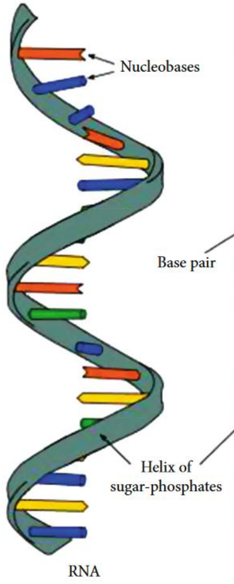
DICTIONARY LEARNING ON DNA DATA ANALYSIS

彭志远

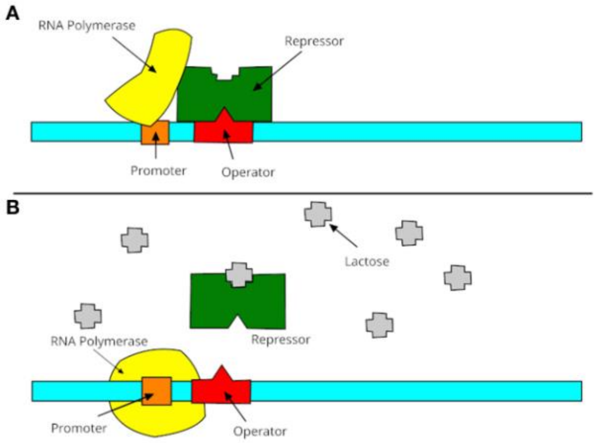
2023/11/22

Why DNA ANALYSIS is so important?

Bases Permutation



Chemistry Property

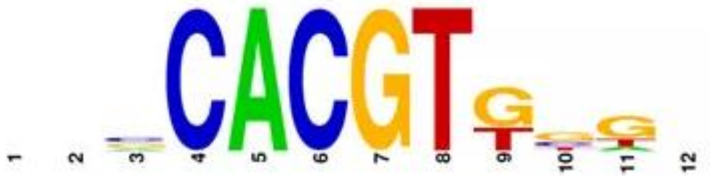


DECIDE

Creature Features
Appearance
Evolution
......

Diseases
Disability
Mental distort

Bases Composition



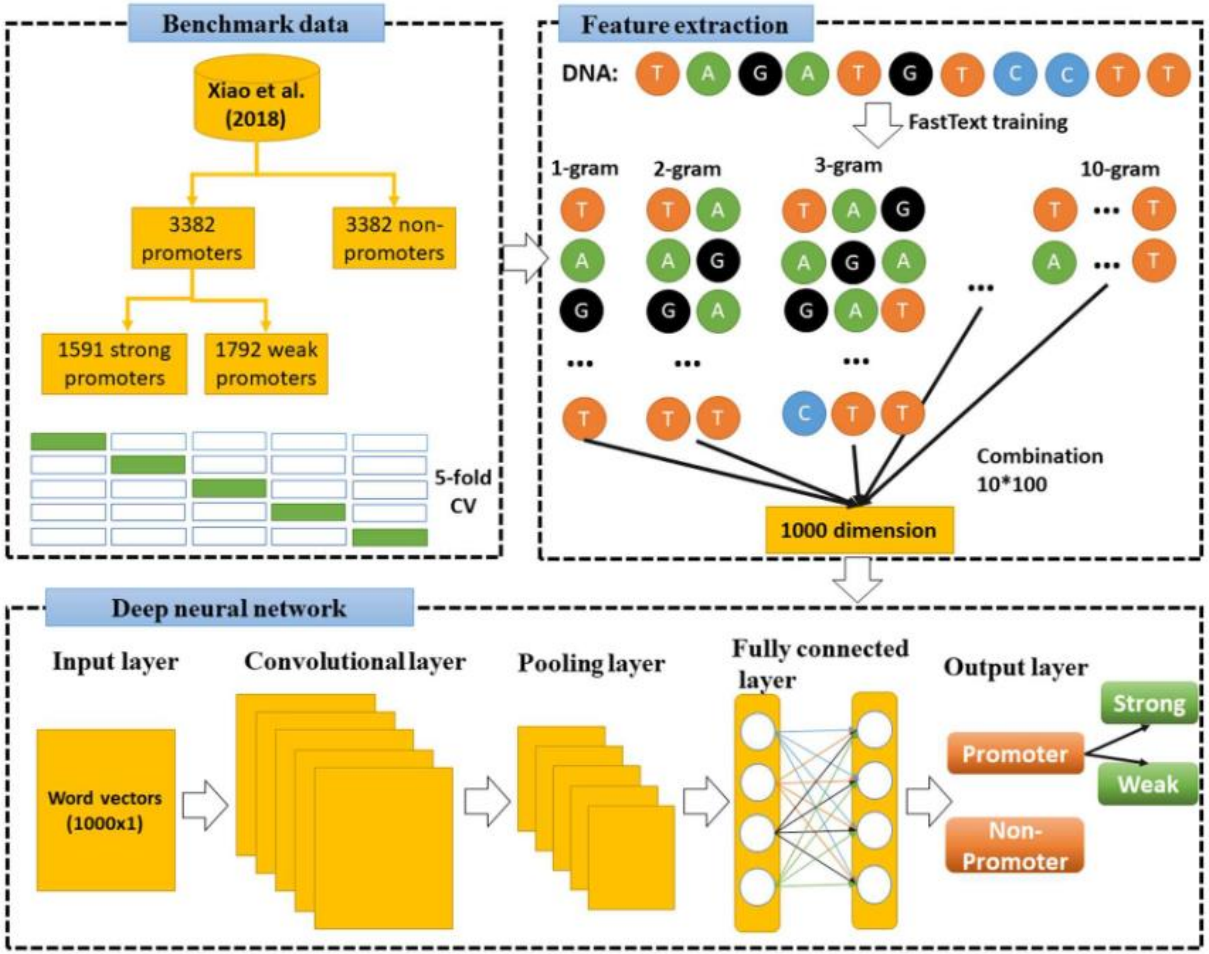
How to quantitate the relationship between the DNA structure and those disasters?

Literature Review

The detection of promoters is an essential problem in genome research for precaution on genetics and human diseases.

The idea is based upon the natural language processing (NLP) field which classifies the text/sentence into its appropriate scenario.

Therefore, we would like to apply it to bioinformatics to interpret the hidden information of DNA sequences (represented by promoters)



Classifying Promoters by Interpreting the Hidden Information of DNA Sequences via Deep Learning and Combination of Continuous FastText N-Grams
2019(Frontiers in Bioengineering and Biotechnology)

DICTIONARY LEARNING

Dictionary learning is a way to find a better sparse mapping matrix by the use of training data.

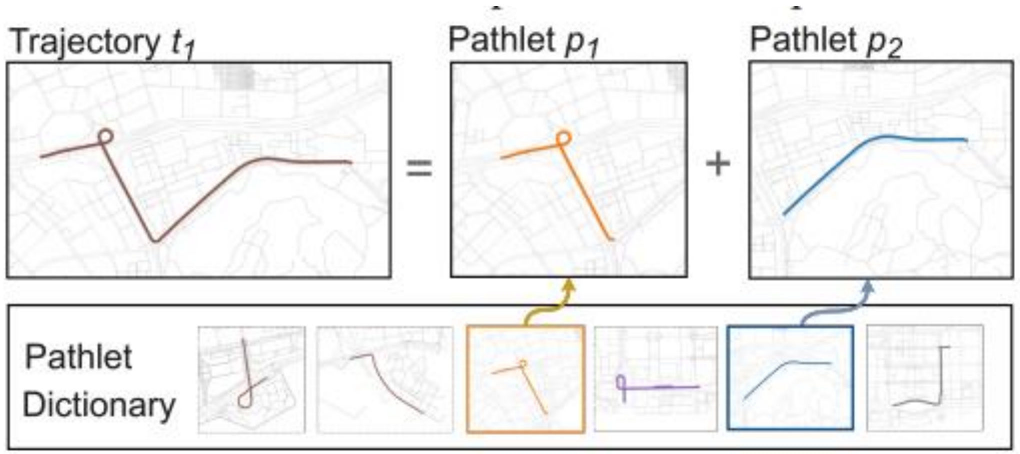
It aims at extracting the essence (low dimensional features) from data for weakening the noises.

$$x\text{: } original\ input\ data$$
$$D\text{: } mapping\ matrix\ (Dictionary)$$
$$\alpha \text{ sparse represention}$$

$$\min_{D,\alpha_i} \sum_{i=1}^{\mathrm{N}} \|\mathrm{x}_i - D\alpha_i\|^2 + \lambda \sum_{i=1}^{N} [\![\alpha_i]\!]_1$$

We can update the D and α alternatively by fixing one and changing the other

Dictionary Learning: Pathlet



$$\min_{R_{i,j}\in\{0,1\}} \quad C(R) = \sum_{i=1}^{|\overline{P}|} max(R_{i,:}) + \lambda * \sum_{i=1}^{|\overline{P}|}\sum_{j=1}^{|T|} |R_{i,j}|$$

$$s.t. DR = M$$

D indicates trajectory j encompasses edge i

*R indicates tra*jectory i takes pathlet j

M indicates edges as a part of pathlet j

Sparsity (Dictionary size)

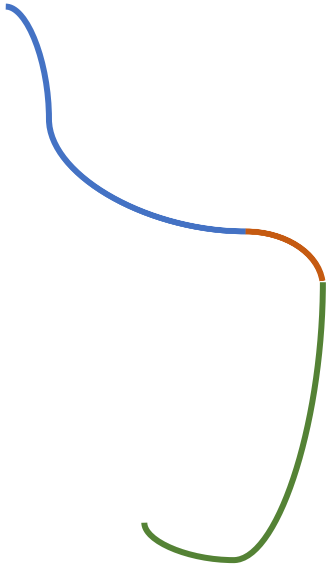
Informative (Representation cost)

Accuracy (Reconstruction loss)

Seems a little Different?

In Roadmap,

Combination of pathlet can ignore the Order, that is, the graph is undirected.



$$ABC = ACB$$

Only focus the occurence

But in DNA sequence,

Combination of segments (k-mer) should take order into consideration, for its biochemical information, that is, the graph is directed.

AATGC

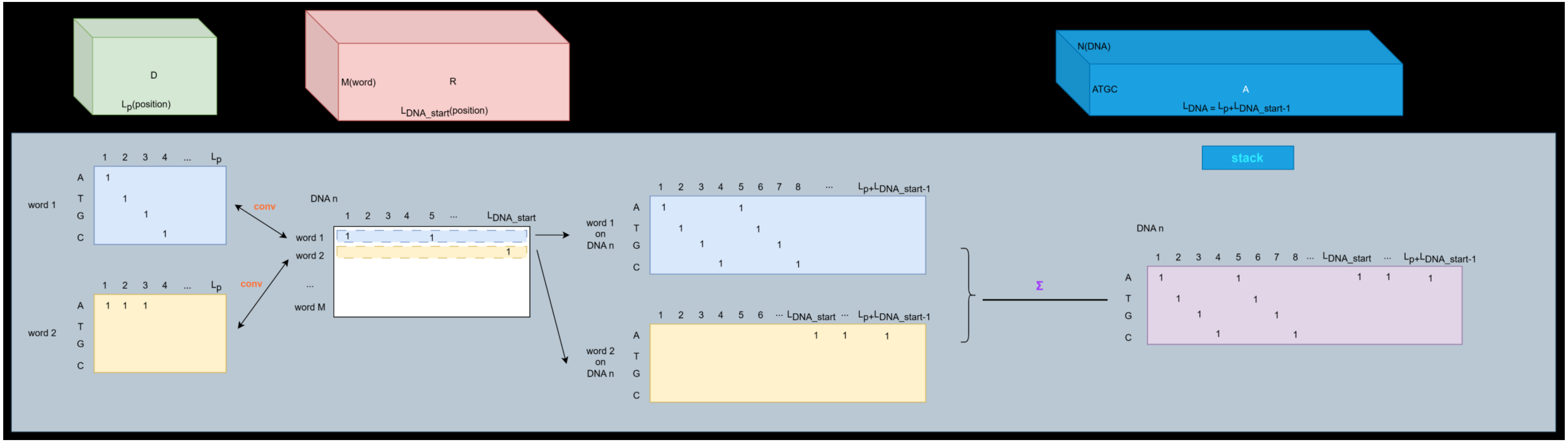
$\neq$

ATAGC

Not only cares about the occurrence
But the permutation

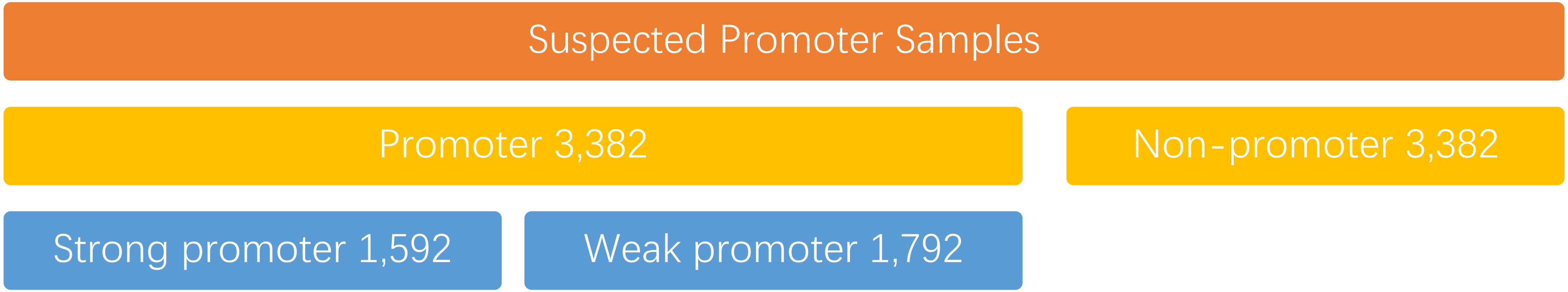
How to add the positional information?

Generalized into 3D scenario



1. Add a dimension to each matrix representing the position of the corresponding element
2. Alter the matrix multiplication into one-dimensional convolution

Benchmark dataset



Suspected Promoter Samples

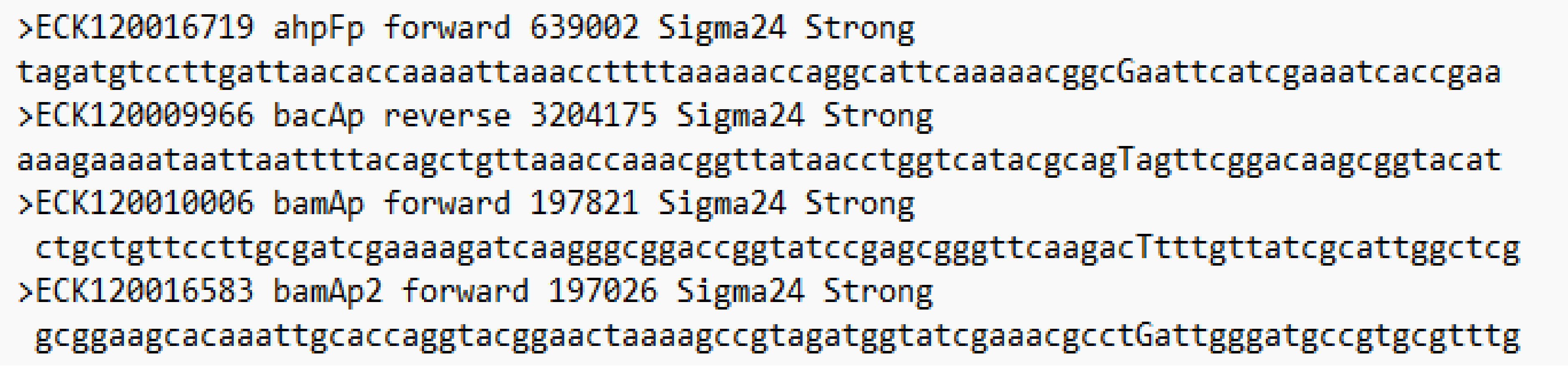
Promoter 3,382

Non-promoter 3,382

Strong promoter 1,592

Weak promoter 1,792

Length: 81
Base: ATGC



```
>ECK120016719 ahpFp forward 639002 Sigma24 Strong
tagatgtccttgattaacaccaaaattaaacctttttaaaaaccaggcattcaaaaacggcGaattcatcgaaatcaccgaa
>ECK120009966 bacAp reverse 3204175 Sigma24 Strong
aaagaaaataattaattttacagctgttaaaccaaacggttataacctggtcatacgcagTagttcggacaagcggtacat
>ECK120010006 bamAp forward 197821 Sigma24 Strong
 ctgctgttccttgcgatcgaaaagatcaagggcggaccggtatccgagcgggttcaagacTtttgttatcgcattggctcg
>ECK120016583 bamAp2 forward 197026 Sigma24 Strong
 gcggaagcacaaattgcaccaggtacggaactaaaagccgtagatggtatcgaaacgcctGattgggatgccgtgcgtttg
```

Experiment Set-up

$$\lambda \; = \; 0.1$$

$$Learning\ Rate \; = \; 0.1$$

$$Epoch\ Number \; = \; 100$$

$$K_{mer}(length\ of\ \ DNA\ segments) = \; 1{\sim}5$$

$$Number\ of\ \ K\_mer \; = \; 1364$$

$$DNA\ Number \; = \; \; 4500$$

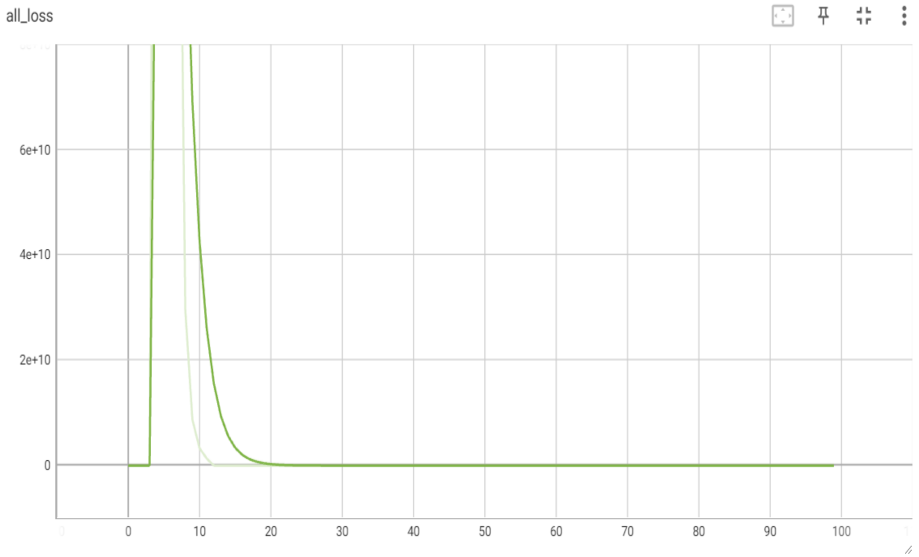
$$DNA\ length \; = \; 81$$

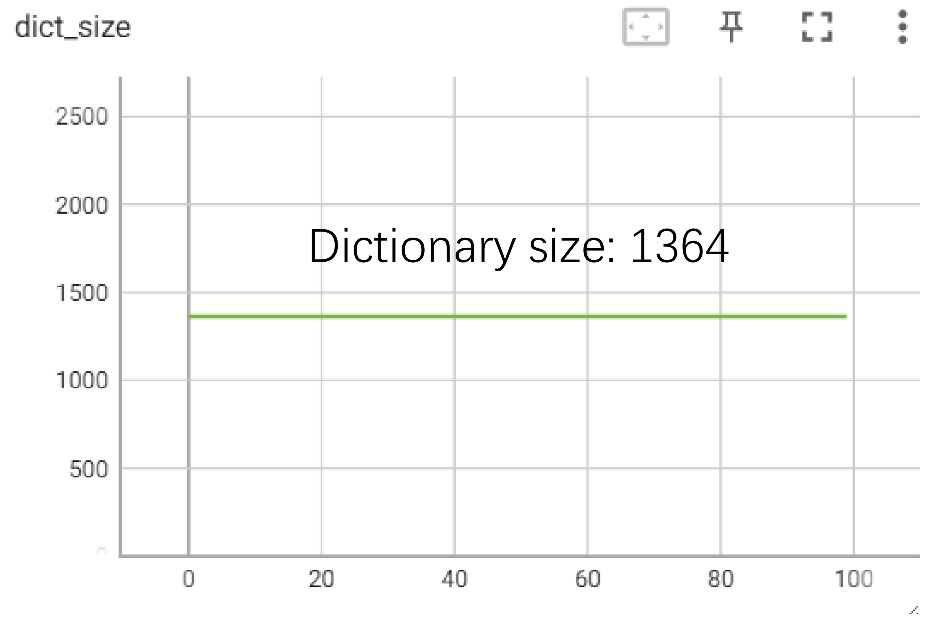
$$R\colon \; DNA\ Number \; * \; Number\ of\ \ K\_mer \; * \; DNA\ length$$

$$W\colon \; Number\ of\ \ K\_mer \; * \; 4 \; * \; MAX(length\ of\ \ K\_mer)$$

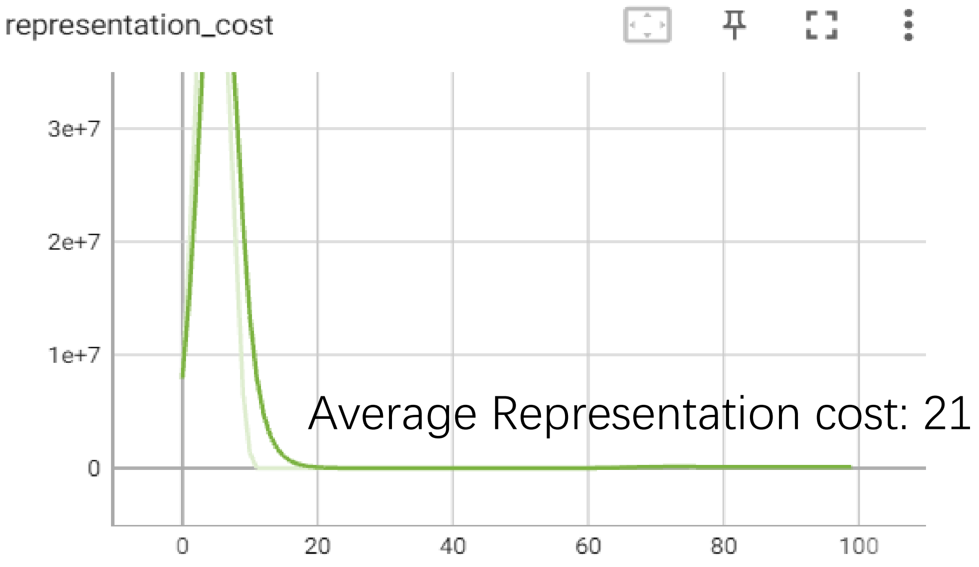
$$D\colon \; DNA\ Number \; * \; 4 \; * \; (l\_dna + l\_word - 1)$$

Numerical result





Dictionary size: 1364



Average Representation cost: 21

Downstream task: promoter classification

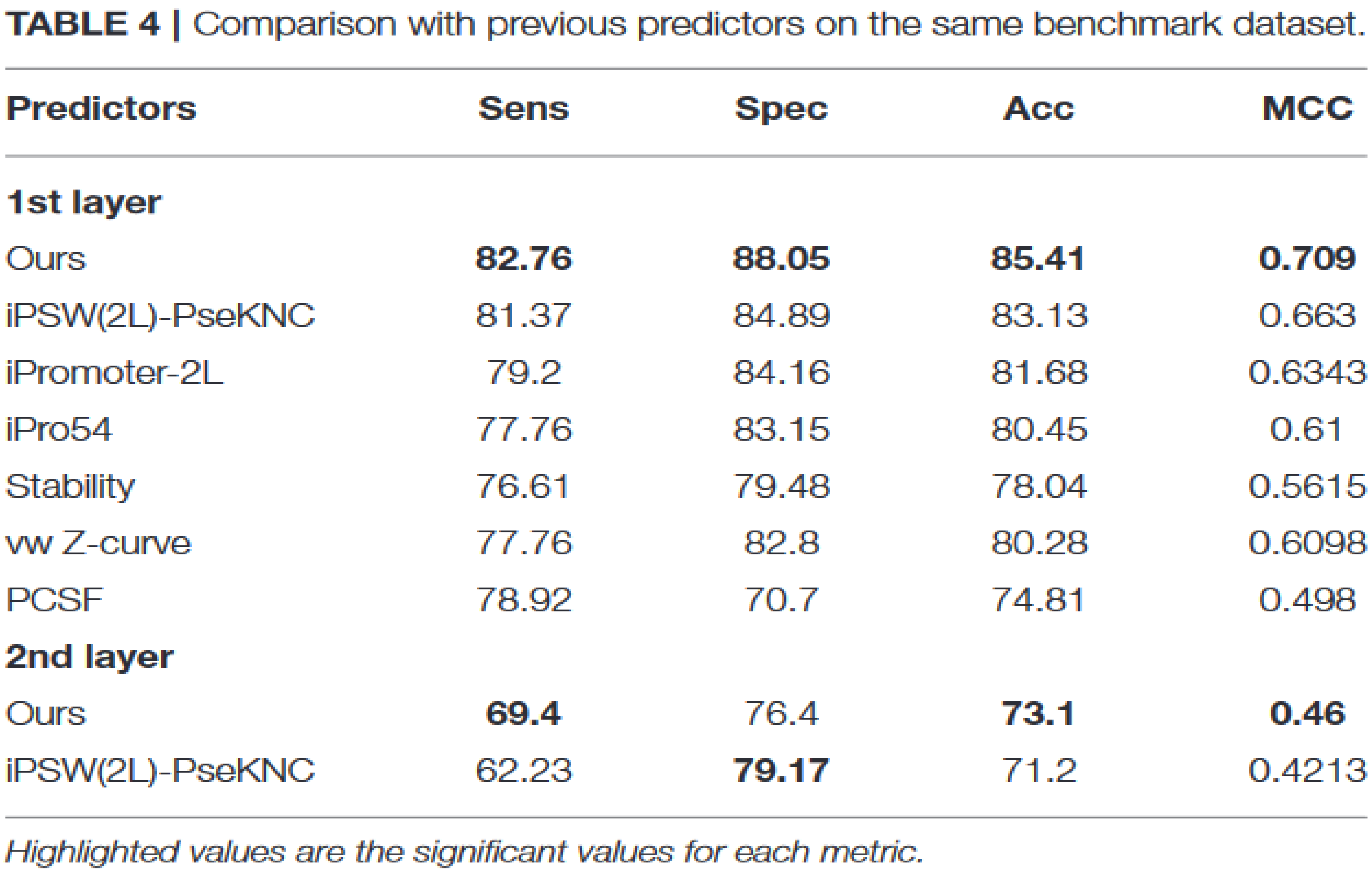


TABLE 4 | Comparison with previous predictors on the same benchmark dataset.

Predictors	Sens	Spec	Acc	MCC
1st layer				
Ours	**82.76**	**88.05**	**85.41**	**0.709**
iPSW(2L)-PseKNC	81.37	84.89	83.13	0.663
iPromoter-2L	79.2	84.16	81.68	0.6343
iPro54	77.76	83.15	80.45	0.61
Stability	76.61	79.48	78.04	0.5615
vw Z-curve	77.76	82.8	80.28	0.6098
PCSF	78.92	70.7	74.81	0.498
2nd layer				
Ours	**69.4**	76.4	**73.1**	**0.46**
iPSW(2L)-PseKNC	62.23	**79.17**	71.2	0.4213

Highlighted values are the significant values for each metric.

1st layer	86.4	76.7	83.2	0.63
2nd layer	79.3	78.1	78.6	0.58
	真正例率	真负例率	准确率	

Challenge

- How to reduce the space computation complexity when using cuda?
- How to quickly get the convolution result between two sparse matrices?