dna2vec:
Consistent vector representations of variable-length k-mers

by Patrick Ng

TABLE OF CONTENT

01 Paper Introduction

Problem: The straightforward vector encoding of k-mer as a one-hot vector is vulnerable to the curse of dimensionality. Worse yet, the distance between any pair of one-hot vectors is equidistant.

Solution: The paper introduces a novel method to train distributed representations of **variable-length** k-mers based on **word2vec** model.

The contribution of the work includes:

- variable-length k-mer embedding model
- experimental evidence of similarity between arithmetic of dna2vec vectors and nucleotides concatenation
- relationship between Needleman-Wunsch alignment and cosine similarity of dna2vec vectors
- construction of nucleotide concatenation analogy using dna2vec arithmetic

02 Training Stages

The training of dna2vec using hg38 dataset is done in four stages:

1. separate genome into long non-overlapping DNA fragments
2. convert long DNA fragments into overlapping variable-length k-mers
3. unsupervised training of an aggregate embedding model using a two-layer neural network
4. decompose aggregated model by k-mer lengths

3 Paper Comprehension

Terminology & Concept [ongoing]

- Distributed Representation
- Categorical Variable
- One-Hot Encoding
- K-Mer Components
- Word2Vec
- Needleman-Wunsch Similarity Score
- Global Alignment Similarity
- hg38 Dataset
- Boxplot
- Cosine Similarity of Vectors

Categorical Data [Ref]

Variables containing **label** values
e.g.:

Person: student, teacher, male, female
Subject: CSE, EEE

Local Representation [Ref1] [Ref2]

One to One relationship between neuron & concept
e.g.: One-Hot Encoding example given above

One-Hot Encoding [Ref]

A method to convert categorical data into numerical data
e.g.:

	student	teacher	male	female
Faizah	**1**	0	0	0
Sir	0	**1**	0	0

Distributed Representation [Ref1] [Ref2]

Many to Many relationship between neuron & concept
e.g.:

	student	teacher	male	female
Faizah	**1**	0	0	**1**
Sir	0	**1**	**1**	0

Global Alignment [Ref1]

End to End Alignment is carried out between sequences.

Needleman-Wunsch Similarity Score [Ref]

The Needleman-Wunsch algorithm is a dynamic programming algorithm for global sequence alignment.

The algorithm assigns a score to every possible alignment and the purpose is to find all possible alignments having the highest score.

In this paper, Needleman-Wunsch similarity score were computed using Biopython's ***align.globalxx*** function, which used

match score = 1
mismatch = 0
gap penalty = 0

Cosine Similarity [Ref1] [Ref2]

magnitude of data does not matter, only orientation matters

$$\vec{a} \cdot \vec{b} = \|\vec{a}\| \|\vec{b}\| \cos\theta$$

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|}$$

Word2Vec [Ref1]

Technique to learn word embeddings using shallow neural network. The objective is to have words with similar context occupy close spatial positions. Mathematically, the cosine of the angle between such vectors should be close to 1, i.e. angle close to 0.

Two kinds of word2vec models:

1. CBOW
2. SKIP-GRAM

In this paper, the author follows skip-gram model.

04 Experiments

1. Similarity and nearest neighbors

$$sim(v,w) = \frac{v \cdot w}{||v||\,||w||}$$

nearest-neighbor of dna2vec vector **v** is a k-mer computed with:

$$NearestNeighbor_k(v) = \underset{s \in \{A,C,G,T\}^k}{\arg\max} \; sim(v, vec(s))$$

2. dna2vec arithmetic and nucleotide concatenation

Summing dna2vec embeddings is related to concatenating k-mers.

The following condition would be a success for 1-NN:

$$NearestNeighbor_6\ (vec(AAC) + vec(TCT)) \in \{AACTCT, TCTAAC\}$$

Likewise, the following would be a success for n-NN:

$$nNearestNeighbors_6\ (vec(AAC) + vec(TCT)) \cap \{AACTCT, TCTAAC\} \neq \emptyset$$

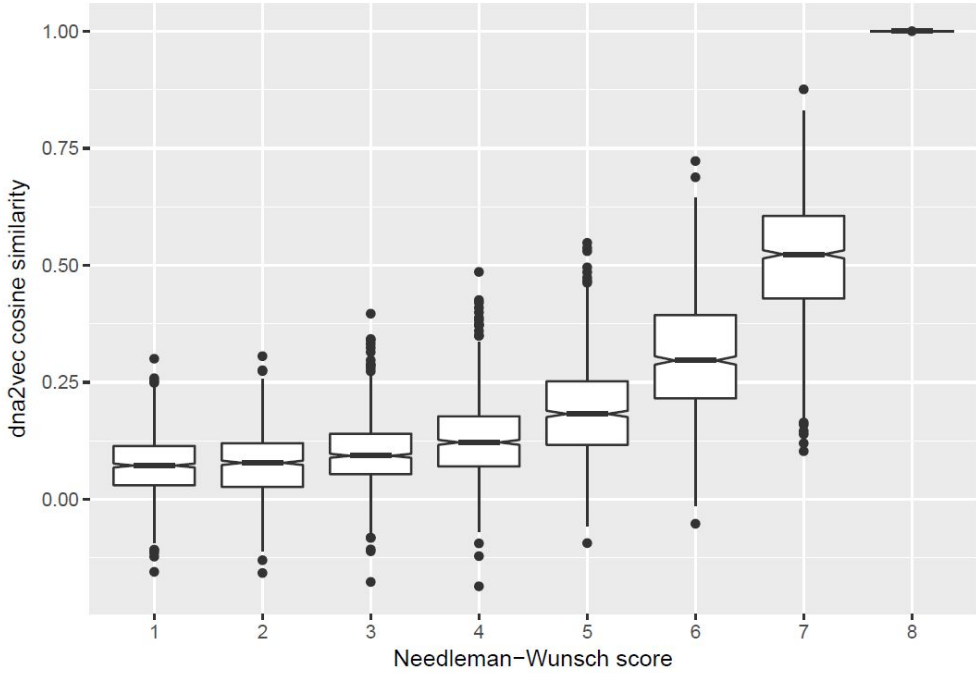
Operands	Concatenated	1-NN	5-NN	10-NN
3-mer + 3-mer	6-mer	28.7%	80.3%	94.6%
3-mer + 4-mer	7-mer	49.9%	90.4%	97.4%
3-mer + 5-mer	8-mer	53.9%	94.0%	98.4%
4-mer + 4-mer	8-mer	73.5%	96.8%	99.2%

Observation of 1000 samples for each operand.
Each row is aggregated from summing the dna2vec vectors of individual pairs of arbitrary x-mer and observing whether each of their string concatenation overlaps with the vector sum's n-nearest k-mer neighbors for each operand.

3. Relationship to global alignment similarity

In Figure 1, The author sampled 1000 pairs of 8-mers for each Needleman-Wunsch score level and plot their Needleman-Wunsch similarity score against dna2vec cosine similarity.

We see that edit distance between two arbitrary k-mers is correlated with the cosine distance of their corresponding dna2vec vectors.

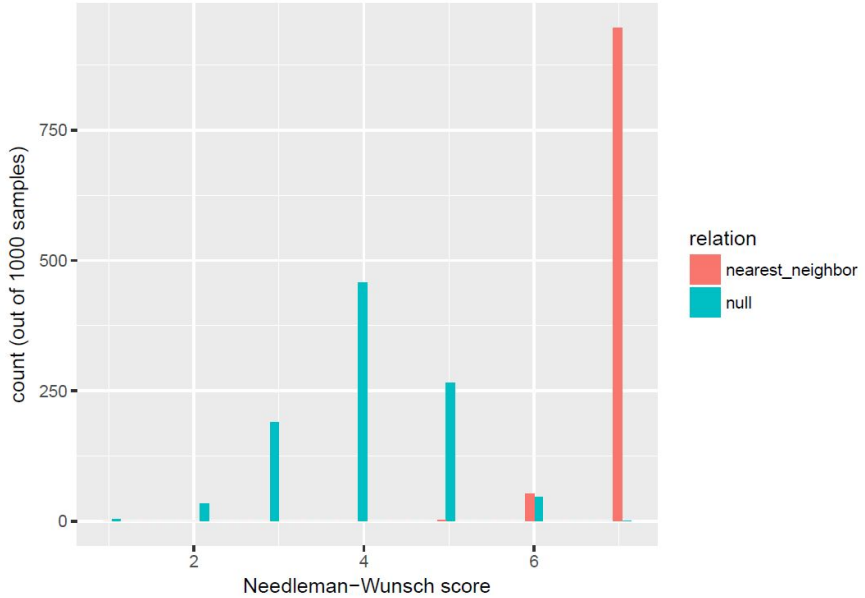


3. Relationship to global alignment similarity

In Figure 2, the author found each nearest neighbor of 1000 8-mers computed the Needleman-Wunsch score for each pair.

They compared the Needleman-Wunsch similarity distribution of k-mer and its nearest dna2vec neighbor against distribution of two random k-mers.

As illustrated in these two figures, they found evidence that the dna2vec nearest-neighbor exhibits alignment similarity.



5 Implementation

Software Requirements [ongoing]

- Git, Git Bash
- Pip, Python 3.7.4
- GCC, GFortran Compilers
- Other Requirements (Around 20)

Python For ML [ongoing]

- Youtube
- Coursera
- Scikit-learn

5 Implementation

Git & GitHub [complete]

- Create Account
- Learn Git Essentials
- Clone the Repo

Training & Analysis [on hold]

- Download the hg38 Dataset
- Train and Analyze Results

Q & A