

# Vector representation of large DNA and protein strings

Bogdan Kirillov   Phan Duc   Evgenij Baraboshkin

Skolkovo Institute of Science and Technology



Skolkovo Institute of Science and Technology

## The key papers to use

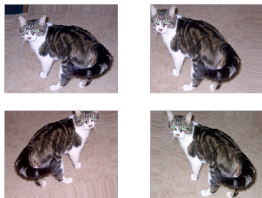
- ▶ Dna2vec: Consistent vector representation of variable-length k-mers, Ng,P – the one we are to discuss here;
- ▶ Distributed Representations for Biological Sequence Analysis, Kimothi et al., ;
- ▶ Continuous Distributed Representation of Biological Sequences for Deep Genomics and Deep Proteomics, Asgari, E. and Mofrad, M;

# Bioinformatics 101

Biological data are complex!

- ▶ Direct Data Augmentation is not really applicable;
- ▶ Working with really large (1Mb and more) DNA sequences is unclear;
- ▶ Unlike basic CV, there usually are no clue about how well we can solve the problem with given training/testing data;
- ▶ A lot of experiments are low-throughput so sample sizes can be limited and the samples may overlap;

Augmentation in Computer Vision

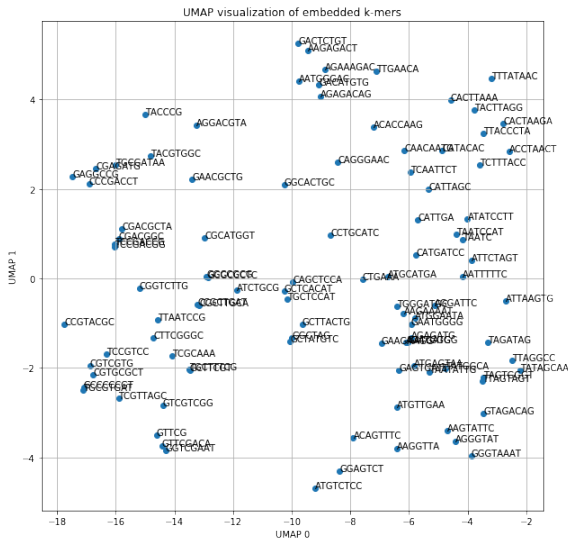


Cropped and mirrored cat pic remains a cat pic

Biological sequence case



You can't say the same here



### Example of k-mer embedding

# Problem of k-mer length choice

```
1 s = "the quick brown fox jumps over the lazy dog"
```

```
1 s.split()
```

```
['the', 'quick', 'brown', 'fox', 'jumps', 'over', 'the', 'lazy', 'dog']
```

```
1 a = "ATTATAGGCGACGATAGCGATAGCGATAGCGATCGTACG"
```

```
1 overlapping(a, [3,4,5])
```

```
array(['ATTA', 'TTATA', 'TAT', 'ATAG', 'TAGGC', 'AGG', 'GGCG', 'GCGA',  
      'CGAC', 'GACG', 'ACG', 'CGAT', 'GATA', 'ATAGC', 'TAGCG', 'AGCGA',  
      'GCGAT', 'CGA', 'GATA', 'ATA', 'TAG', 'AGCG', 'GCGA', 'CGAT',  
      'GAT', 'ATA', 'TAGC', 'AGC', 'GCGA', 'CGATC', 'GATCG', 'ATC',  
      'TCG', 'CGTA', 'GTAC', 'TACG', 'ACG'], dtype='<U5')
```

No words in DNA

## dna2vec training procedure

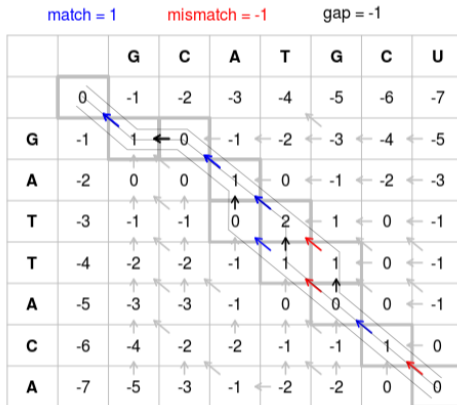
1. separate genome into long non-overlapping DNA fragments based on gaps;
2. convert long DNA fragments into overlapping variable-length k-mers;
3. unsupervised training of an aggregate embedding model using a two-layer neural network;
4. decompose aggregated model by k-mer lengths.



Example of gap in genome

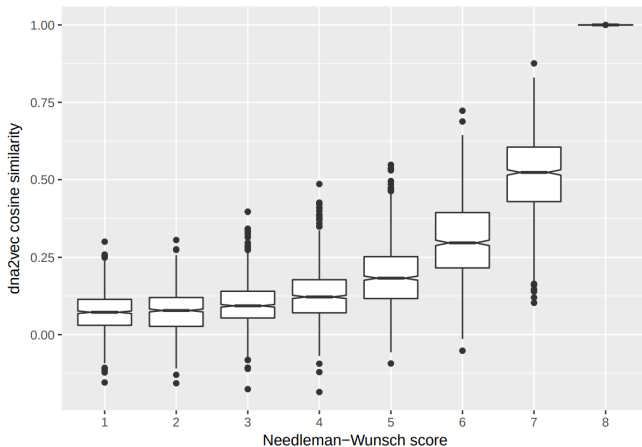
# Alignment in Bioinformatics

Needleman-Wunsch



Needleman-Wunsch algorithm (Wikipedia)

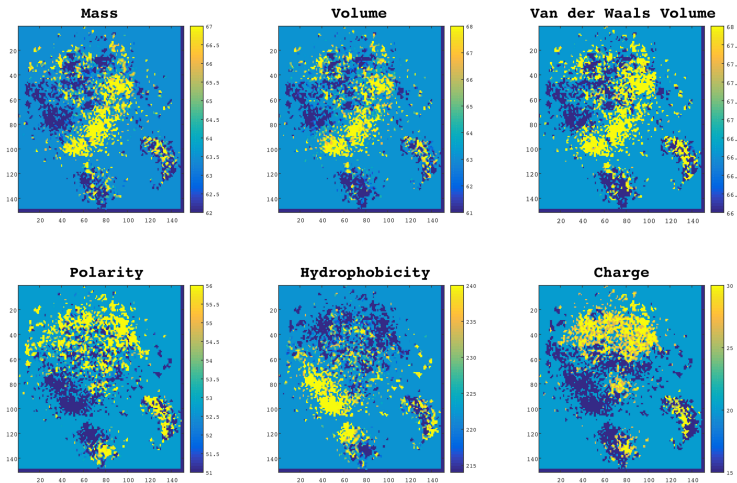
# Alignment similarity



Cosine similarity of dna2vec and NW score correlate with  $r=0.83$



# Possible applications



Distribution of physical properties in embedding space (Asgari and Mofrad)