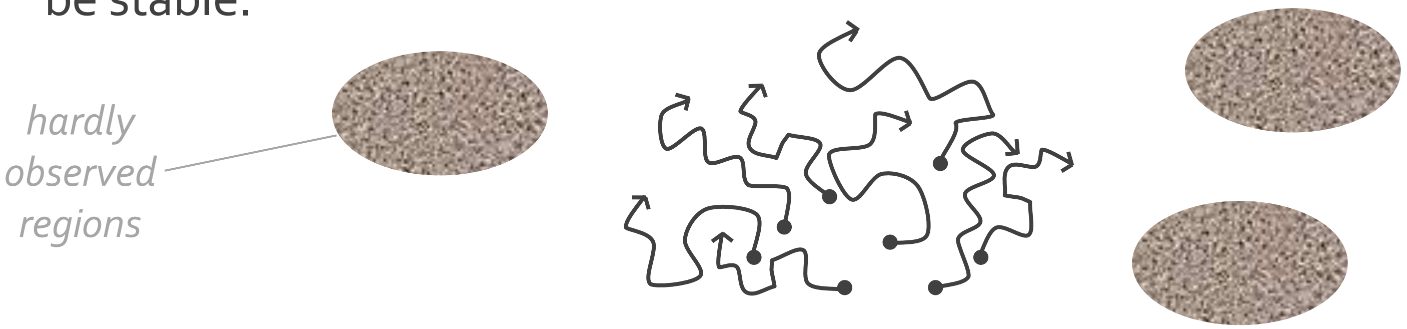
Boosted Contrastive Divergence (1/2)

- Contrastive divergence training is looped over all mini-batches and known to be stable.



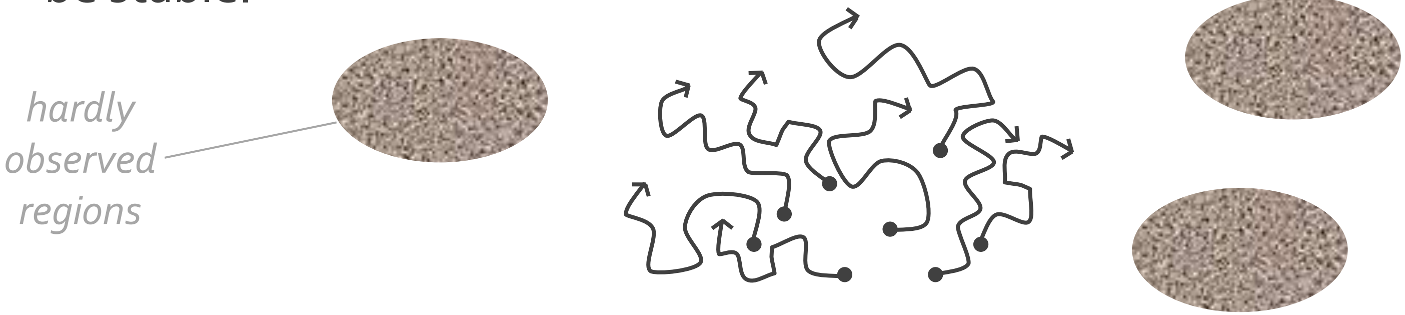
Boosted Contrastive Divergence (1/2)

- Contrastive divergence training is looped over all mini-batches and known to be stable.



Boosted Contrastive Divergence (1/2)

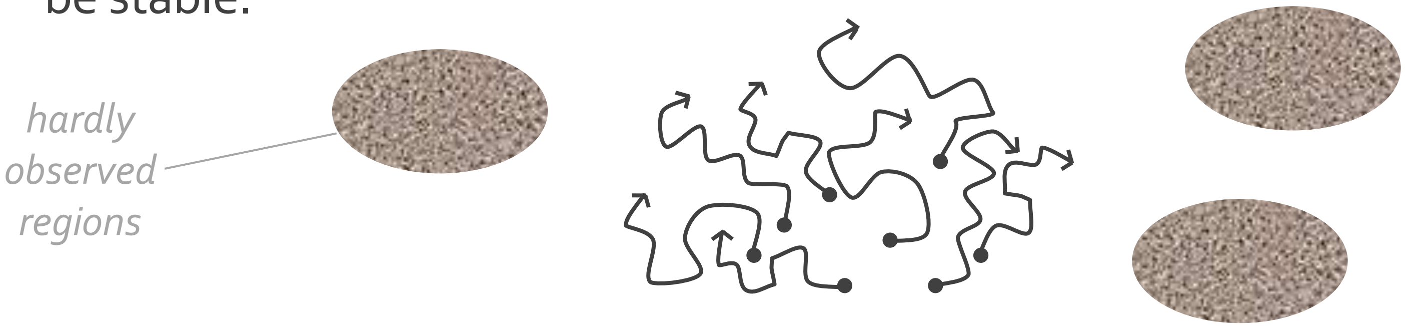
- Contrastive divergence training is looped over all mini-batches and known to be stable.



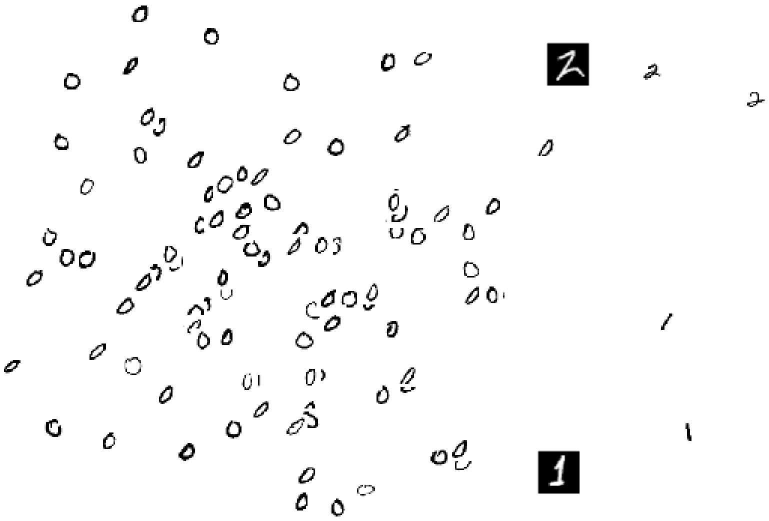
- However, for a class-imbalance distribution, we need to **assign higher weights to rare samples** in order to **jump to unseen examples** by Gibbs chains.

Boosted Contrastive Divergence (1/2)

- Contrastive divergence training is looped over all mini-batches and known to be stable.

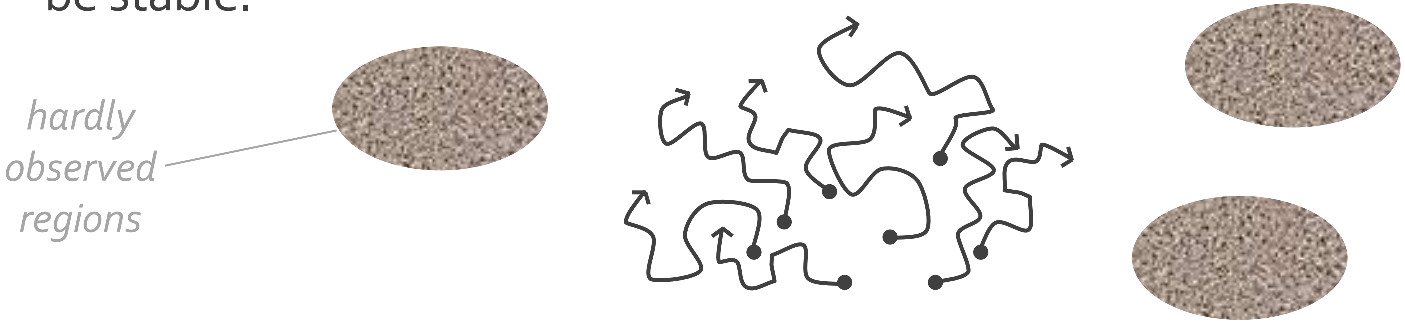


- However, for a class-imbalance distribution, we need to **assign higher weights to rare samples** in order to **jump to unseen examples** by Gibbs chains.

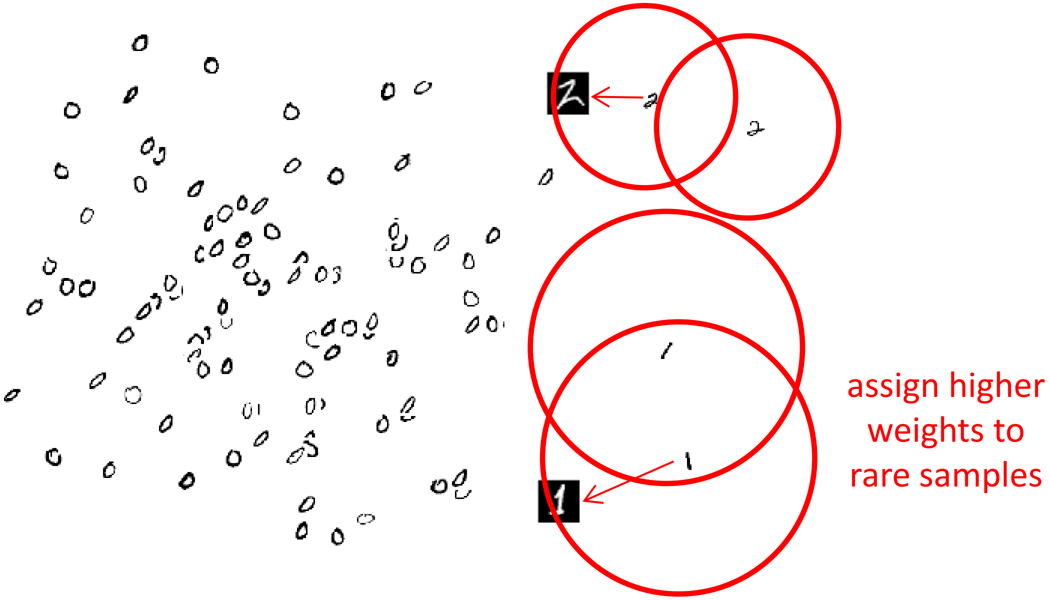


Boosted Contrastive Divergence (1/2)

- Contrastive divergence training is looped over all mini-batches and known to be stable.

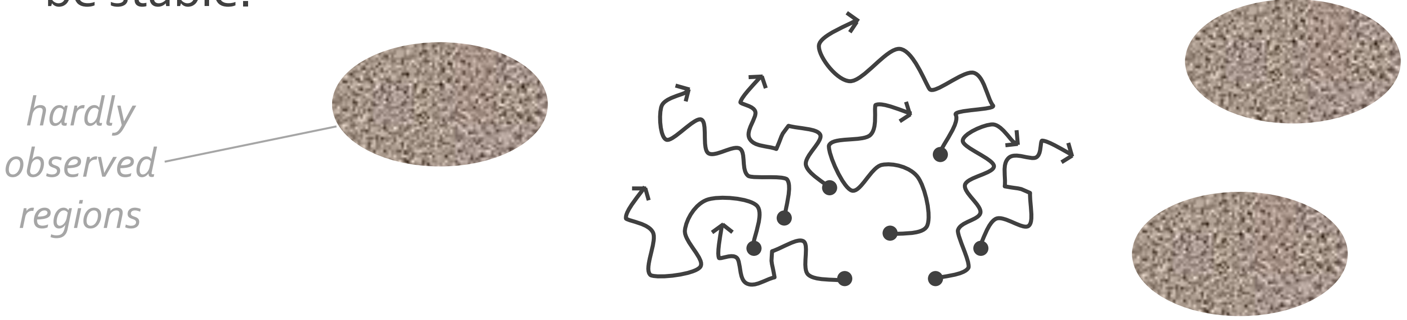


- However, for a class-imbalance distribution, we need to **assign higher weights to rare samples** in order to **jump to unseen examples** by Gibbs chains.

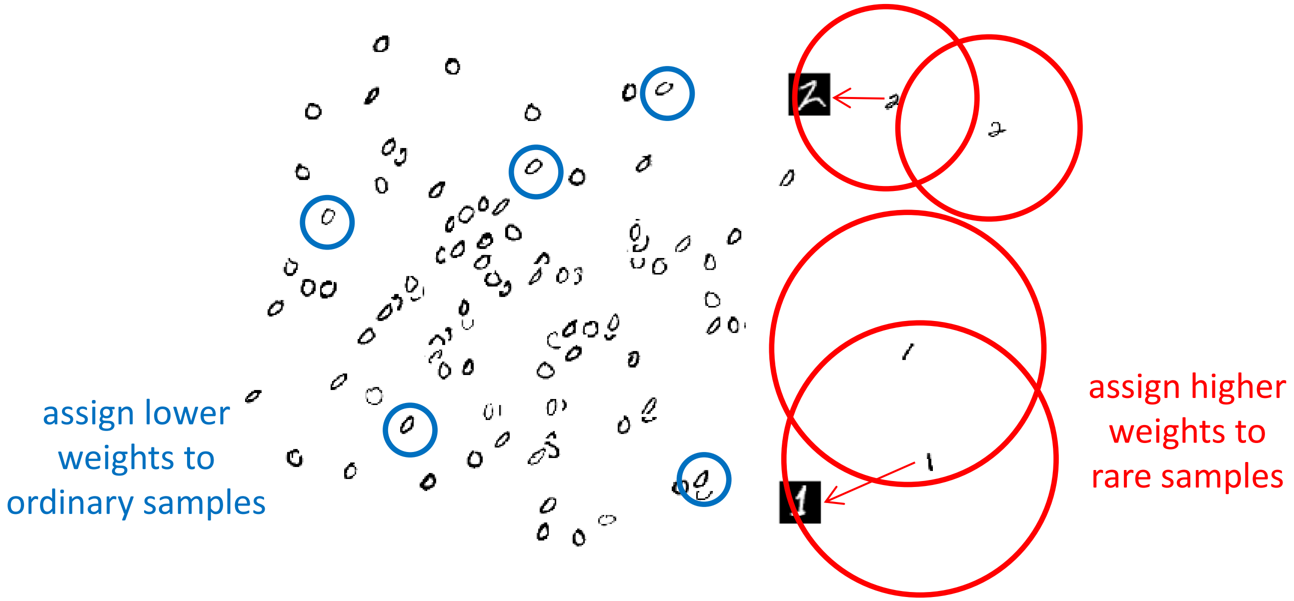


Boosted Contrastive Divergence (1/2)

- Contrastive divergence training is looped over all mini-batches and known to be stable.



- However, for a class-imbalance distribution, we need to **assign higher weights to rare samples** in order to **jump to unseen examples** by Gibbs chains.

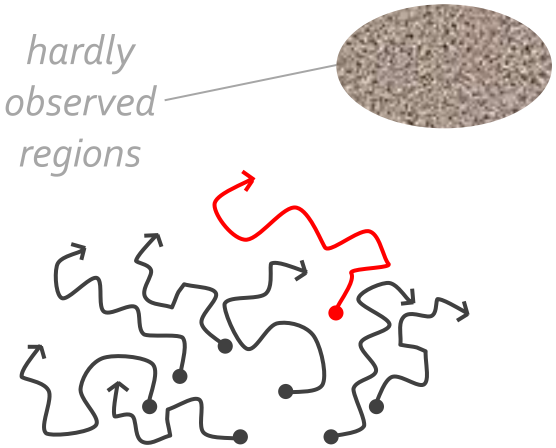

assign lower weights to ordinary samples
assign higher weights to rare samples

Boosted Contrastive Divergence (2/2)

- If we assign the same weight to all the data, the performance of Gibbs sampling would degrade in the regions that are hardly observed.
- Whenever sampling, we therefore **re-weight each observation by the energy** of its reconstruction $E(\boldsymbol{v}_n{}^{(k)}, \boldsymbol{h}_n{}^{(k)})$.

Boosted Contrastive Divergence (2/2)

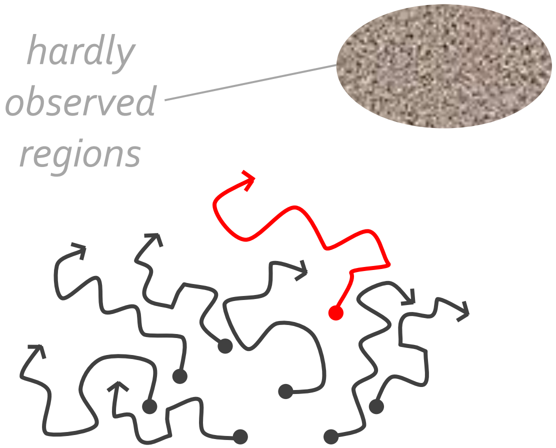
- If we assign the same weight to all the data, the performance of Gibbs sampling would degrade in the regions that are hardly observed.
- Whenever sampling, we therefore **re-weight each observation by the energy** of its reconstruction $E(\boldsymbol{v}_n{}^{(k)}, \boldsymbol{h}_n{}^{(k)})$.



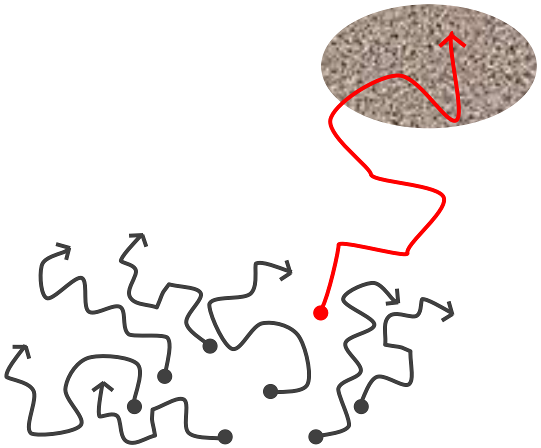
Relative locations of samples and corresponding Markov chains by CD

Boosted Contrastive Divergence (2/2)

- If we assign the same weight to all the data, the performance of Gibbs sampling would degrade in the regions that are hardly observed.
- Whenever sampling, we therefore **re-weight each observation by the energy** of its reconstruction $E({\boldsymbol{v}_n}^{(k)}, {\boldsymbol{h}_n}^{(k)})$.



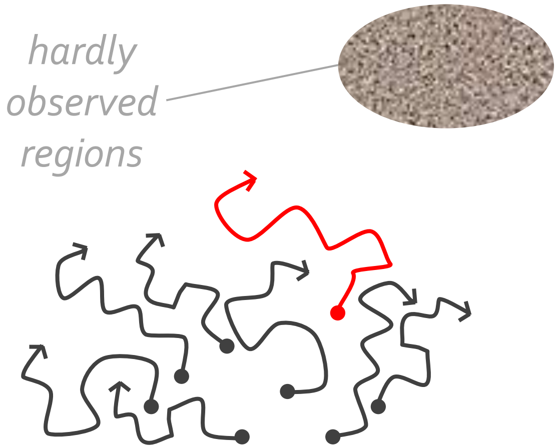
Relative locations of samples and corresponding Markov chains by CD



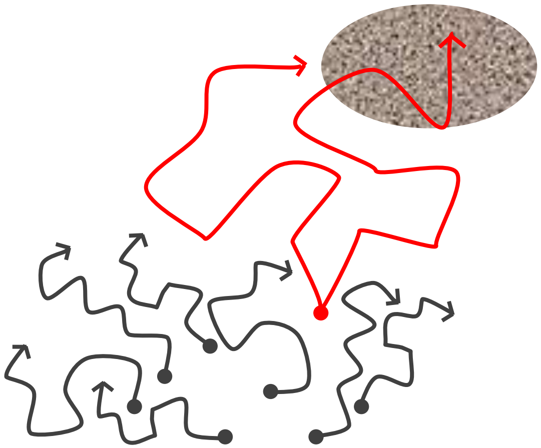
Relative locations of samples and corresponding Markov chains by the proposed

Boosted Contrastive Divergence (2/2)

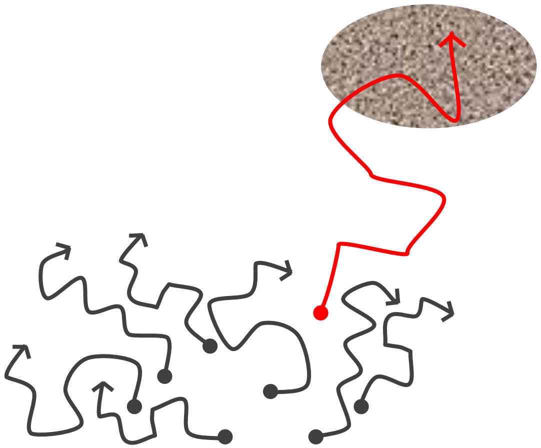
- If we assign the same weight to all the data, the performance of Gibbs sampling would degrade in the regions that are hardly observed.
- Whenever sampling, we therefore **re-weight each observation by the energy** of its reconstruction $E({\boldsymbol{v}_n}^{(k)}, {\boldsymbol{h}_n}^{(k)})$.



Relative locations of samples and corresponding Markov chains by CD



Relative locations of samples and corresponding Markov chains by PT



Relative locations of samples and corresponding Markov chains by the proposed

Categorical Gradient

- For biological sequences, 1-hot encoding is widely used (Baldi & Brunak, 2001).
    - A, C, G, and T are encoded by 1000, 0100, 0010, and 0001, respectively.
    - In encoded binary vectors, 75% of the elements are zero.

Categorical Gradient

- For biological sequences, 1-hot encoding is widely used (Baldi & Brunak, 2001).
    - A, C, G, and T are encoded by 1000, 0100, 0010, and 0001, respectively.
    - In encoded binary vectors, 75% of the elements are zero.

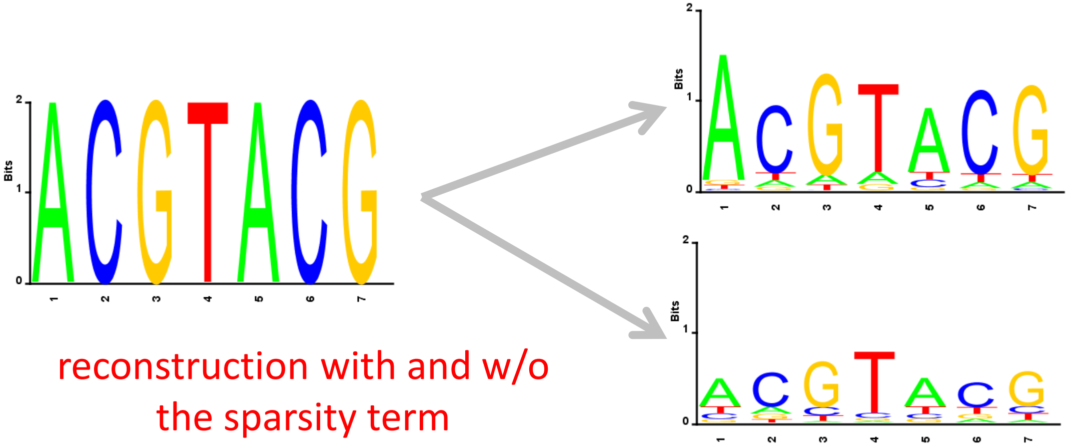
- To resolve sparsity of 1-hot encoding vectors, we devise a **new regularization** technique that incorporates prior knowledge on the sparsity.

Categorical Gradient

- For biological sequences, 1-hot encoding is widely used (Baldi & Brunak, 2001).
  - A, C, G, and T are encoded by 1000, 0100, 0010, and 0001, respectively.
  - In encoded binary vectors, 75% of the elements are zero.

- To resolve sparsity of 1-hot encoding vectors, we devise a **new regularization** technique that incorporates prior knowledge on the sparsity.

$$\min_{W,\mathbf{b},\mathbf{c}} \mathbf{E}\left[-\sum_{n=1}^{N} \log P(\mathbf{v}_n)\right] + \lambda_c \underbrace{\phi(\mathbf{v}_n),}_{\text{sparsity term}} \qquad \phi(\mathbf{v}) = \frac{1}{2}\sum_{i=1}^{m}\Big(\sum_{j=1}^{n_c} v^{(k)}_{n_c(i-1)+j} - 1\Big)^2$$



reconstruction with and w/o the sparsity term

Categorical Gradient

- For biological sequences, 1-hot encoding is widely used (Baldi & Brunak, 2001).
  - A, C, G, and T are encoded by 1000, 0100, 0010, and 0001, respectively.
  - In encoded binary vectors, 75% of the elements are zero.

- To resolve sparsity of 1-hot encoding vectors, we devise a **new regularization** technique that incorporates prior knowledge on the sparsity.

$$\min_{W,\mathbf{b},\mathbf{c}} \mathbf{E}\left[-\sum_{n=1}^{N} \log P(\mathbf{v}_n)\right] + \lambda_c \boxed{\phi(\mathbf{v}_n),} \qquad \phi(\mathbf{v}) = \frac{1}{2}\sum_{i=1}^{m}(\sum_{j=1}^{n_c} v^{(k)}_{n_c(i-1)+j} - 1)^2$$

sparsity term

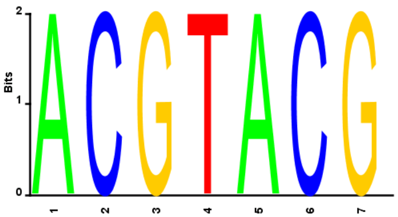
derived from the sparsity term

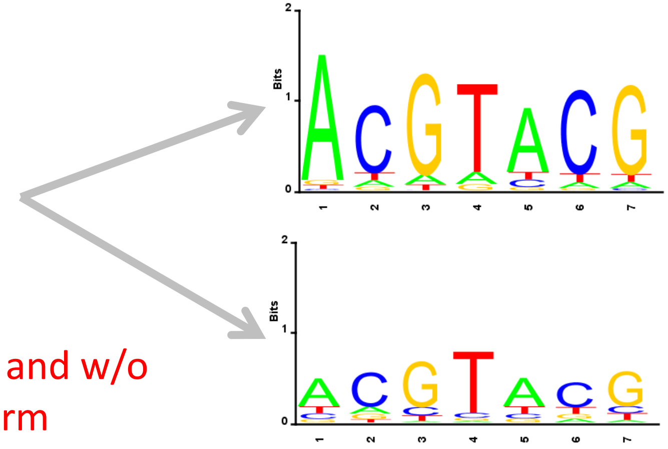
$$\frac{\partial L}{\partial W} \approx \text{Eq. } (2) + \boxed{\frac{1}{N}\sum_{n=1}^{N} f(\mathbf{v}_n^{(k)})\mathbf{h}_n^{(k-1)}} \tag{3}$$

$$\frac{\partial L}{\partial \mathbf{b}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{v}_n^{(0)} - \mathbf{v}_n^{(k)} + \boxed{f(\mathbf{v}_n^{(k)})}\right) \tag{4}$$

$$\frac{\partial L}{\partial \mathbf{c}} \approx \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{h}_n^{(0)} - \mathbf{h}_n^{(k)}\right) \tag{5}$$

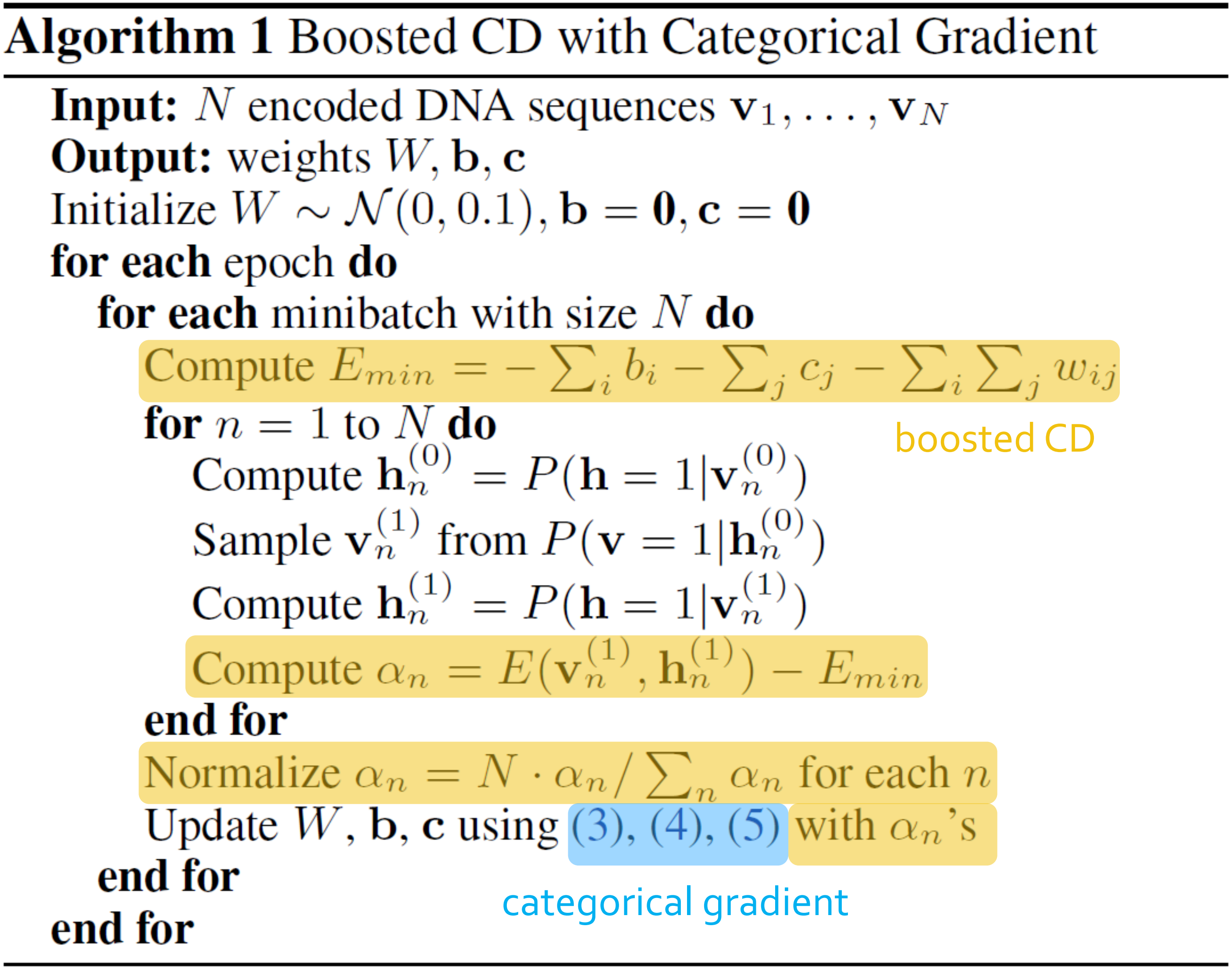
$$f(\mathbf{v}) = \mathbf{v} \circ (1-\mathbf{v}) \circ g(\mathbf{v}), \;\; g(\mathbf{v})_i = \sum_{j=1}^{n_c} v_{n_c[\frac{i-1}{n_c}]+j} - 1$$





reconstruction with and w/o the sparsity term

Proposed Training Algorithm



Algorithm 1 Boosted CD with Categorical Gradient

```
Input: $N$ encoded DNA sequences $\mathbf{v}_1, \ldots, \mathbf{v}_N$
Output: weights $W$, $\mathbf{b}$, $\mathbf{c}$
Initialize $W \sim \mathcal{N}(0, 0.1)$, $\mathbf{b} = \mathbf{0}$, $\mathbf{c} = \mathbf{0}$
for each epoch do
    for each minibatch with size $N$ do
        Compute $E_{min} = -\sum_i b_i - \sum_j c_j - \sum_i \sum_j w_{ij}$
        for $n = 1$ to $N$ do
            Compute $\mathbf{h}_n^{(0)} = P(\mathbf{h} = 1 | \mathbf{v}_n^{(0)})$
            Sample $\mathbf{v}_n^{(1)}$ from $P(\mathbf{v} = 1 | \mathbf{h}_n^{(0)})$
            Compute $\mathbf{h}_n^{(1)} = P(\mathbf{h} = 1 | \mathbf{v}_n^{(1)})$
            Compute $\alpha_n = E(\mathbf{v}_n^{(1)}, \mathbf{h}_n^{(1)}) - E_{min}$
        end for
        Normalize $\alpha_n = N \cdot \alpha_n / \sum_n \alpha_n$ for each $n$
        Update $W$, $\mathbf{b}$, $\mathbf{c}$ using (3), (4), (5) with $\alpha_n$'s
    end for
end for
```

boosted CD

categorical gradient

Outline

- Motivation
- Preliminary
- Boosted contrastive divergence
- Categorical restricted Boltzmann machine
- Experiment results
- Conclusion

Results

Effects of categorical gradient

Effects of boosting

Effects on the splicing prediction

- Data preparation:
  - **Real human DNA** sequences with known boundary information.

Results

Effects of categorical gradient
Effects of boosting
Effects on the splicing prediction

- Data preparation:
  - **Real human DNA** sequences with known boundary information.

CGTAGCAGCGATACGTACCGATCGTCACTATCATCGAGGTACGAGAGATCGATCGGCAACG

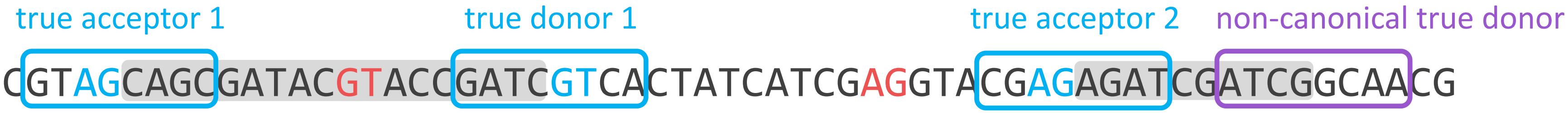
Results

Effects of categorical gradient

Effects of boosting

Effects on the splicing prediction

- Data preparation:
  - **Real human DNA** sequences with known boundary information.



true acceptor 1

true donor 1

true acceptor 2

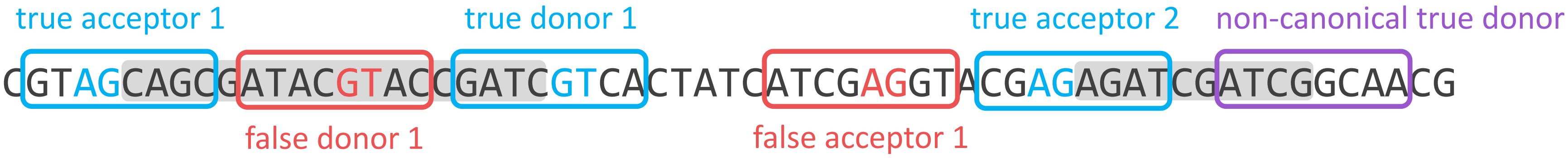
non-canonical true donor

CGTAGCAGCGATACGTACCGATCGTCACTATCATCGAGGTACGAGAGATCGATCGGCAACG

Results

Effects of categorical gradient
Effects of boosting
Effects on the splicing prediction

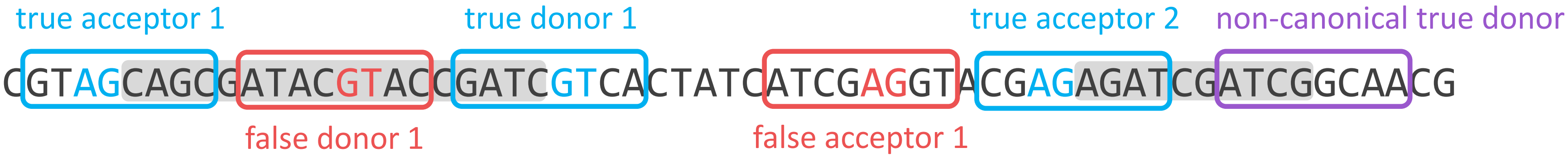
- Data preparation:
  - **Real human DNA** sequences with known boundary information.



Results

Effects of categorical gradient
Effects of boosting
Effects on the splicing prediction

- Data preparation:
  - **Real human DNA** sequences with known boundary information.



  - GWH dataset: **2-class** (boundary or not).
  - UCSC dataset: **3-class** (acceptor, donor, or non-boundary).

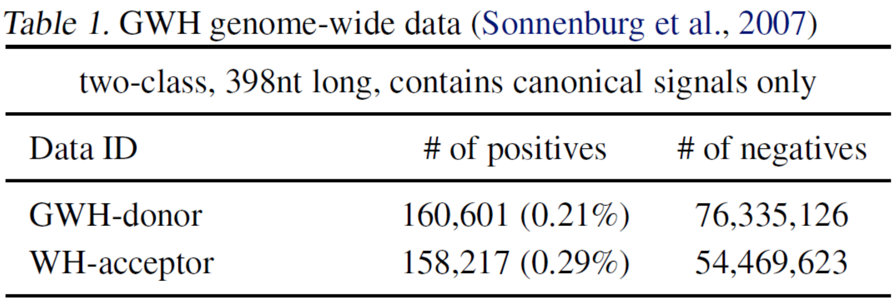


Table 1. GWH genome-wide data (Sonnenburg et al., 2007)

two-class, 398nt long, contains canonical signals only		
Data ID	# of positives	# of negatives
GWH-donor	160,601 (0.21%)	76,335,126
WH-acceptor	158,217 (0.29%)	54,469,623

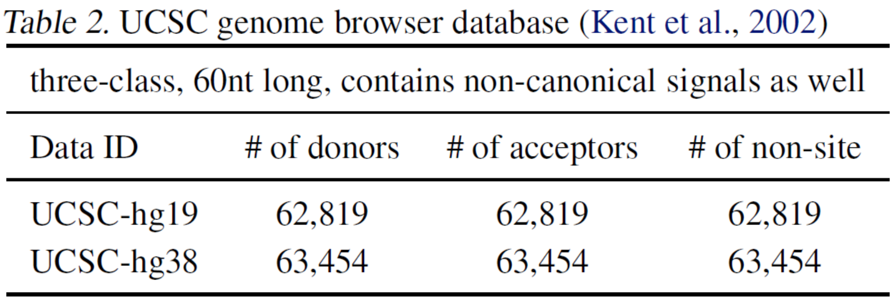
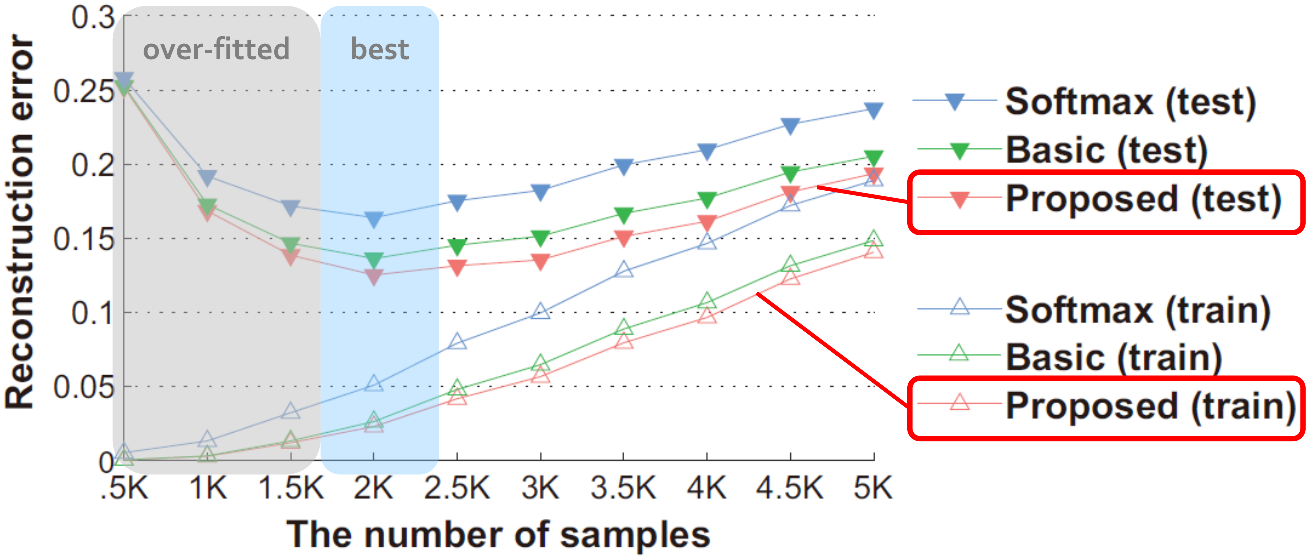


Table 2. UCSC genome browser database (Kent et al., 2002)

three-class, 60nt long, contains non-canonical signals as well			
Data ID	# of donors	# of acceptors	# of non-site
UCSC-hg19	62,819	62,819	62,819
UCSC-hg38	63,454	63,454	63,454

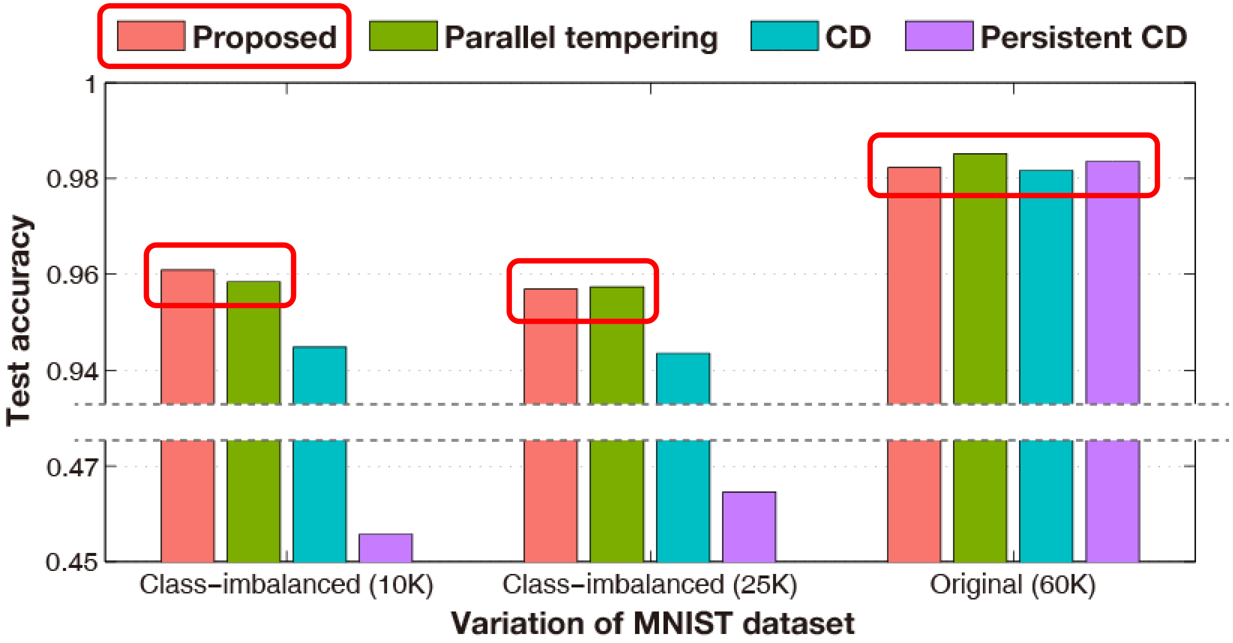
Results: Effects of Categorical Gradient



Data: chromosome 19 in GWH-donor
Sequence Length: 200nt
(800 dimension)
# of iterations: 500
Learning rate: 0.1
L2-decay: 0.001

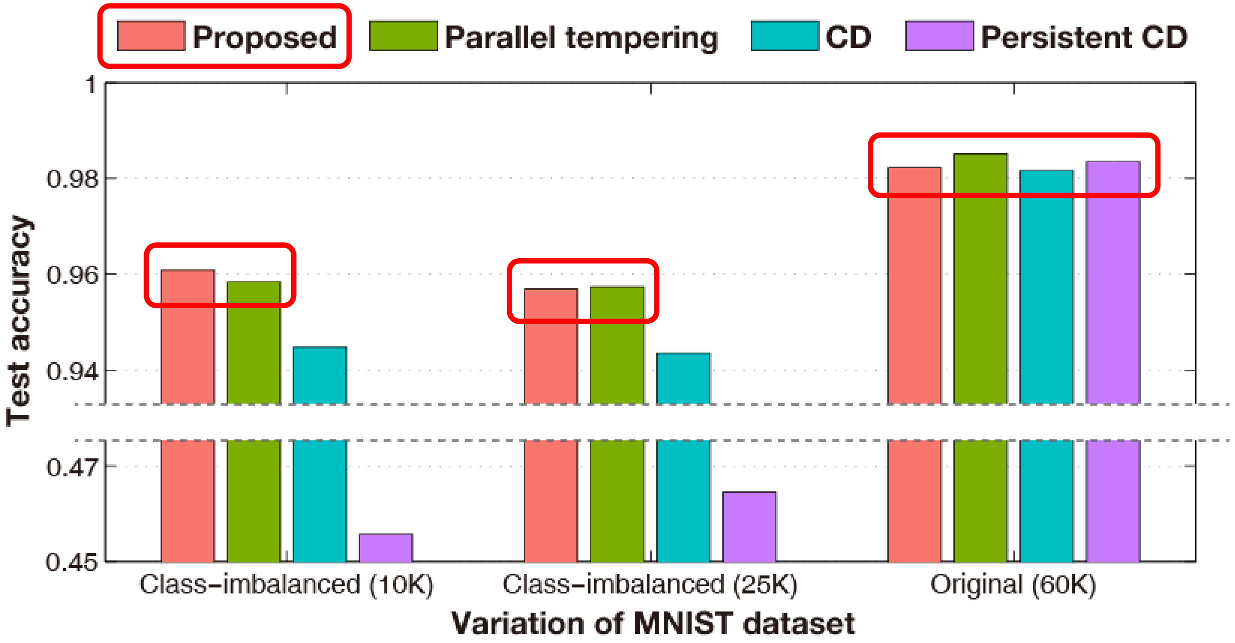
- The proposed method shows the **best** performance in terms of **reconstruction error** for both training and testing.
- Compare to the softmax approach, the proposed regularized RBM succeeds in **achieving lower error by slightly sacrificing the probability sum constraint**.

Results: Effects of Boosting



- For **simulating a class-imbalance** situation
  - we randomly dropped samples with different drop rates for different classes.

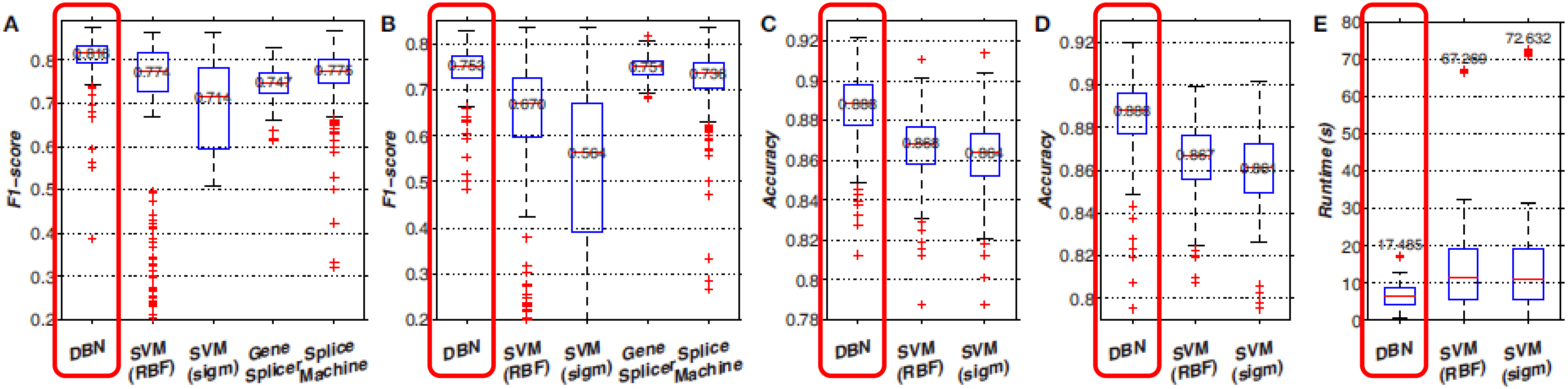
Results: Effects of Boosting



- For **simulating a class-imbalance** situation
  - we randomly dropped samples with different drop rates for different classes.

	Description	Training cost	Noise handling	Class-imbalance handling
CD (Hinton, Neural Comp. 2002)	Standard and widely used	-	-	-
Persistent CD (Tieleman, ICML 2008)	Use of a single Markov chain	-	☺	-
Parallel tempering (Cho et al., IJCNN 2010)	Simultaneous Markov chains generation	☹	☺	☺
Proposed boosted CD	**Reweighting samples**	**-**	☺	☺

Results: Improved Performance and Robustness

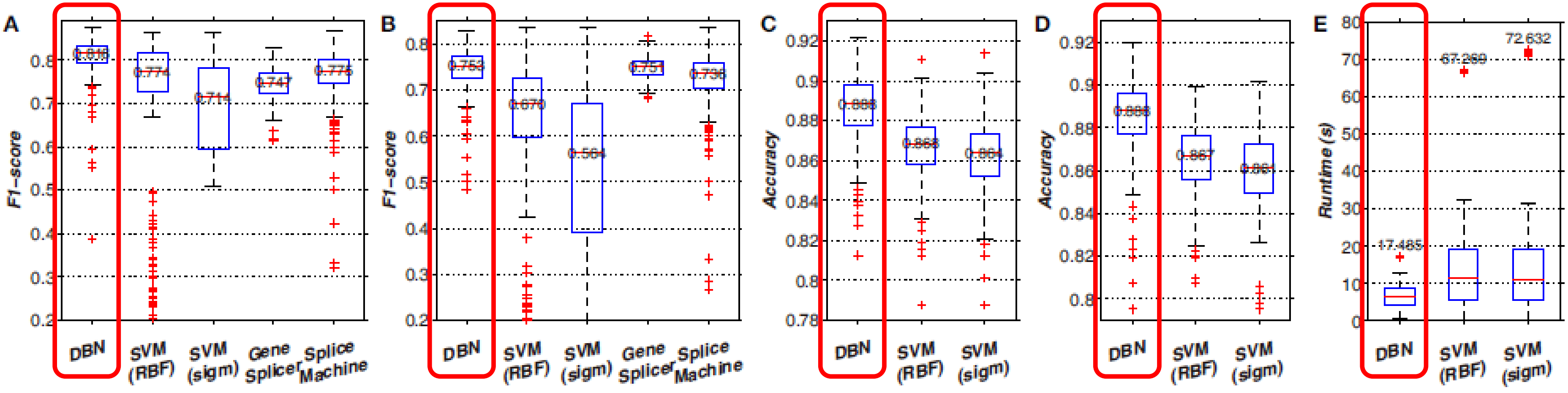


2-class classification performance

3-class classification

Runtime

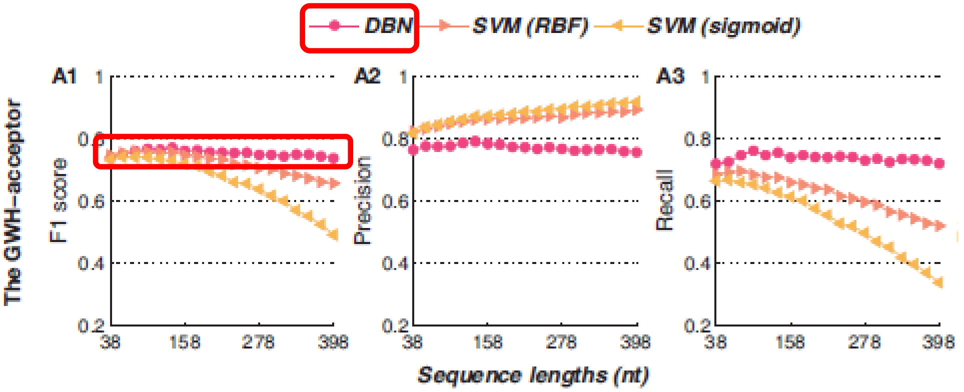
Results: Improved Performance and Robustness



2-class classification performance

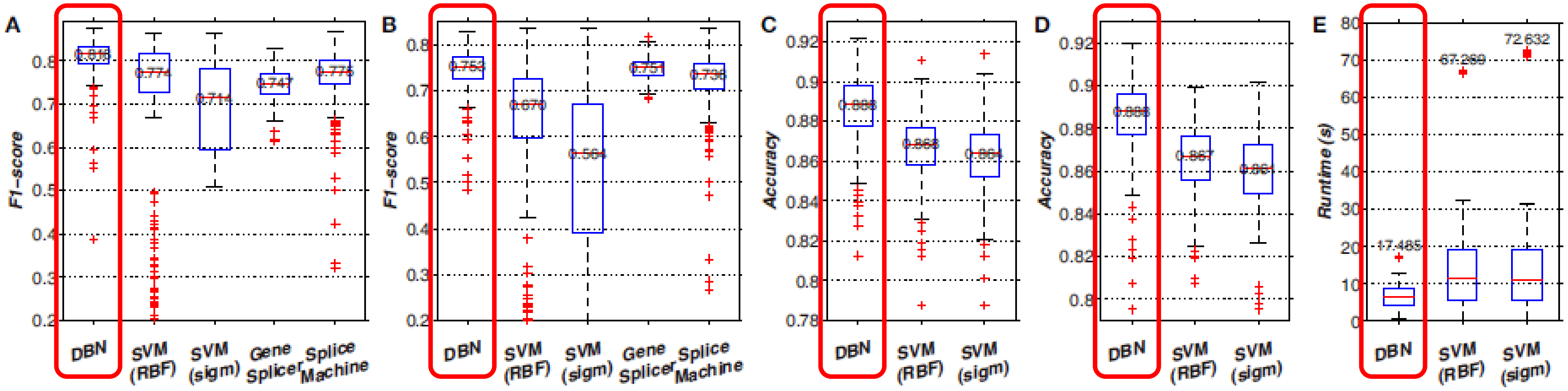
3-class classification

Runtime



Insensitivity to sequence lengths

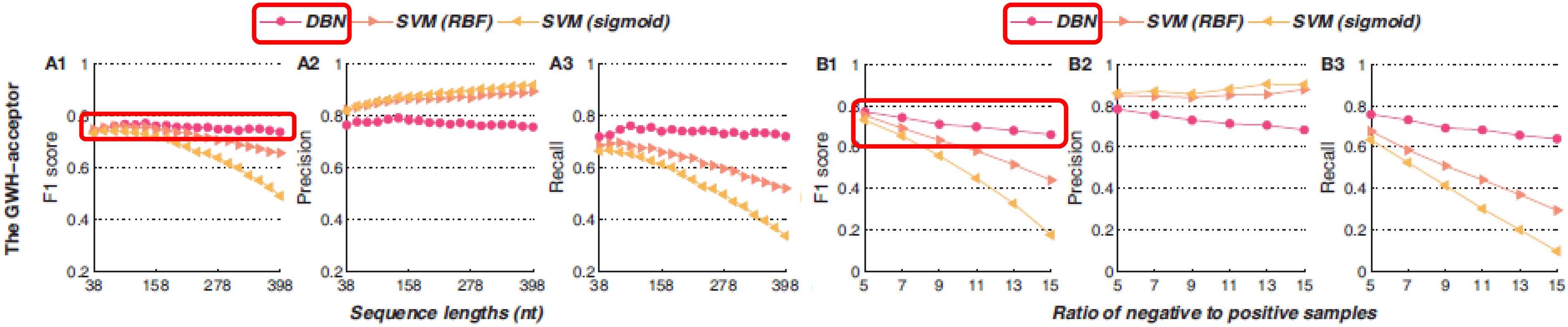
Results: Improved Performance and Robustness



2-class classification performance

3-class classification

Runtime

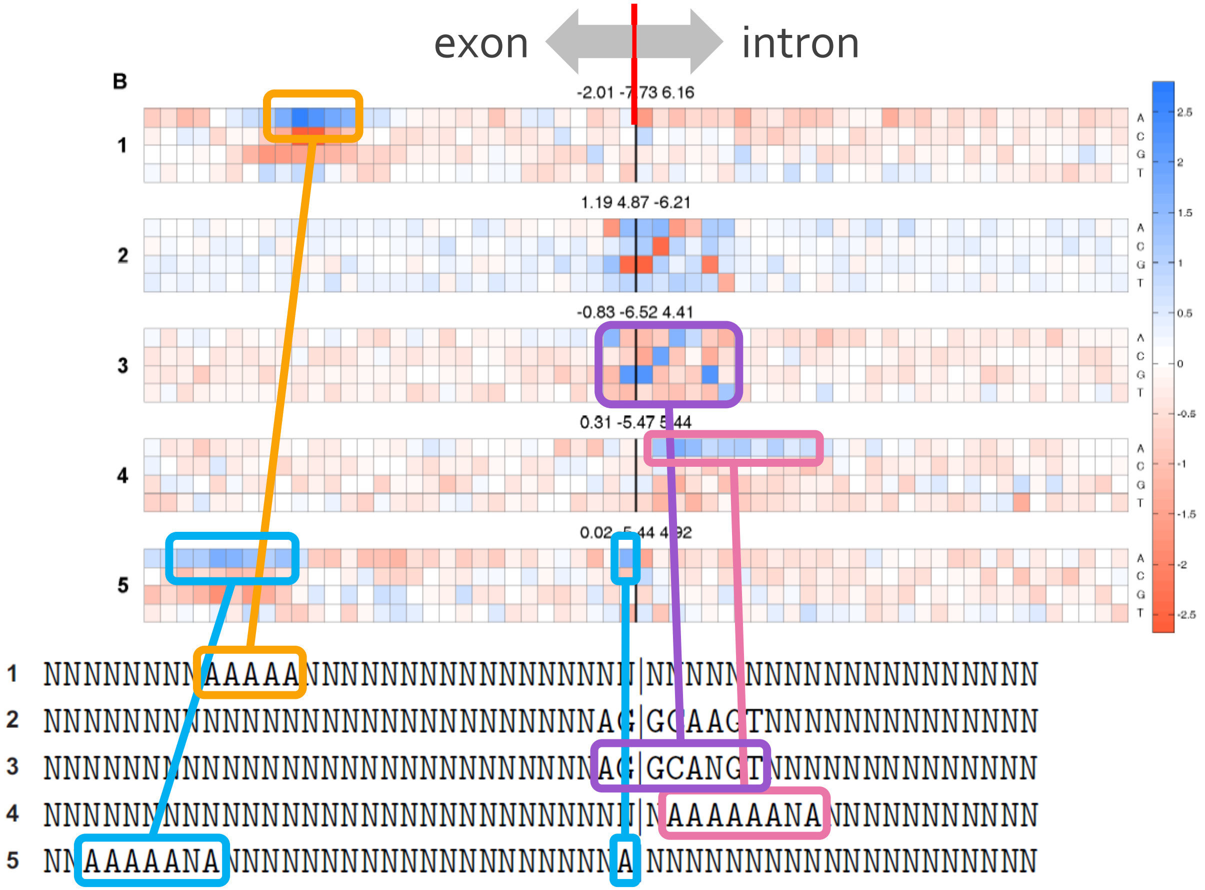


Insensitivity to sequence lengths

Robustness to negative samples

Results: Identification of Non-Canonical Splice Sites

- **(Important biological finding)** non-canonical splicing can arise if:
  - Introns contain GCA or NAA sequences at their boundaries.
  - Exons include contiguous A's around the boundaries.

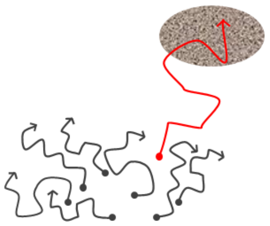


We used 162,951 examples excluding canonical splice sites.

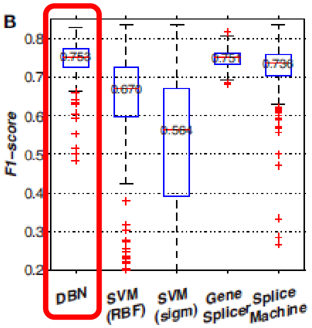
Conclusion

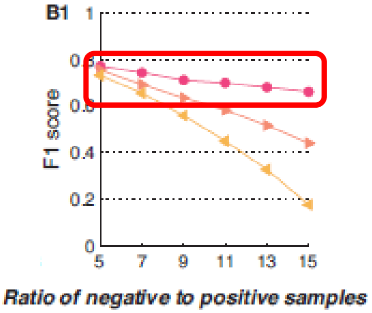
- We proposed a **new RBM training method called boosted CD** with categorical gradients that improves conventional CD for class-imbalanced data.
  - **Significant boosts in splicing prediction** in terms of accuracy and runtime.
  - Increased **robustness** to high-dimensional **class-imbalanced** data.

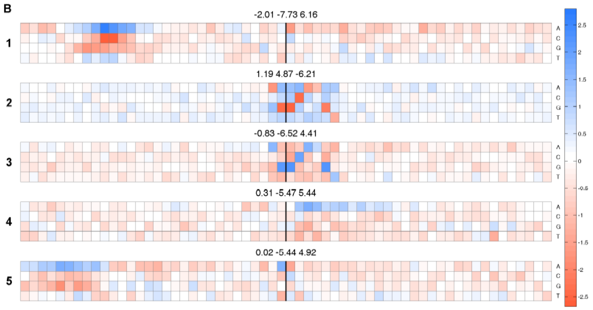
- The proposed scheme shows the ability to detect subtle **non-canonical splicing signals** that often could not be identified by traditional methods.
  - Future work: additional validation using various class-imbalance datasets.



$$\phi(\mathbf{v}) = \frac{1}{2}\sum_{i=1}^{m}(\sum_{j=1}^{n_c} v^{(k)}_{n_c(i-1)+j} - 1)^2$$







Acknowledgements

- Our lab members

June 2, 2015

- Financial supports

NRF National Research Foundation of Korea

BrainKorea21 PLUS

IITP 정보통신기술진흥센터 Institute for Information & communications Technology Promotion

SAMSUNG ELECTRONICS

- ICML 2015 travel scholarship

ICML Lille

International Conference on Machine Learning

Acknowledgements

- Our lab members

June 2, 2015

- Financial supports

NRF National Research Foundation of Korea

BrainKorea21 PLUS

IITP 정보통신기술진흥센터 Institute for Information & communications Technology Promotion

SAMSUNG ELECTRONICS

- ICML 2015 travel scholarship

ICML Lille

International Conference on Machine Learning

Backup: Comparison with Recurrent Neural Networks (RNNs)

To be placed

- The proposed DBN showed xx% higher performance in terms of the F1-score.
- RNN is appropriate for sequence modeling. However, splicing signals are often too far from the boundaries and hard to maintain splicing information.