Application of Word2Vec to Represent Biological Sequences

Li Ka Shing Faculty of Medicine, HKU

Xu Hang

PhD. Candidate

2/4/2018

Content

1. Background

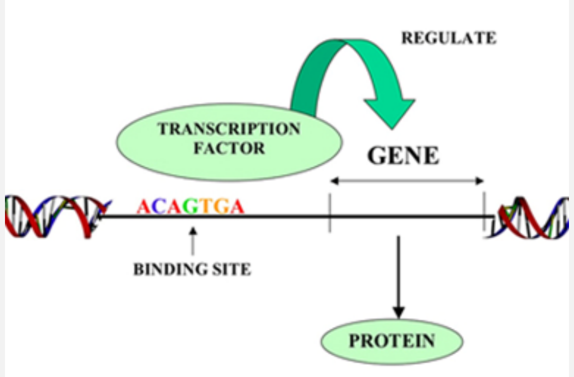
2. Principle of Word2Vec

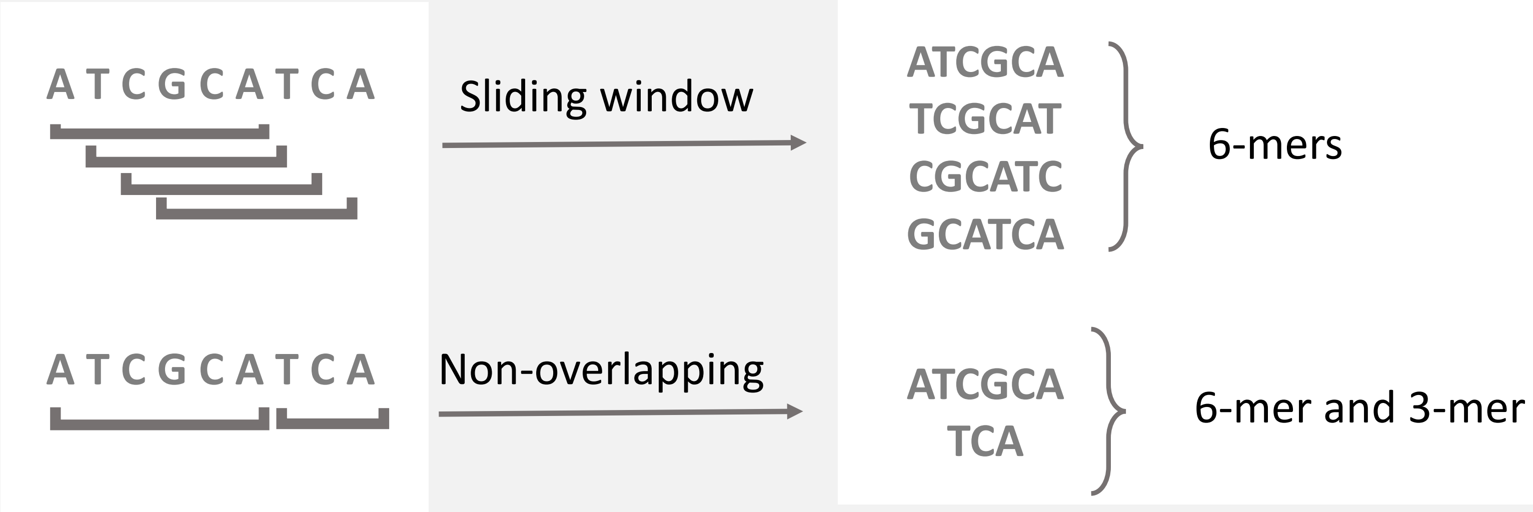
3. Pipeline of dna2vec

4. Performance Evaluations and Discussions.

Background

1. Biological Problems
   - Long DNA sequences are usually investigated
   - K-mer representation plays important role in splitting DNA sequences





A T C G C A T C A

Sliding window

ATCGCA
TCGCAT
CGCATC
GCATCA

6-mers

A T C G C A T C A

Non-overlapping

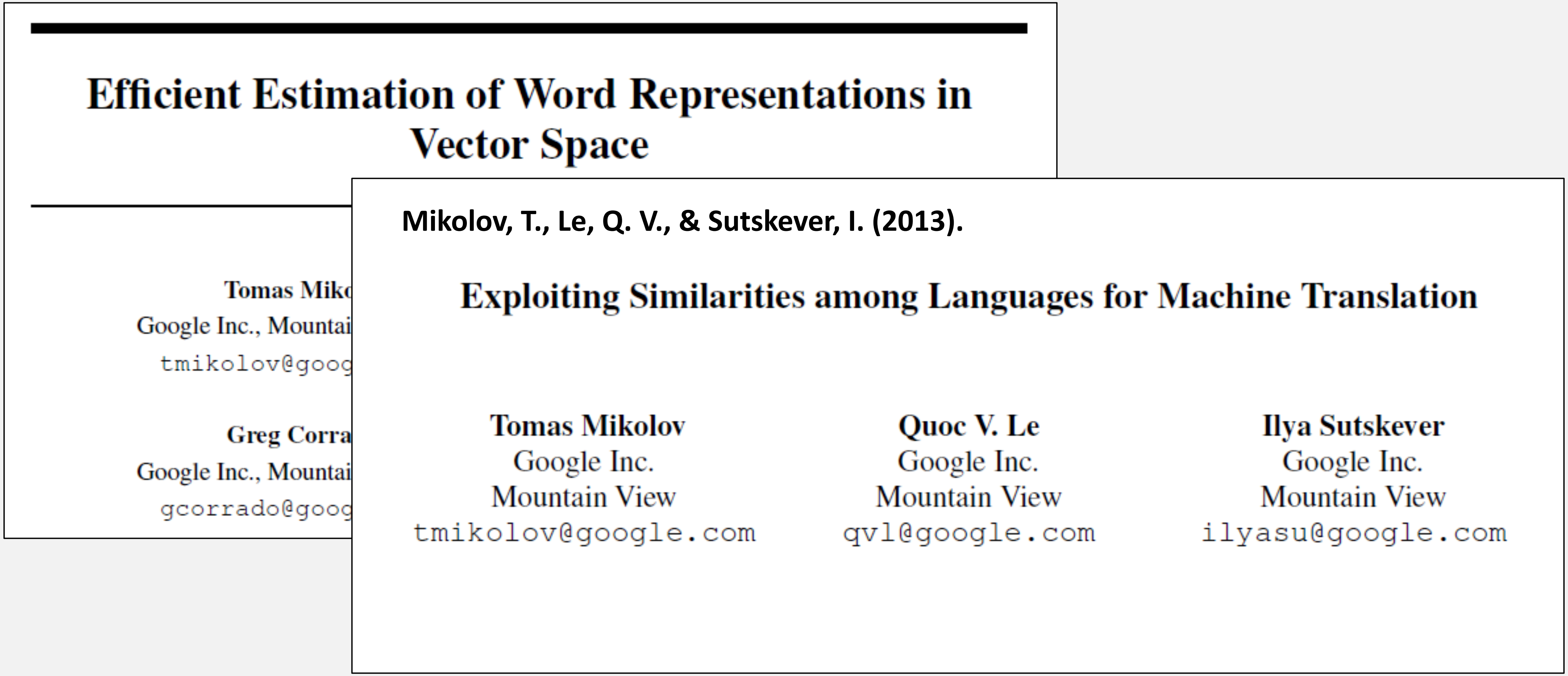
ATCGCA
TCA

6-mer and 3-mer

2. Encoding for k-mer: one-hot vector
   - Simple to understand
   - High dimension: $4^6 = 4096$
   - The distance between all paired vectors is equivalent

Word2Vec

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013).



Efficient Estimation of Word Representations in Vector Space

Tomas Miko
Google Inc., Mountai
tmikolov@goog

Greg Corra
Google Inc., Mountai
gcorrado@goog

Mikolov, T., Le, Q. V., & Sutskever, I. (2013).

Exploiting Similarities among Languages for Machine Translation

Tomas Mikolov
Google Inc.
Mountain View
tmikolov@google.com

Quoc V. Le
Google Inc.
Mountain View
qvl@google.com

Ilya Sutskever
Google Inc.
Mountain View
ilyasu@google.com

Word2Vec

1. Model description
   - The vocabulary size is V.
   - Input layer: $\{x_1, \dots, x_V\}$
   - Hidden layer: $h_{N\times N}$
   - Output layer: $\{y_1, \dots, y_V\}$
   - Two matrix: $W_{V\times N}$, $W_{N\times V}$
2. Optimization Target
   - Given one context word x, the model can properly predict the word y
3. Important intermediate product
   - The row vector in $W_{V\times N}$ can be used as word vector

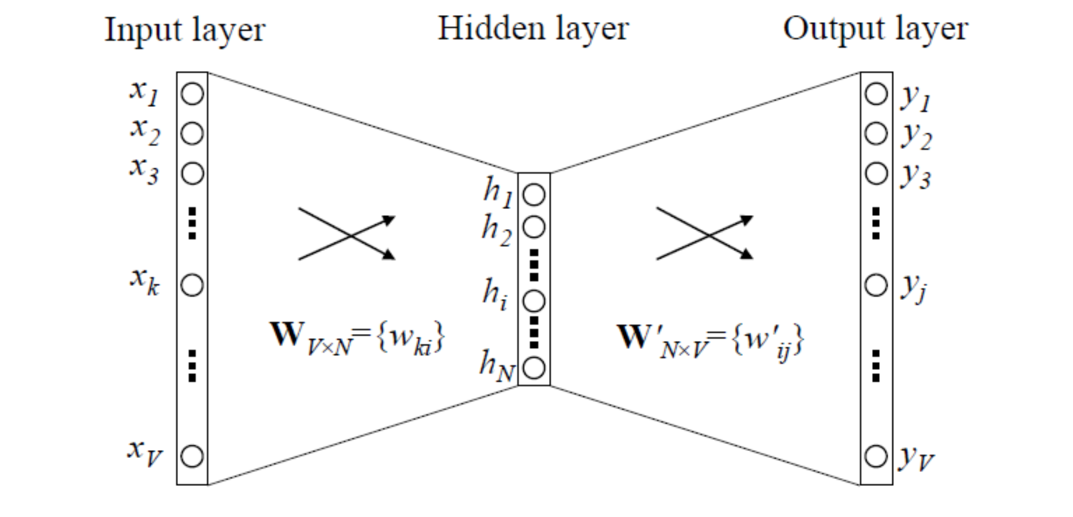
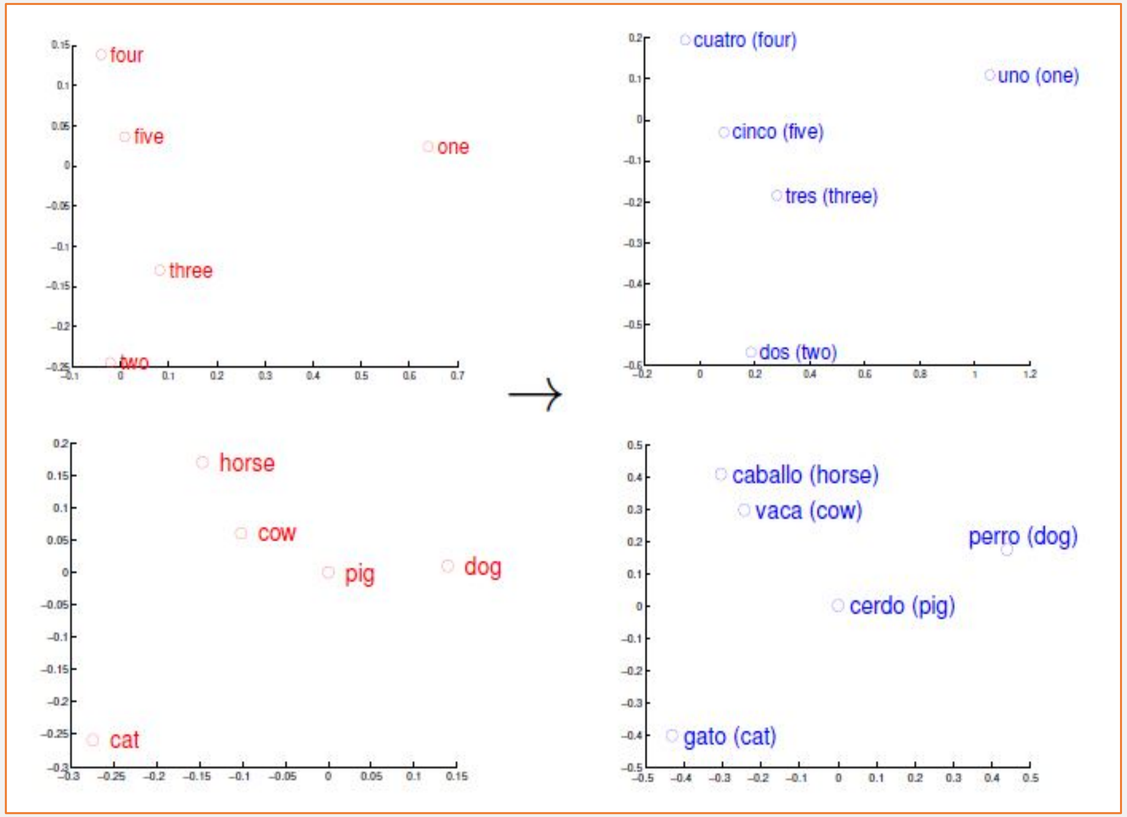


Figure 1: A simple CBOW model with only one word in the context

Rong, X. (2014).

Word2Vec

3. Important intermediate product
   - The row vector in $W_{V \times N}$ can be used as word vector
4. Application of word vector in translation
   - English to Spanish
   - These concepts have similar geometric arrangements in both spaces



Mikolov, T., Le, Q. V., & Sutskever, I. (2013).

dna2vec

1. Analogy between DNA and Nature Language

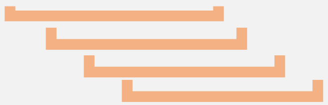
Nature Language	DNA
Words	K-mer
Sentences	DNA fragments
Corpus	Part or whole genome

2. Pipeline of training dna2vec
    1. Preparing corpus
        - Prepare a genome which contains long DNA contig (chromosome) (>1M)
        - Randomly select DNA fragments from contigs (<1k)
        - Use sliding-window or non-overlapping to split DNA fragments into k-mers
    2. Use gensim (python package) to train word2vec model with corpus

Different Strategies of Establishing Corpus

Word Embedding (same k)

1.Seq A T C G C A T C A



2.Word

ATCGCA
TCGCAT
CGCATC
GCATCA

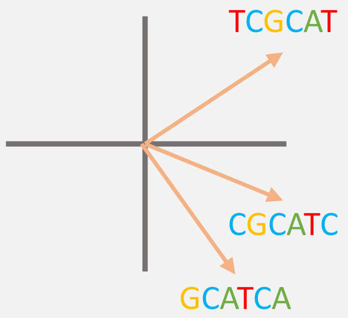
3. Sentence

ATCGCA | TCGCAT | CGCATC | GCATCA
This is a sentence

4. Corpus

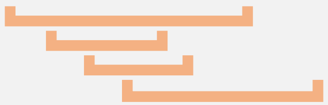
1). ATCGCA | TCGCAT | CGCATC | GCATCA
2). TCGCAT | CGCATC | GCATCG | CATCAG
3). CGCATC | GCATCG | CATCGC | ATCAGC
4). ...

5. dna2vec



Word Embedding (variable k)

1.Seq A T C G C A T C A



2.Word

ATCGCAT
TCGC
CGC
GCATCA

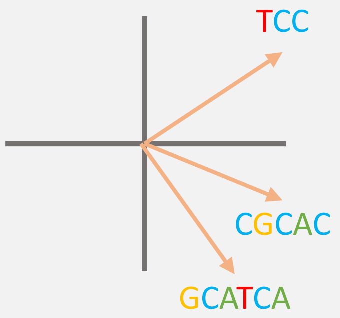
3. Sentence

ATCGCA | TCGC | CGC | GCATCA
This is a sentence

4. Corpus

1). ATCGCA | TCGC | CGCAT | GCATCA
2). TCGCAT | CGCA | GCATC | CATCAG
3). CGCA | GCATCG | CATCGC | ATCG
4). ...

5. dna2vec



Methods to Evaluate dna2vec

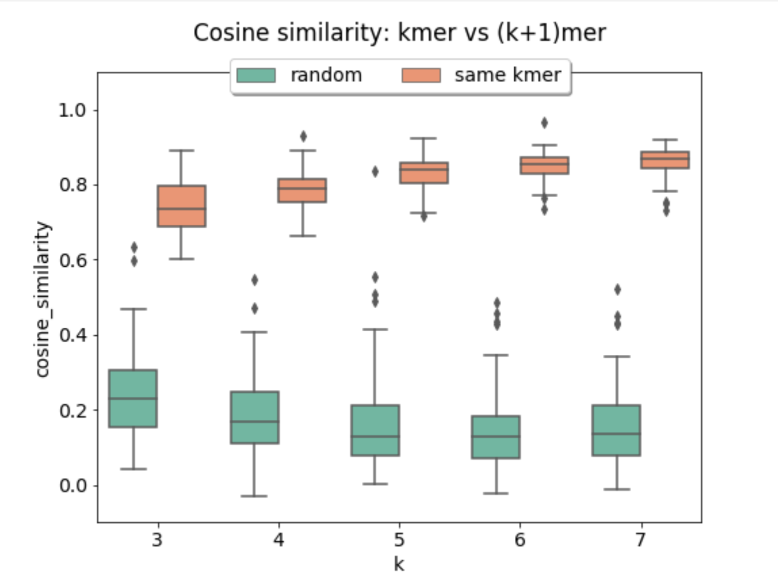
- Similarity between k-mers
  - ➢ $vec(king) - vec(man) + vec(woman) \approx vec(queen)$
  - ➢ $vec(ACT) \approx vec(ACTC)$
  - ➢ $vec(ACT) \approx vec(ACTA) + vec(ACTT) + vec(ACTC) + vec(ACTG)$
  - ➢ $vec(ACTCTG) \approx vec(ACT) + vec(CTG)$
- Application:
  - ➢ The performance of dna2vec should be significantly higher than one-hot encoding

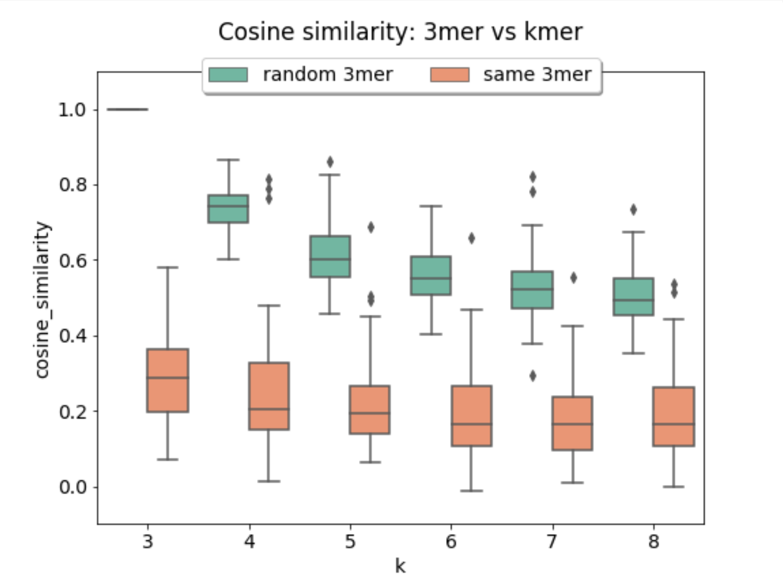
dna2vec Reflects Similarity Between K-mers

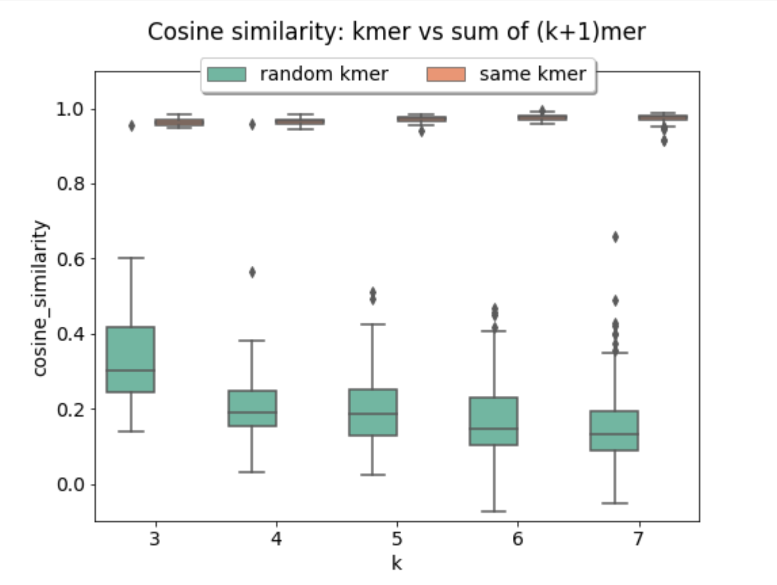
Three tests of cosine similarity:

1. $v(\mathrm{kmer}) \sim v(\mathrm{kmer} + \{A, T, C, G\}^1)$ for $\mathrm{kmer} \in \{A, T, C, G\}^k$: ACT ~ ACTC
2. $v(\mathrm{kmer}) \sim v(\mathrm{kmer} + \{A, T, C, G\}^n)$ for $\mathrm{kmer} \in \{A, T, C, G\}^k$: ACT ~ ACTCTCAC
3. $v(\mathrm{kmer}) \sim v(\mathrm{kmer} + A) + v(\mathrm{kmer} + T) + v(\mathrm{kmer} + C) + v(\mathrm{kmer} + G)$ for $\mathrm{kmer} \in \{A, T, C, G\}^k$

   ACT ~ ACTA+ACTT+ACTC+ACTG



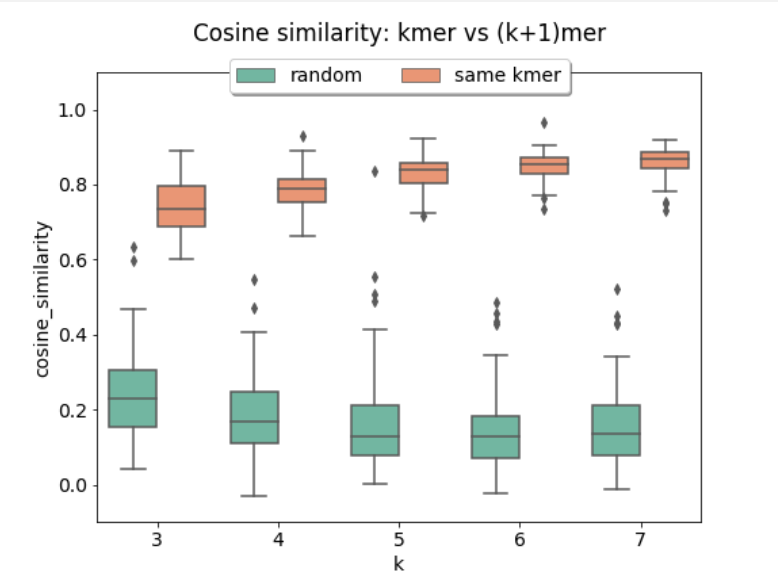


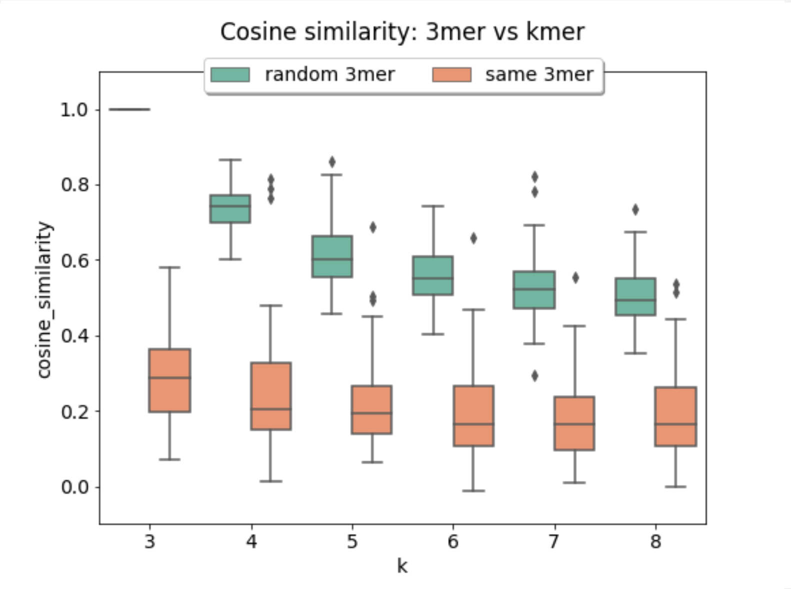


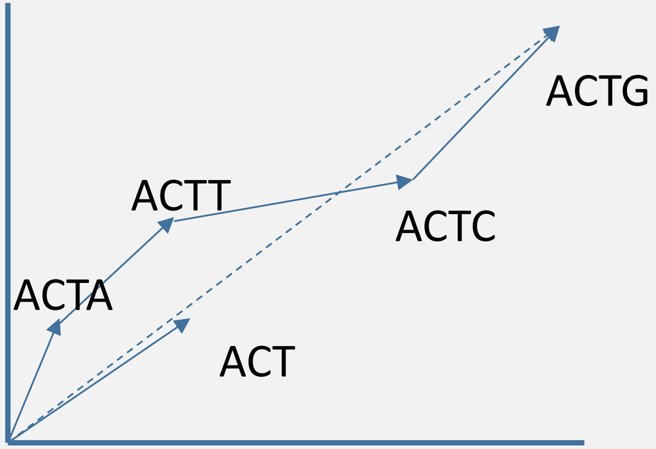
dna2vec Reflects Similarity Between K-mers

Three tests of cosine similarity:

1. $v(\mathrm{kmer}) \sim v(\mathrm{kmer} + \{A, T, C, G\}^1)$ for $\mathrm{kmer} \in \{A, T, C, G\}^k$: ACT ~ ACTC
2. $v(\mathrm{kmer}) \sim v(\mathrm{kmer} + \{A, T, C, G\}^n)$ for $\mathrm{kmer} \in \{A, T, C, G\}^k$: ACT ~ ACTCTCAC
3. $v(\mathrm{kmer}) \sim v(\mathrm{kmer} + A) + v(\mathrm{kmer} + T) + v(\mathrm{kmer} + C) + v(\mathrm{kmer} + G)$ for $\mathrm{kmer} \in \{A, T, C, G\}^k$
   ACT ~ ACTA+ACTT+ACTC+ACTG

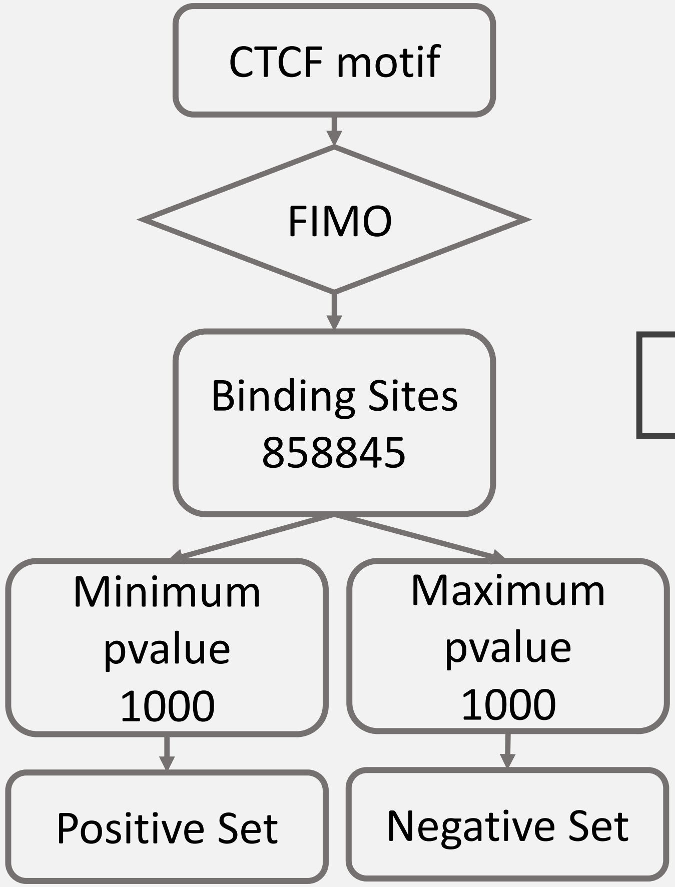

Cosine similarity: kmer vs (k+1)mer


Cosine similarity: 3mer vs kmer



dna2vec Increase the Performance of Downstream Analysis

1. Prepare P/N dataset



CTCF motif

FIMO

Binding Sites 858845

Minimum pvalue 1000

Maximum pvalue 1000

Positive Set

Negative Set

2. Three classifier

Model 1

$$\{N, A, T, C, G\}^{L} \rightarrow \{0,1,2,3,4\}^{L}$$

$$\{0,1,2,3,4\}^{L} \xrightarrow{SVM(linear)} \{0,1\}$$

Model 2

$$\{N, A, T, C, G\}^{L} \rightarrow \{0,1,2,3,4\}^{L}$$

$$\{0,1,2,3,4\}^{L} \xrightarrow{SVM(Gaussian)} \{0,1\}$$

Model 3

$$\{N, A, T, C, G\}^{L} \xrightarrow{Divide} \{N, A, T, C, G\}^{w*(L-w+1)}$$

$$\xrightarrow{dna2vec} R^{d*(L-w+1)}$$

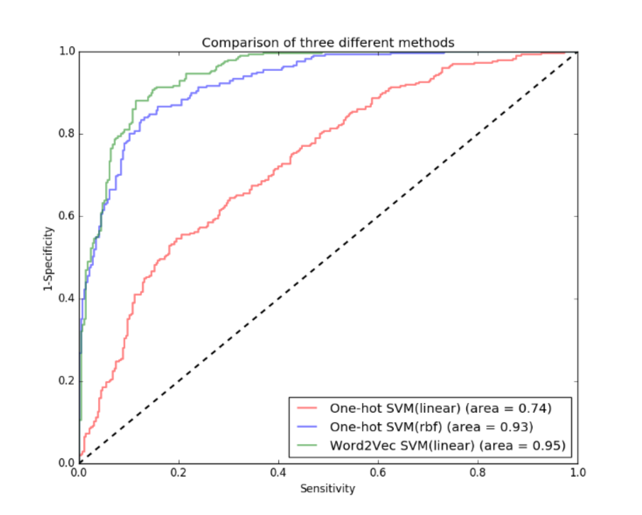
$$R^{d*(L-w+1)} \xrightarrow{SVM(linear)} \{0,1\}$$

3. Performance

Table: Accuracy

Encoding	Method	Accuracy
One-hot	SVM-linear	0.66
One-hot	SVM-Gaussian	0.85
Word2vec	SVM-linear	0.87

ROC Curve



Summary

- The training method of dna2vec model was presented
- The representation of k-mers with dna2vec was shown to be able to reflect the similarity between k-mers
- The performance of classifier trained with dna2vec was proved to be better than SVM (Gaussian Kernel)