🧬 Exploring WORD2VEC models for Capturing the Similarity of Codon Embeddings

Anđa Denić, Jelena Pejić, Aleksandar Trokicić

University of Niš
Faculty of Sciences
and Mathematics

Words as high dimensional vectors

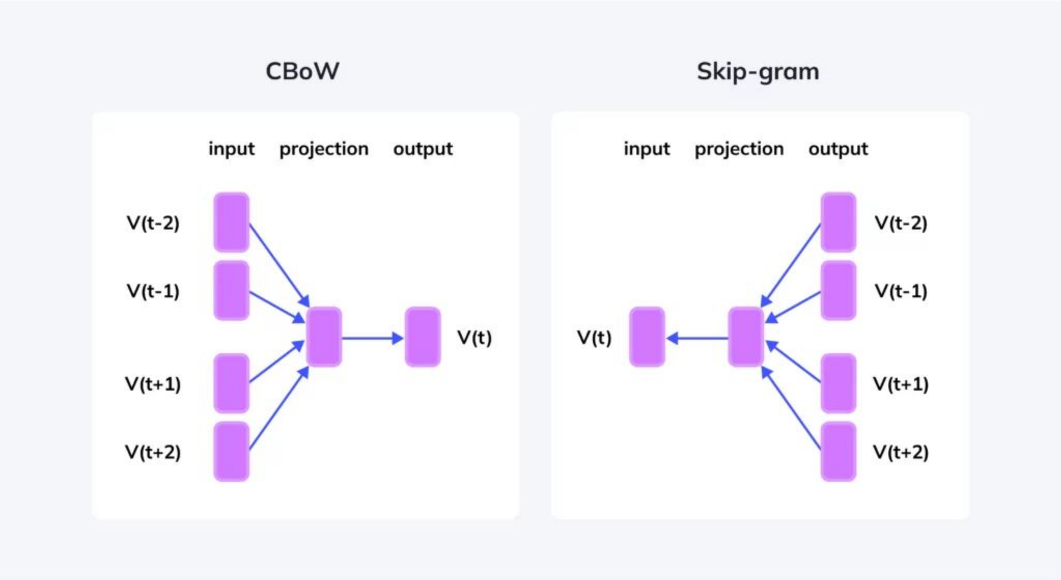
☕

vec(‘cup’) = [1.22, 44.444, …, 5.78]

vec(‘coffee’) = [3.23, 50.4, …, 5.78]

Word2vec

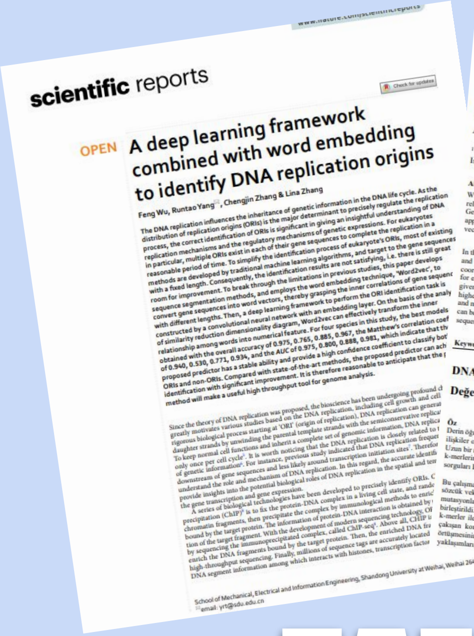
model that learns a vector representation of words



Word2vec

Word meaning is based on the **surrounding words** (**context**).

Today is a **beautiful** **and** **sunny** **day** **in** **Kragujevac**, **the** **fourth** largest city in Serbia.



www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

A deep learning framework combined with word embedding to identify DNA replication origins

Feng Wu, Runtao Yang, Chengjin Zhang & Lina Zhang

The DNA replication influences the inheritance of genetic information in the DNA life cycle. As the distribution of replication origins (ORIs) is the major determinant to precisely regulate the replication process, the correct identification of ORIs is significant in giving an insightful understanding of DNA replication mechanisms and the regulatory mechanisms of genetic expressions. For eukaryotes in particular, multiple ORIs exist in each of their gene sequences to complete the replication in a reasonable period of time. To simplify the identification process of eukaryote's ORIs, most of existing methods are developed by traditional machine learning algorithms, and target to the gene sequences with a fixed length. Consequently, the identification results are not satisfying, i.e. there is still great room for improvement. To break through the limitations in previous studies, this paper develops sequence segmentation methods, and employs the word embedding technique, 'Word2vec', to convert gene sequences into word vectors, thereby grasping the inner correlations of gene sequenc with different lengths. Then, a deep learning framework to perform the ORI identification task is constructed by a convolutional neural network with an embedding layer. On the basis of the analy of similarity reduction dimensionality diagram, Word2vec can effectively transform the inner relationship among words into numerical feature. For four species in this study, the best models obtained with the overall accuracy of 0.975, 0.765, 0.885, 0.967, the Matthew's correlation coef of 0.940, 0.530, 0.771, 0.934, and the AUC of 0.975, 0.800, 0.888, 0.981, which indicate that th proposed predictor has a stable ability and provide a high confidence coefficient to classify bot ORIs and non-ORIs. Compared with state-of-the-art methods, the proposed predictor can ach identification with significant improvement. It is therefore reasonable to anticipate that the p method will make a useful high throughput tool for genome analysis.

Since the theory of DNA replication was proposed, the bioscience has been undergoing profound c greatly motivates various studies based on the DNA replication, including cell growth and cell rigorous biological process starting at 'ORI' (origin of replication), DNA replication can generat daughter strands by unwinding the parental template strands with the semiconservative replica To keep normal cell functions and inherit a complete set of genomic information, DNA replica only once per cell cycle[1]. It is worth noticing that the DNA replication is closely related to t of genetic information[2]. For instance, previous study indicated that DNA replication freque downstream of gene sequences and less likely around transcription initiation sites[3]. Therefor understand the role and mechanism of DNA replication. In this regard, the accurate identifi provide insights into the potential biological roles of DNA replication in the spatial and tem the gene transcription and gene expression.

A series of biological technologies have been developed to precisely identify ORIs. C precipitation (ChIP)[4] is to fix the protein-DNA complex in a living cell state, and rand chromatin fragments, then precipitate the complex by immunological methods to enric bound by the target protein. The information of protein-DNA interaction is obtained by tion of the target fragment. With the development of modern sequencing technology, OI by sequencing the immunoprecipitated complex, called ChIP-seq[5]. Above all, CHIP i enrich the DNA fragments bound by the target protein. Then, the enriched DNA fra high-throughput sequencing. Finally, millions of sequence tags are accurately located DNA segment information among which interacts with histones, transcription factor

School of Mechanical, Electrical and Information Engineering, Shandong University at Weihai, Weihai 264200, China. email: yrt@sdu.edu.cn

Research Article
Journal of Intelligent Systems: Theory and Applications 3(1) 2020: 1-6

Journal of Intelligent Systems: Theory

Assessment of Mutation Susceptibility in DNA Sequences w Word Vectors

Alper Yilmaz [1]

[1]Yildiz Technical University, Faculty of Chemical and Metallurgical Engineering, Department of Bioengineeri Istanbul/ Turkey

Abstract

With the advent of natural language processing (NLP) techniques empowered with deep learning approaches, more detailec relationships between words have been unraveled. Word2Vec is quite robust in discovering contextual and semantic relationships. Genome being a long text, is subject to similar studies to unravel yet to be discovered relationships between DNA k-mers. Dna2vec applies Word2Vec approach to whole genome so that DNA k-mers are represented as vectors. The cosine similarity queries on DNA vectors reveal unusual relationships between DNA k-mers.

In this study, we examined DNA sequence based prediction of mutation susceptibility. Initially, we generated word vectors for human and mouse genome via dna2vec. On the other hand, we retrieved coordinates of common and all mutations from dbSNP. For each coordinate, we extracted 8 nucleotide k-mers intersecting mutations and results are aggregated such a way that number of mutations for each 8-mer has been tabulated. These results are incorporated with dna2vec cosine similarity data. Our results showed that for a given k-mer, k-mers with highest cosine similarity coincide with highest mutation count k-mer. In other words, the neighbor with the highest cosine similarity for a k-mer was also seen to be the neighbor overlapping the mutation count. As a result of our studies, human and mouse, dna2vec vs. mutation overlap is 80% and 70%, respectively. In conclusion, dna2vec and other word embedding approaches can be used to reveal mutation or variation characteristics of genomes without sequencing or experimental data, solely using the genome sequence itself. This might pave the way for understanding the underlying mechanism or dynamics of mutations in genomes.

Keywords: mutation, word2vec, dna2vec, k-mer, cosine-similarity.

DNA Dizilerinde Kelime Vektörleri ile Mutasyon Yatkınlığının Değerlendirilmesi

Öz

Derin öğrenme yaklaşımları ile güçlendirilen doğal dil işleme (NLP) tekniklerinin ortaya çıkmasıyla, kelimeler arasındaki daha ayrıntılı ilişkiler ortaya çıkarılmıştır. Bu açıdan word2vec yöntemi bağlamsal ve anlamsal ilişkileri keşfetme konusunda oldukça gelişmiştir. Uzun bir metin olan genom, DNA k-merleri arasındaki ilişkileri henüz keşfedememiş olan benzer çalışmalara tabidir. Dna2vec, DNA k-merlerinin vektör olarak gösterilmesi için tüm genoma word2vec yaklaşımını uygular. DNA vektörleri üzerindeki kosinüs benzerlik sorguları DNA k-merleri arasında olağandışı ilişkiler ortaya koymaktadır.

Bu çalışmada, mutasyon duyarlılığının DNA dizisi temelli tahmini incelendi. Başlangıçta, insan ve fare genomu için dna2vec yoluyla sözcük vektörleri üretildi. Diğer yandan, ortak ve tüm mutasyonların koordinatlarını dbSNP'den alındı. Her koordinat için, kesişen mutasyonlar halinde 8 nükleotidlik k-merler çıkarıldı ve sonuçlar toplandı. Bu sonuçlar dna2vec kosinüs benzerlik verileri ile birleştirildi. Sonuçlarımız, belirli bir k-mer için, en yüksek kosinüs benzerliğine sahip k-merlerin en yüksek mutasyon sayısına sahip k-merler ile çakıştığını göstermiştir. Başka bir deyişle, bir k-mer için kosinüs benzerliği en yüksek komşunun, mutasyon sayısıyla çakışan komşu olduğu da görülmüştür. Çalışmalarımız sonucunda insanda ve farede, k-mer kosinüs benzerliği ve mutasyon örtüşmesinin oranları sırasıyla %80 ve %70 olduğu görülmüştür. Bu çalışmalar sonucunda dna2vec ve diğer kelime gömme yaklaşımları, sadece genom dizisinin kendisi kullanılarak, dizileme veya deney verileri olmadan genomların mutasyon veya varyasyon

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

Patrick Ng
`ppn3@cs.cornell.edu`

Abstract

One of the ubiquitous representation of long DNA sequence is dividing it into shorter k-mer components. Unfortunately, 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. This is particularly problematic when applying the latest machine learning algorithms to solve problems in biological sequence analysis. In this paper, we propose a novel method to train distributed representations of variable-length k-mers. Our method is based on the popular word embedding model *word2vec*, which is trained on a shallow two-layer neural network. Our experiments provide evidence that the summing of dna2vec vectors is akin to nucleotides concatenation. We also demonstrate that there is correlation between Needleman-Wunsch similarity score and cosine similarity of dna2vec vectors.

1 Introduction

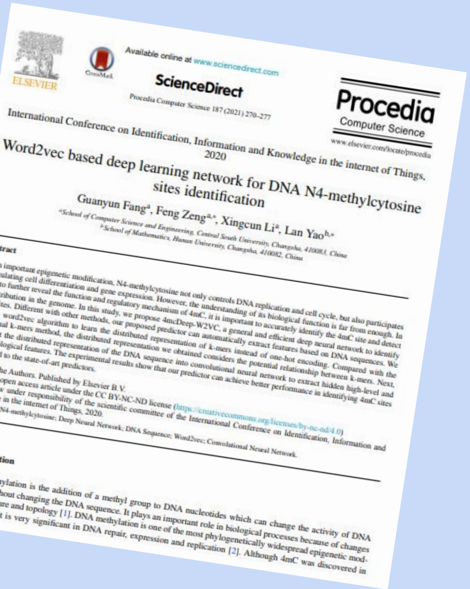
The usage of k-mer representation has been a popular approach in analyzing long sequence of DNA fragmen The k-mer representation is simple to understand and compute. Unfortunately, its straightforward vect encoding as a one-hot vector (i.e. bit vector that consists of all zeros except for a single dimension vulnerable to curse of dimensionality. Specifically, its one-hot vector has dimension exponential to the le of k. For example, an 8-mer needs a bit vector of dimension $4^8 = 65536$. This is problematic when app the latest machine learning algorithms to solve problems in biological sequence analysis, due to the fac most of these tools prefer lower-dimensional continuous vectors as input (Suykens and Vandewalle Angermueller et al., 2016; Turian et al., 2010). Worse yet, the distance between any arbitrary pair of vectors is equidistant, even though `ATGGC` should be closer to `ATGGG` than `CACGA`.

1.1 Word embeddings

The Natural Language Processing (NLP) research community has a long tradition of using bag-of- one-hot vector, where its dimension is equal to the vocabulary size. Recently, there has been an using *word embeddings* as inputs to machine learning algorithms, especially in the deep learnin (Mikolov et al., 2013b; LeCun et al., 2015; Bengio et al., 2013). Word embeddings are vectors of that are distributed representations of words.

A popular training technique for word embeddings, word2vec (Mikolov et al., 2013a), con 2-layer neural network that is trained on the current word and its surrounding context wo 2.3). This reconstruction of context of words is loosely inspired by the linguistic concept *hypothesis*, which states that words that appear in the same context have similar meani Deep learning algorithms applied with word embeddings have had dramatic improveme machine translation (Sutskever et al., 2014; Bahdanau et al., 2014; Cho et al., 2014), sum et al., 2016), sentiment analysis (Kim, 2014; Dos Santos and Gatti, 2014) and image c et al., 2015).

One of the most fascinating properties of word2vec is that its vector arithmetic can solve s analogies (Mikolov et al., 2013c,a). They showed that $vec(king) - vec(man) + vec($ is interpreted as finding In particular, the analogy task `man:king :: woman:???` is closest to $vec(w)$ under cosine distance. Fu $vec(king) - vec(man) + vec(woman)$ is closest to $vec(w)$ for past-tense relation $vec(capture) - vec(captured)$ 2014) showed that analogy works



ELSEVIER

CrossMark

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 187 (2021) 270–277

Procedia
Computer Science

www.elsevier.com/locate/procedia

International Conference on Identification, Information and Knowledge in the internet of Things, 2020

Word2vec based deep learning network for DNA N4-methylcytosine sites identification

Guanyun Fang[a], Feng Zeng[a,*], Xingcun Li[a], Lan Yao[b,*]

[a]School of Computer Science and Engineering, Central South University, Changsha, 410083, China
[b]School of Mathematics, Hunan University, Changsha, 410082, China

Abstract

As an important epigenetic modification, N4-methylcytosine not only controls DNA replication and cell cycle, but also participates in regulating cell differentiation and gene expression. However, the understanding of its biological function is far from enough. In order to further reveal the function and regulatory mechanism of 4mC, it is important to accurately identify the 4mC site and detect its distribution in the genome. In this study, we propose 4mcDeep-W2VC, a general and efficient deep neural network to identify 4mC sites. Different with other methods, our proposed predictor can automatically extract features based on DNA sequences. We use the word2vec algorithm to learn the distributed representation of k-mers instead of one-hot encoding. Compared with the traditional k-mers method, the distributed representation we obtained considers the potential relationship between k-mers. Next, we input the distributed representation of the DNA sequence into convolutional neural network to extract hidden high-level and more biological features. The experimental results show that our predictor can achieve better performance in identifying 4mC sites compared to the state-of-art predictors.

© 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the International Conference on Identification, Information and Knowledge in the internet of Things, 2020.

Keywords: N4-methylcytosine; Deep Neural Network; DNA Sequence; Word2vec; Convolutional Neural Network.

1. Introduction

DNA methylation is the addition of a methyl group to DNA nucleotides which can change the activity of DNA fragments without changing the DNA sequence. It plays an important role in biological processes because of changes in DNA structure and topology [1]. DNA methylation is one of the most phylogenetically widespread epigenetic modifications and it is very significant in DNA repair, expression and replication [2]. Although 4mC was discovered in

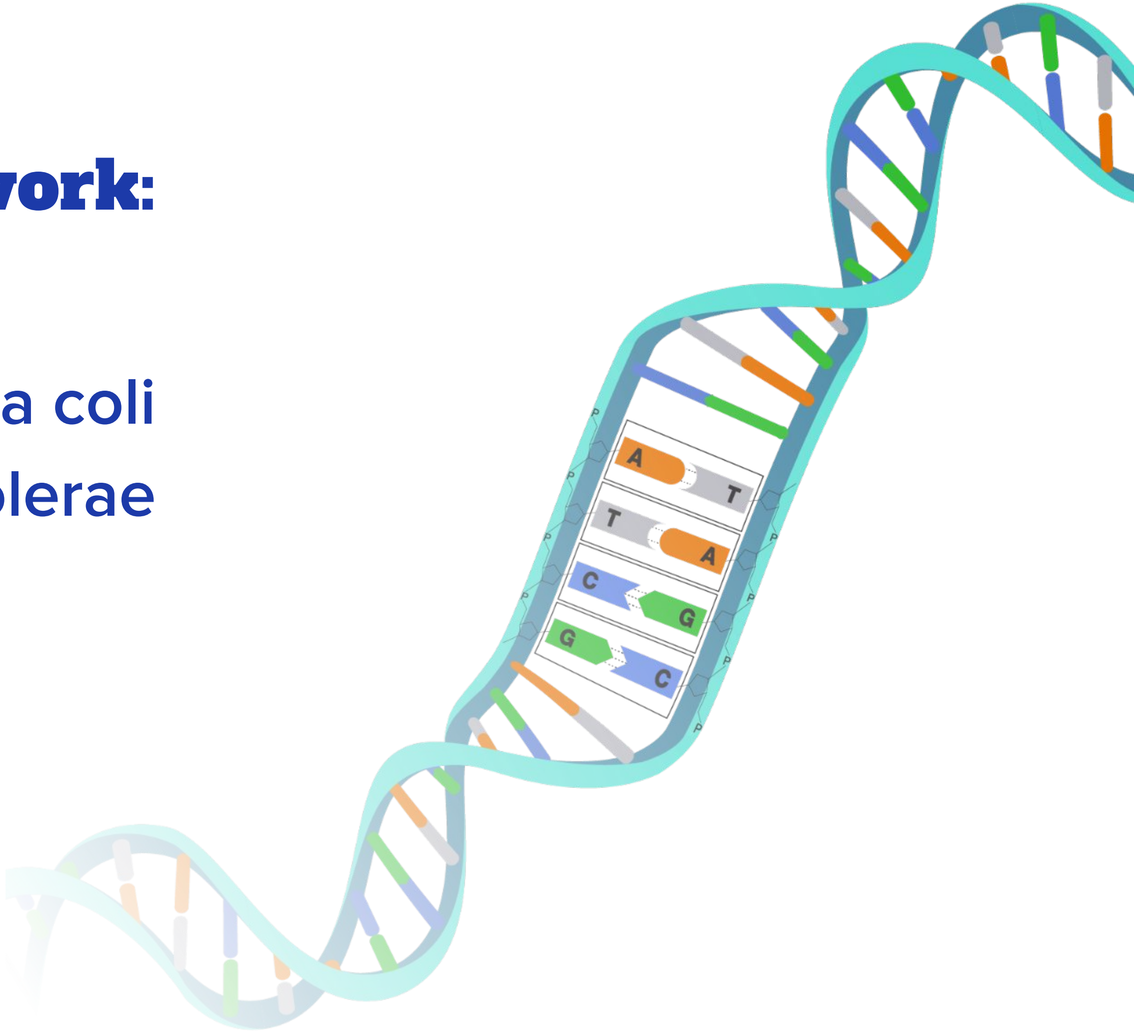
Word2vec in bioinformatics

Question:

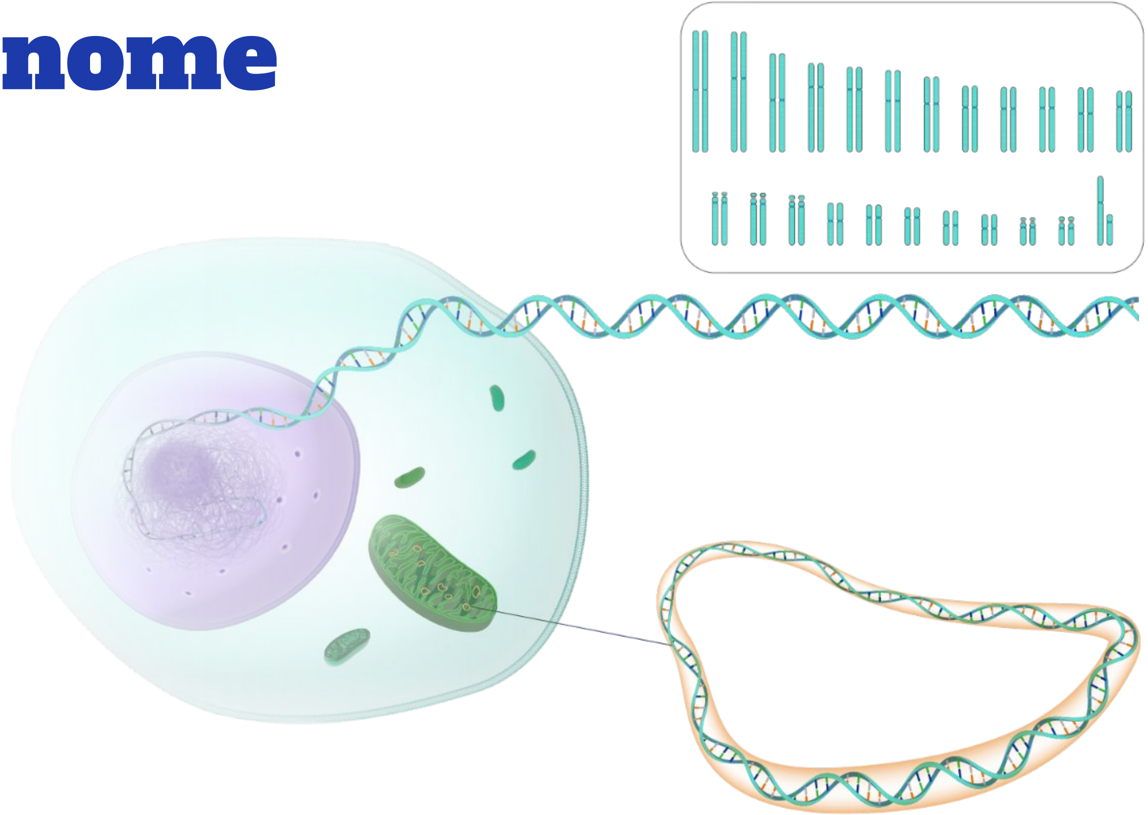
AI can `understand` language, can it `understand` DNA? 🤔

Our work:

- 🧬Escherichia coli
- 🧬Vibrio cholerae

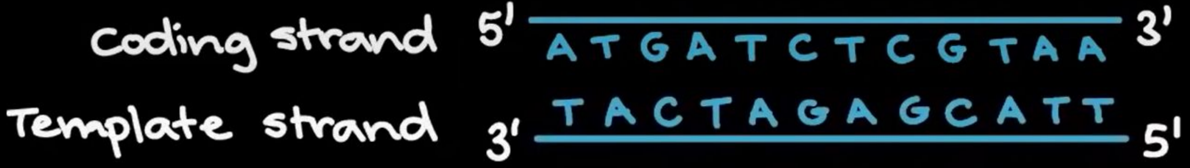


Genome

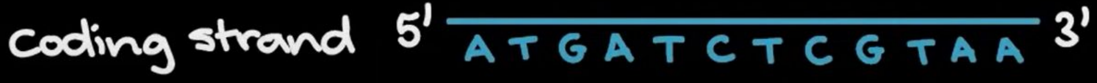


Genome

DNA strands:

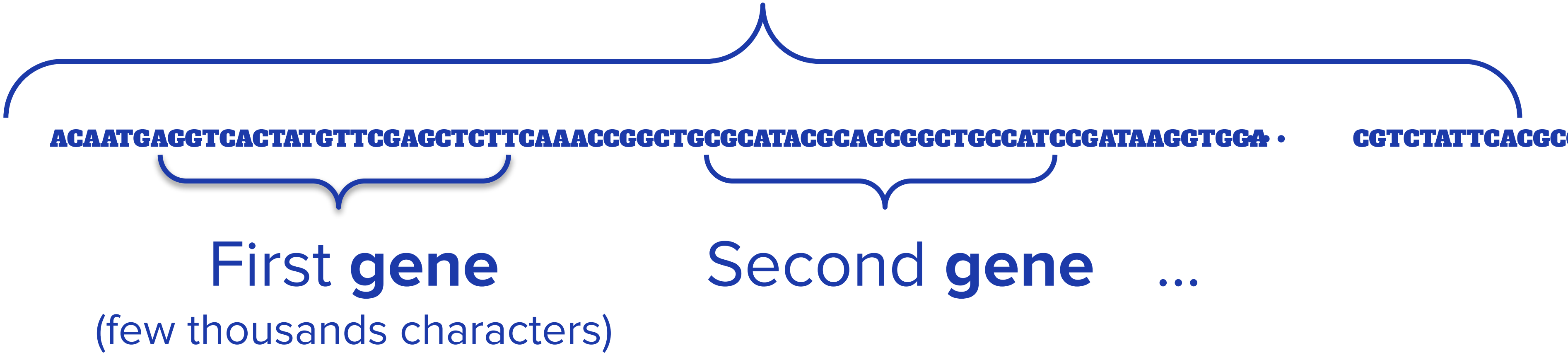


DNA sequence representation:



Genome

(few millions characters)



ACAATGAGGTCACTATGTTCGAGCTCTTCAAACCGGCTGCGCATACGCAGCGGCTGCCATCCGATAAGGTGGA • • CGTCTATTCACGGC

First **gene**

(few thousands characters)

Second **gene** …

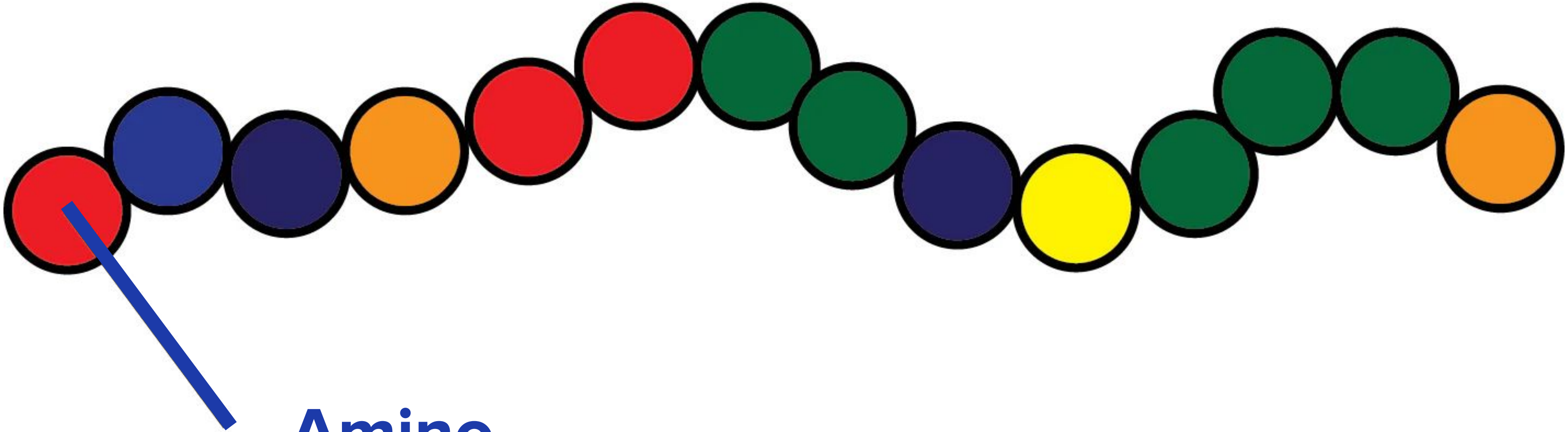
ACAATGAGGTCACTATGTTCGAGCTCTTCAAACCGGCTGCGCATACGCAGCGGCTGCCATCCGATAAGGTGGA• •   CGTCTATTCACGGC

Genes

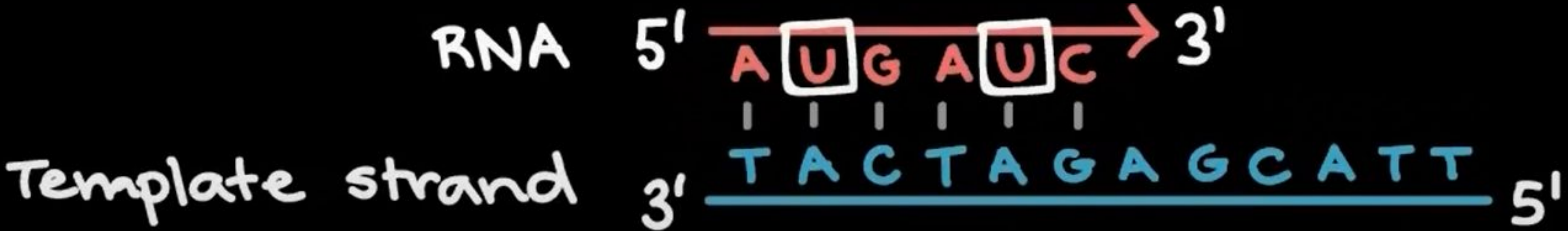
code for

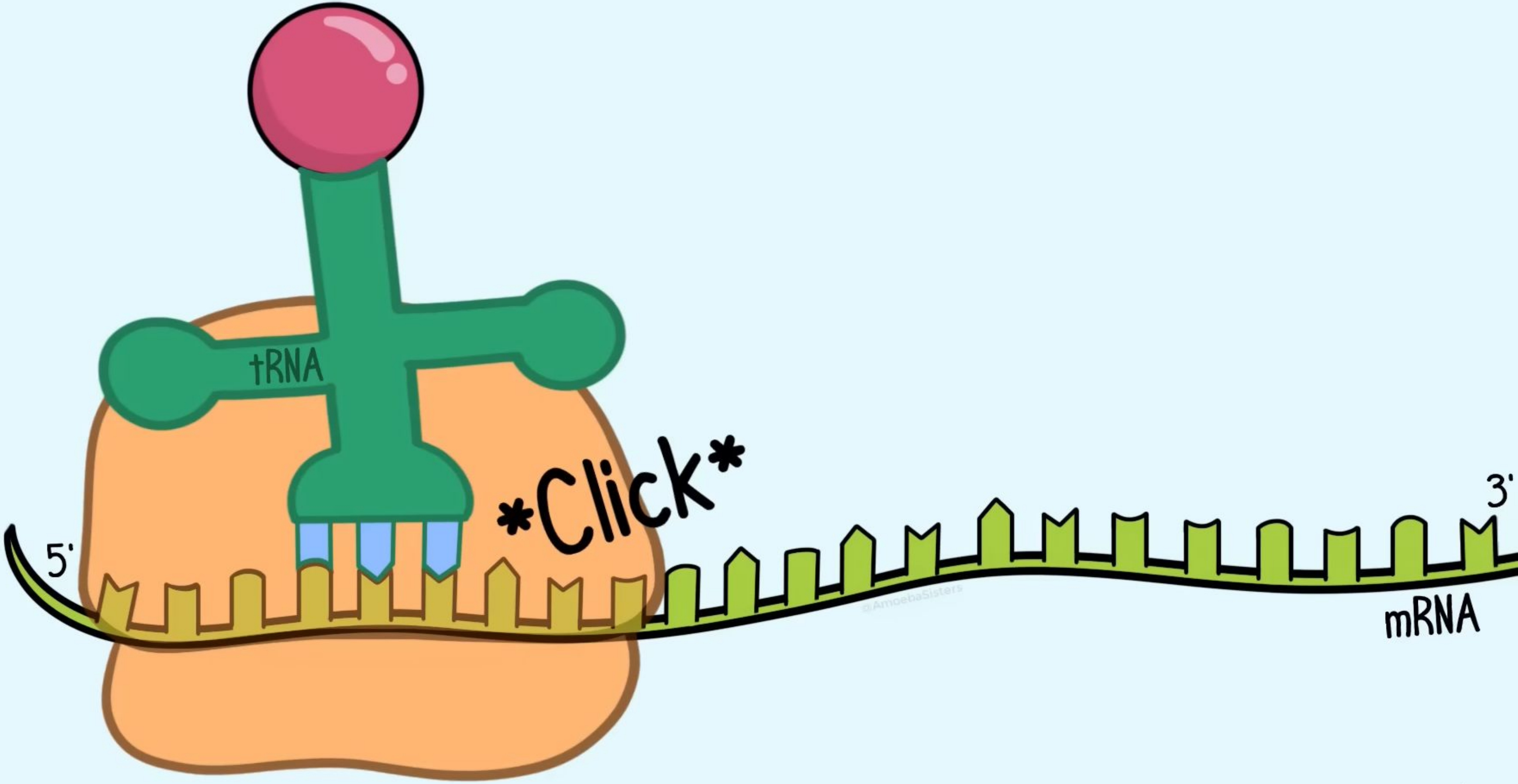
proteins

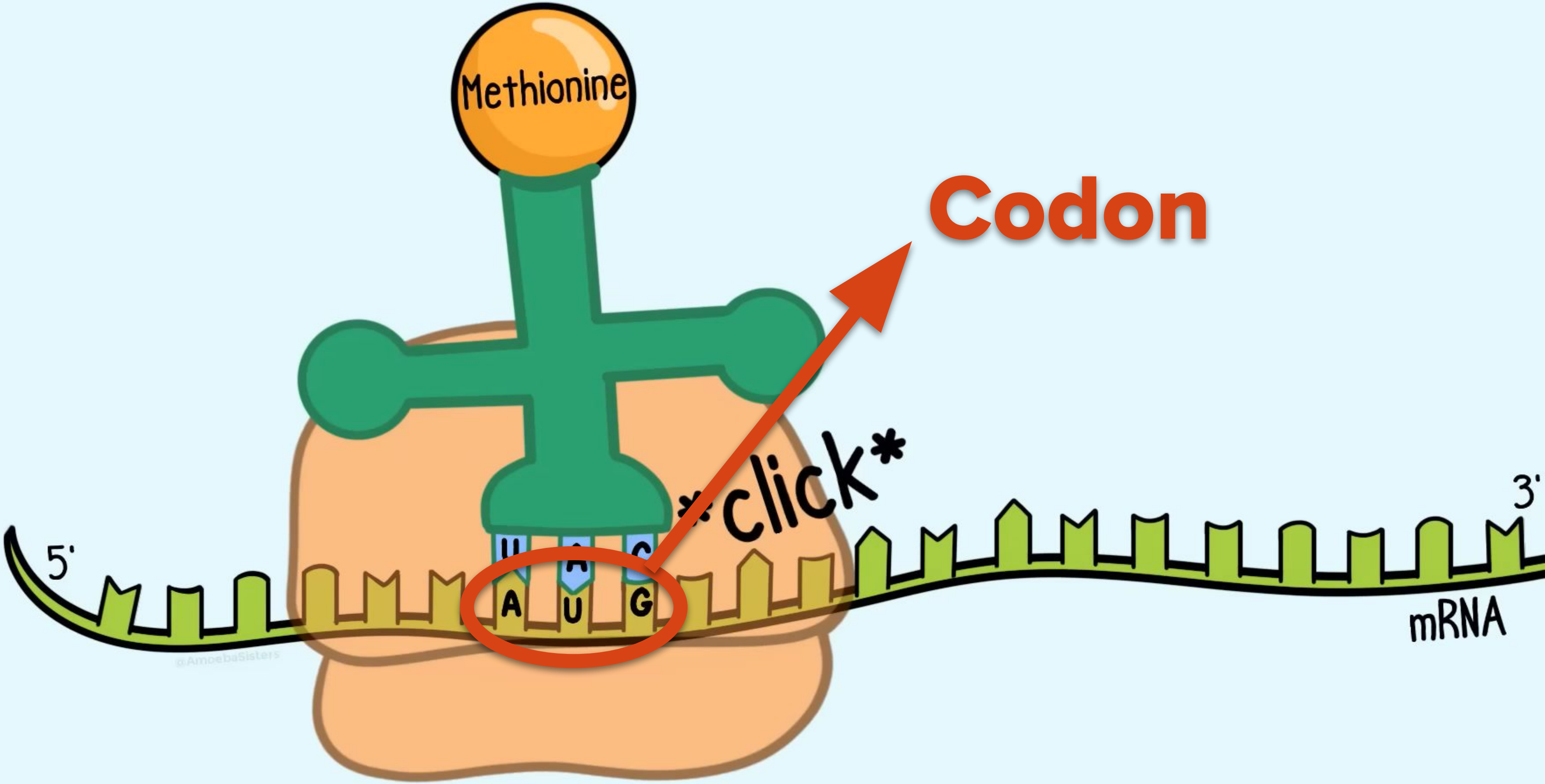
Protein



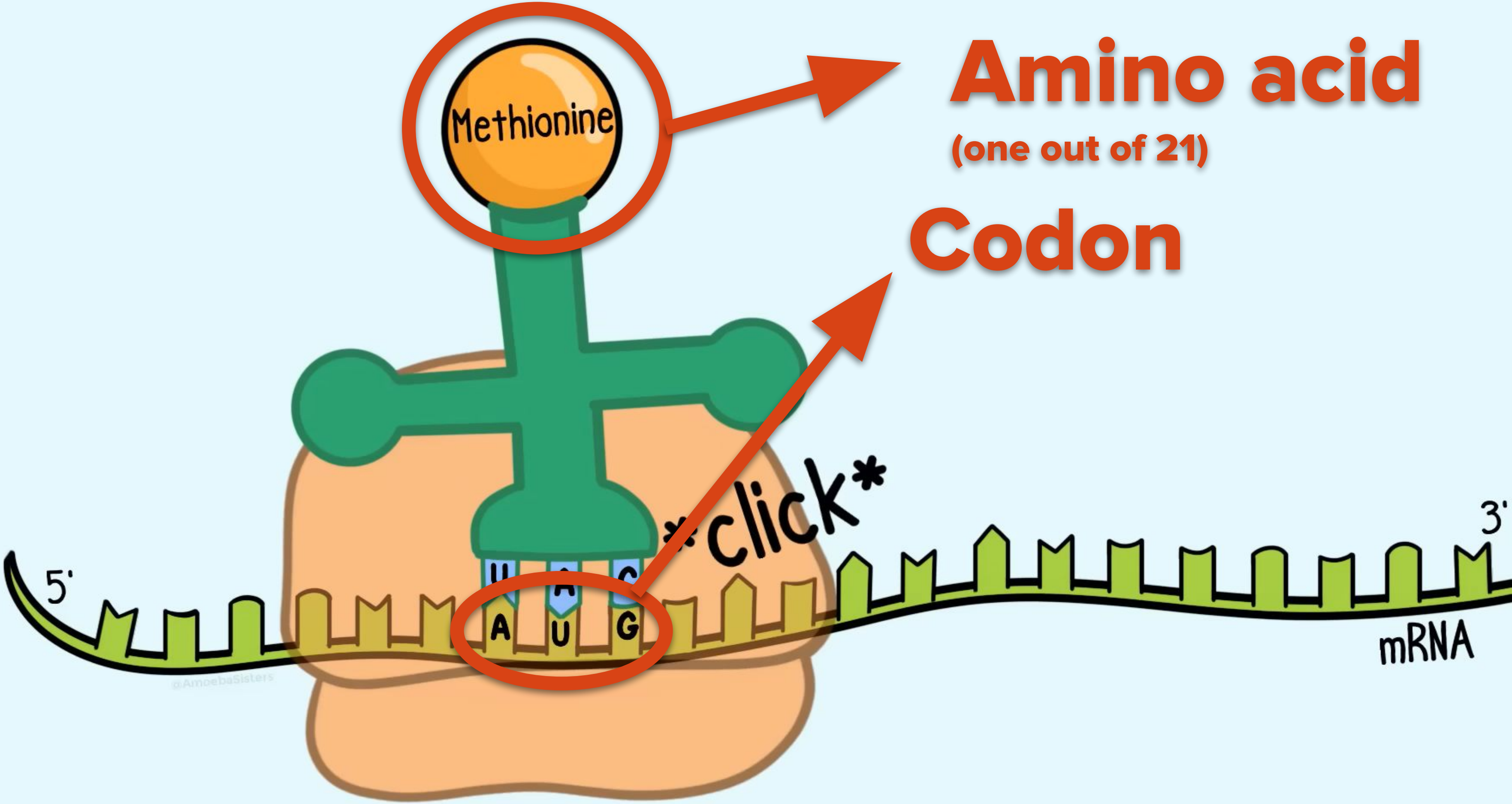
Amino acid


RNA
5'
A U G A U C
3'
Template strand
3'
T A C T A G A G C A T T
5'


tRNA
Click
5'
3'
mRNA
@AmoebaSisters
The Amoeba Sisters



Codon



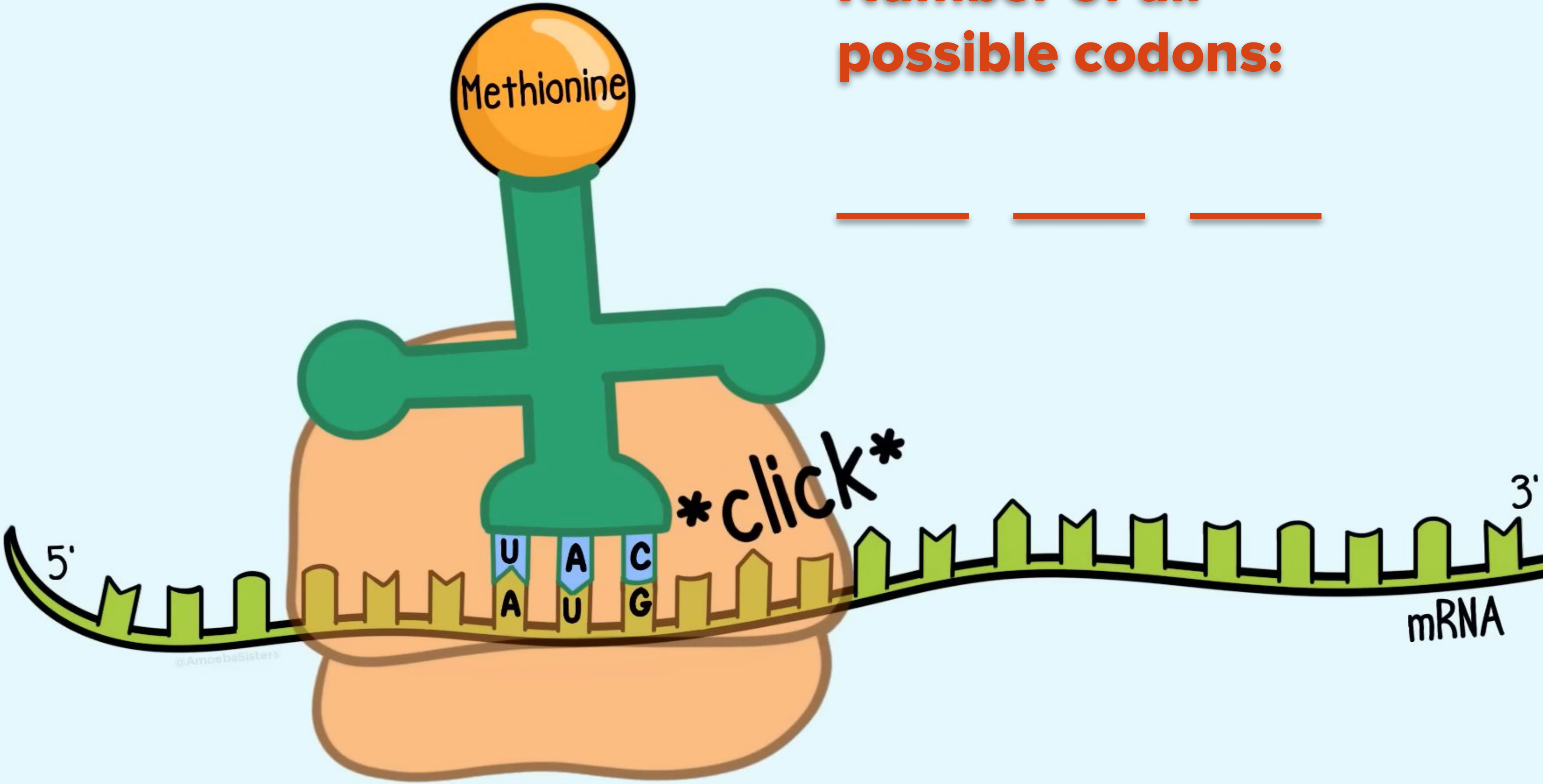
Amino acid

(one out of 21)

Codon

Number of all possible codons:

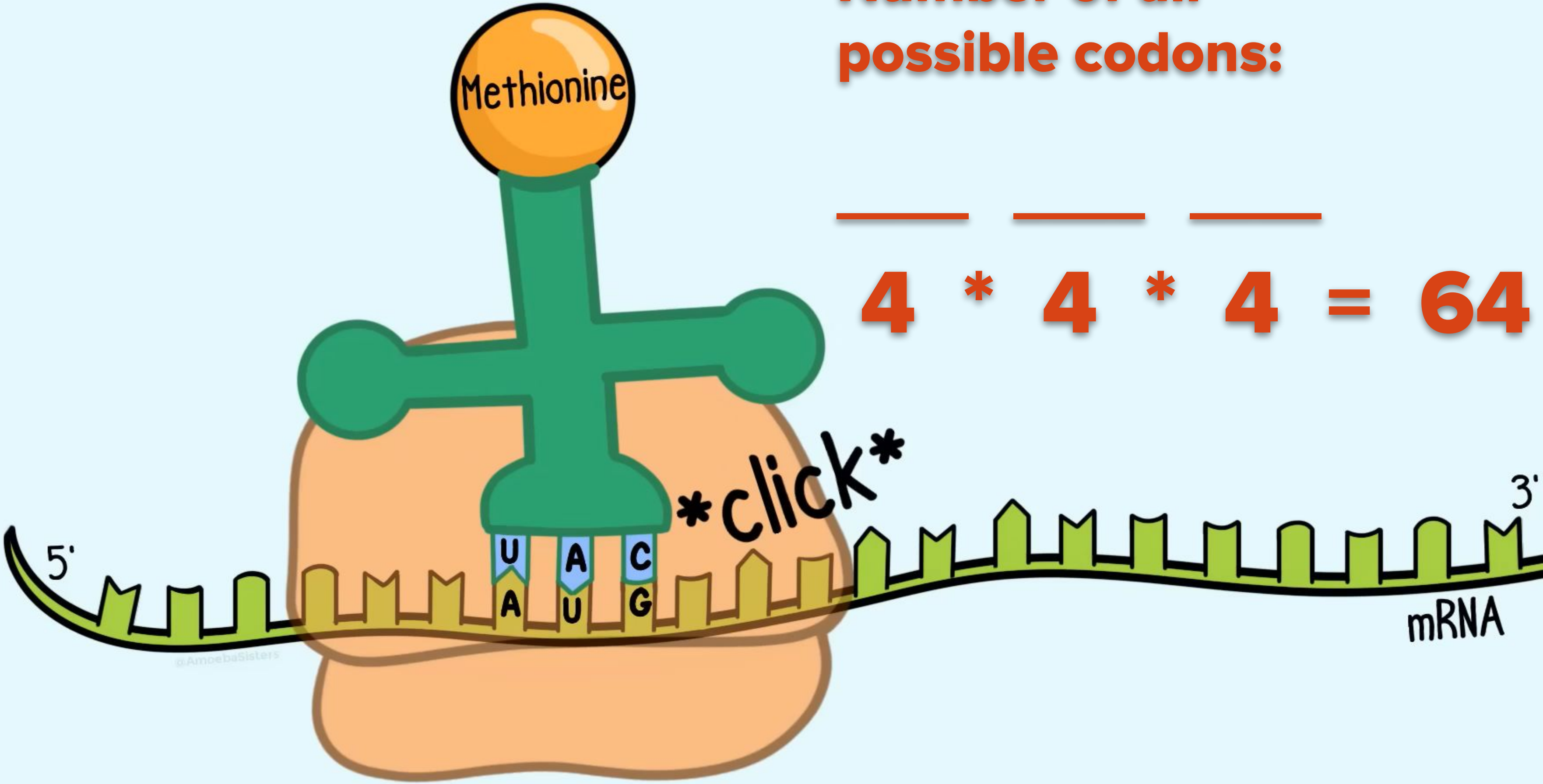
___ ___ ___

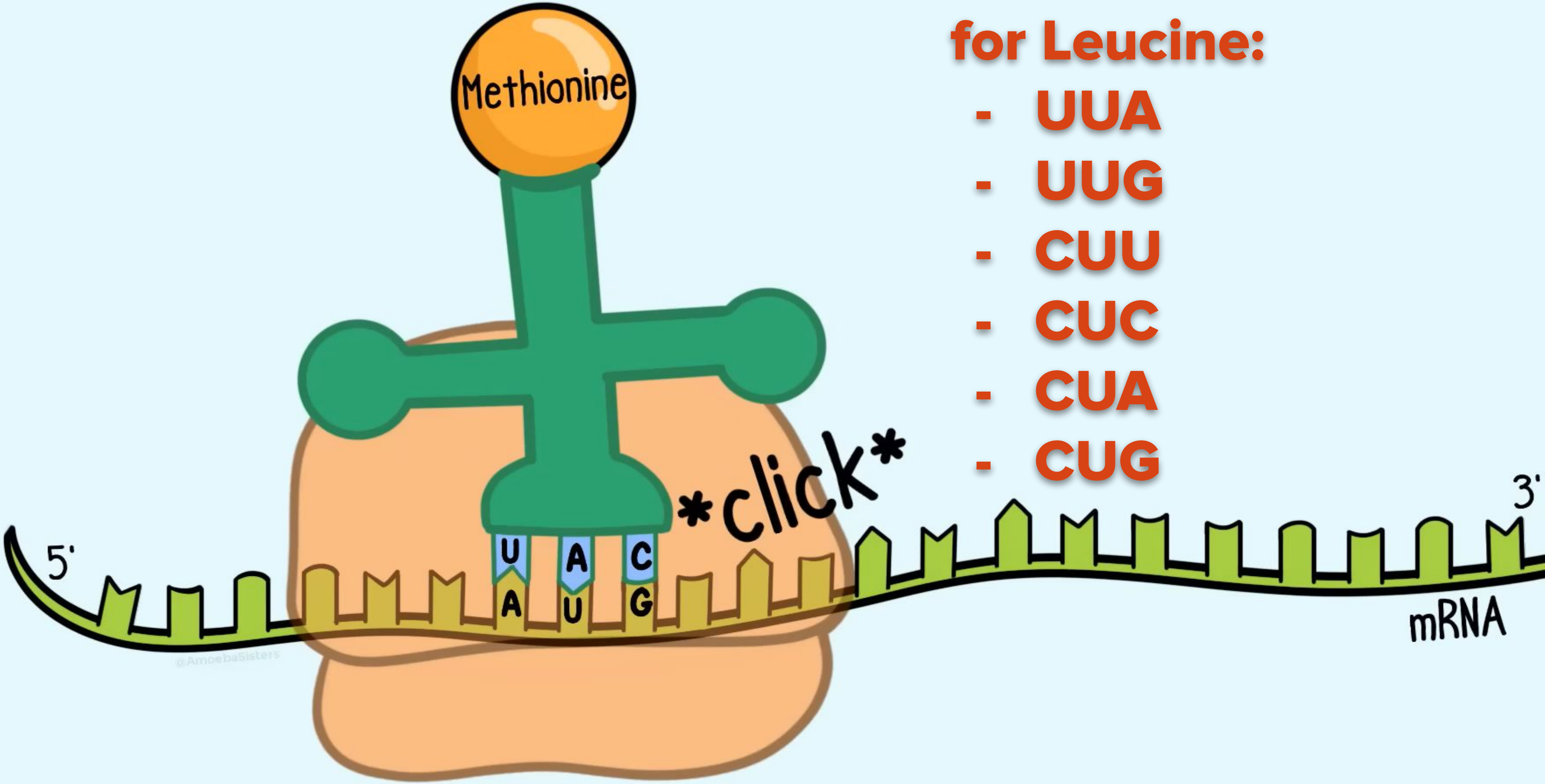


Number of all possible codons:

___ ___ ___

4 * 4 * 4 = 64

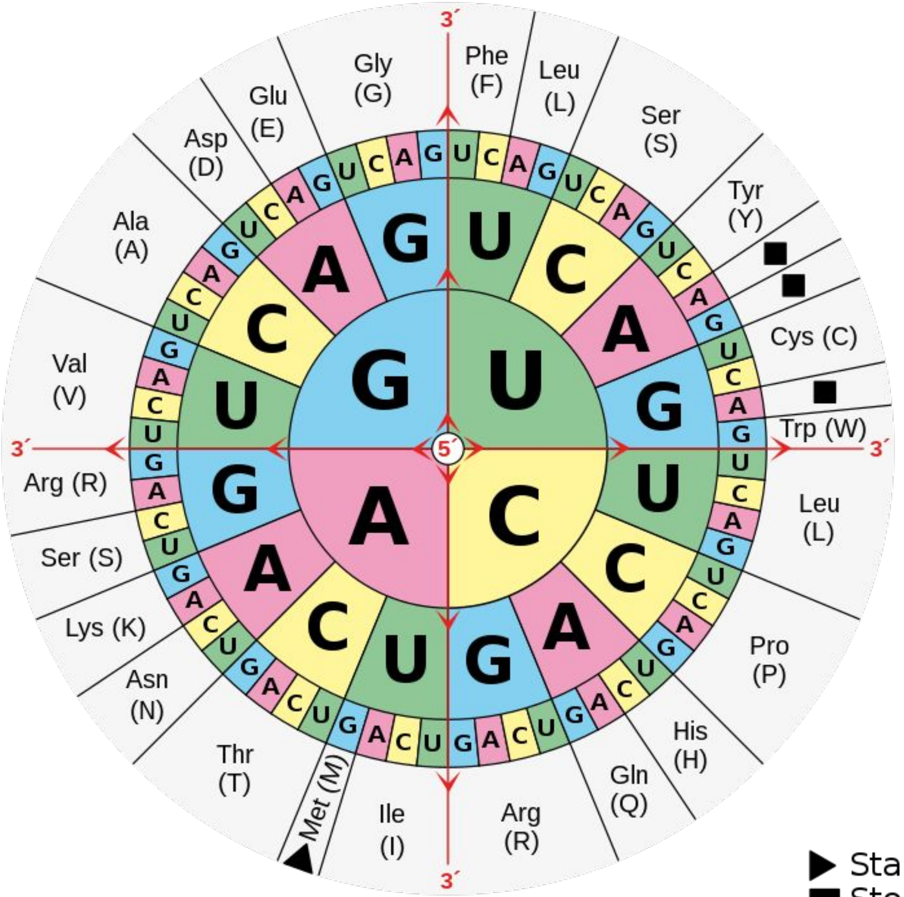




Codons that code for Leucine:

- UUA
- UUG
- CUU
- CUC
- CUA
- CUG

Codon chart:



▶ Start
■ Stop

Context in NLP

Context of a **word 'day'**:

Today is a **beautiful** **and** **sunny** **day** **in** **Kragujevac**, **the** **fourth** largest city in Serbia.

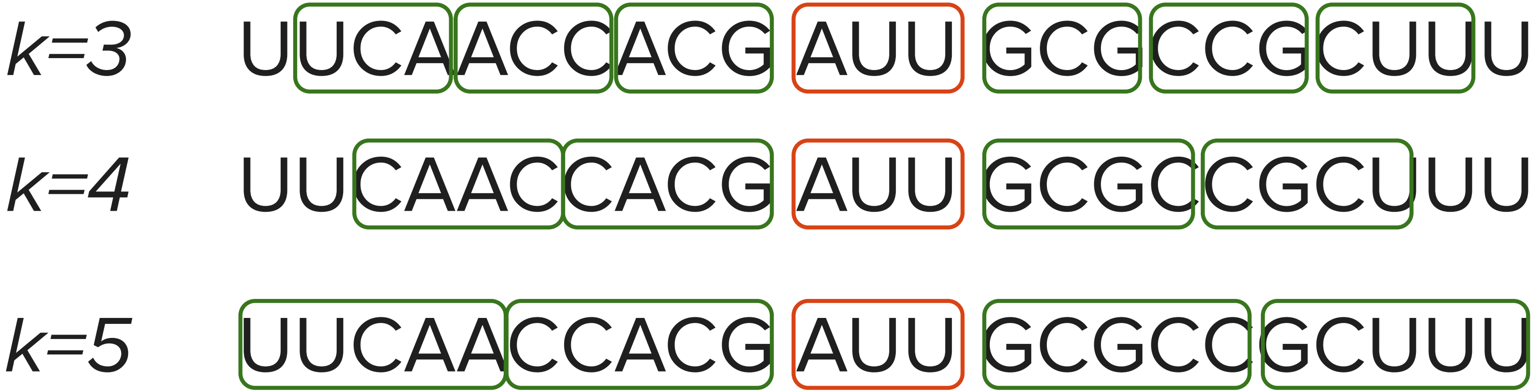
Context in bioinformatics?

Context of a **codon 'AUU'**?

UUCAACCACG AUU GCGCCGCUUU

Context in bioinformatics?

Context of a **codon ‘AUU’**?



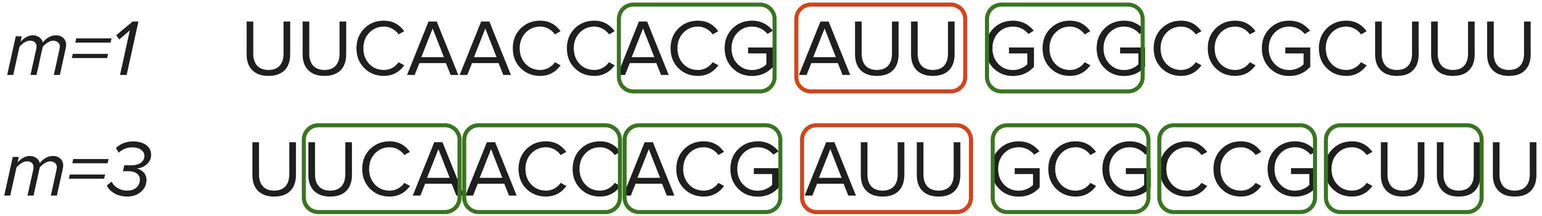
k=3 UUCAACCACG AUU GCGCCGCUUUU

k=4 UUCAACCACG AUU GCGCCGCUUU

k=5 UUCAACCACG AUU GCGCCGCUUU

Context in bioinformatics?

Context of a **codon ‘AUU’**?

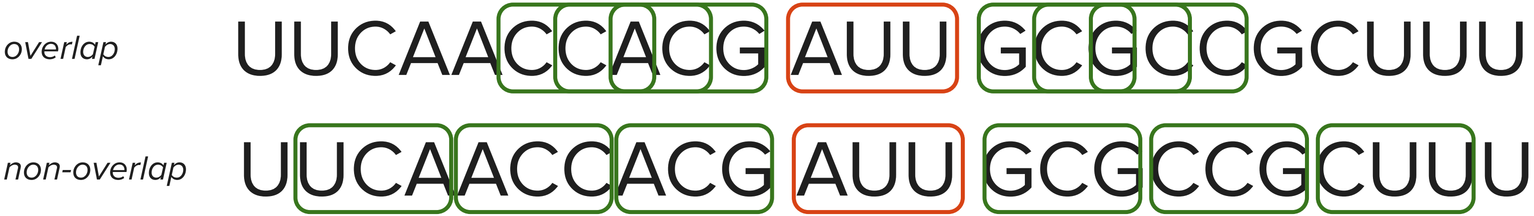


m=1 UUCAACCACG AUU GCGCCGCUUU

m=3 UUCAACCACG AUU GCGCCGCUUU

Context in bioinformatics?

Context of a **codon ‘AUU’**?



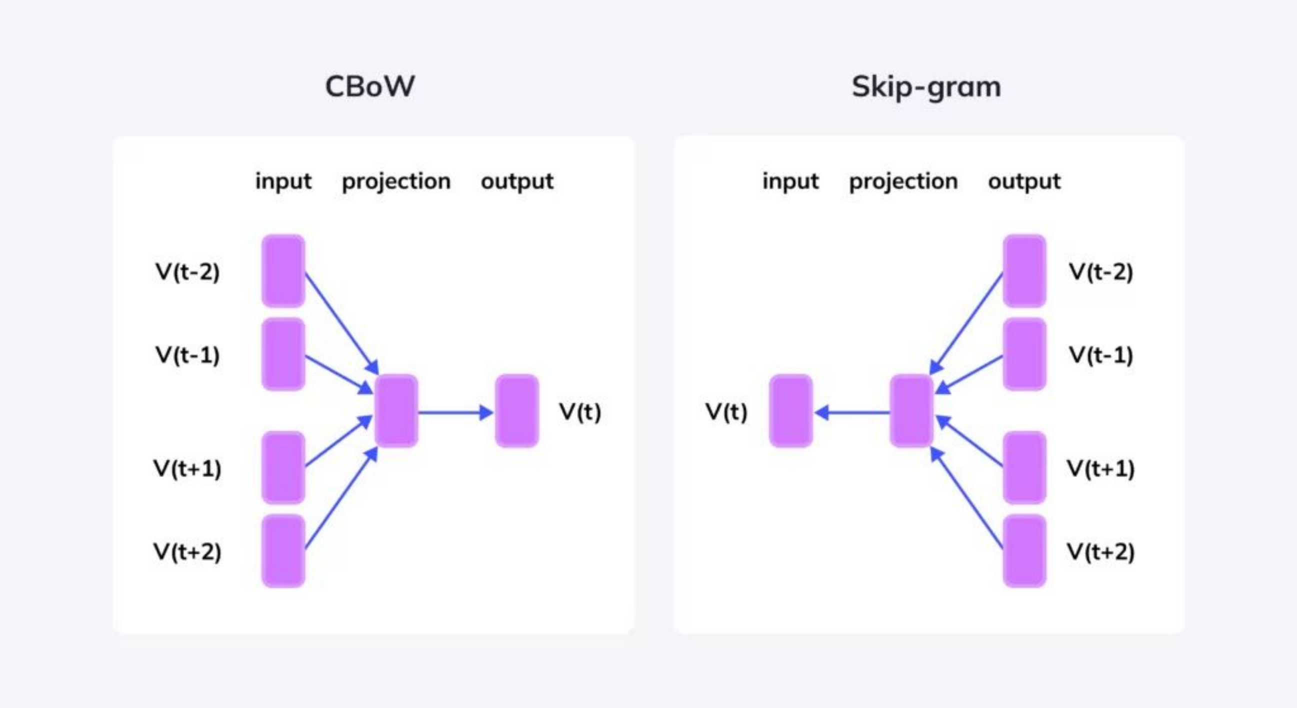
overlap UUCAACCACG AUU GCGCCGCUUU

non-overlap UUCAACCACG AUU GCGCCGCUUU

What is the best context?

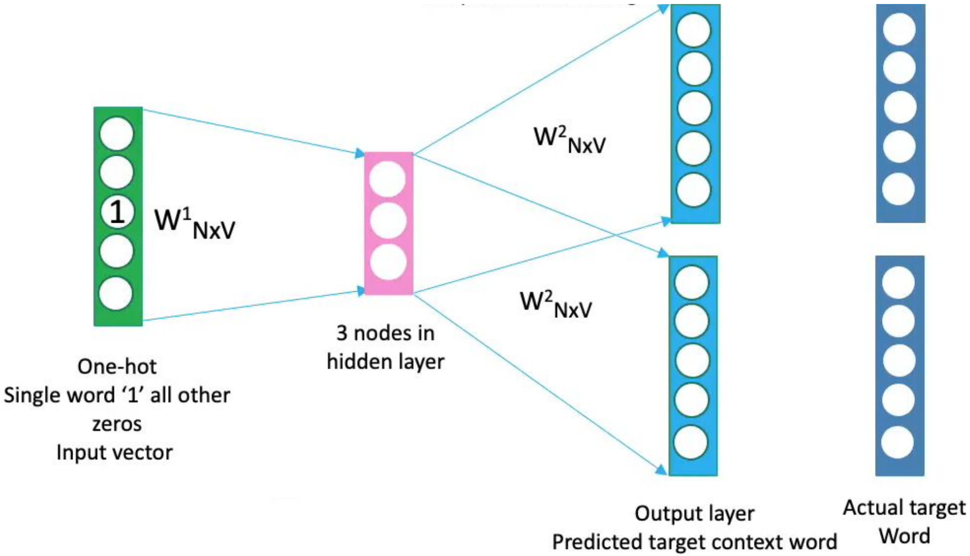
We trained 60 different word2vec models

Two word2vec architectures



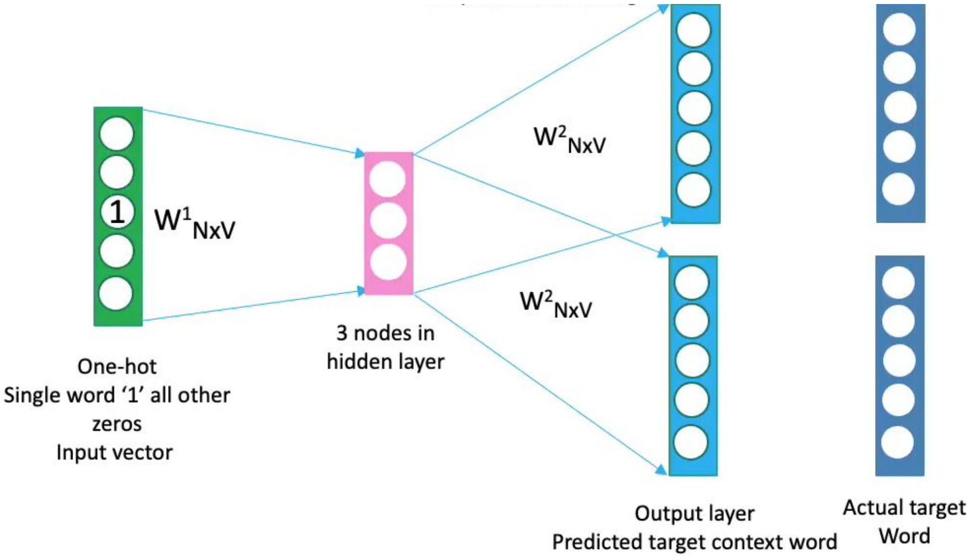
Skip gram architecture

optimize the objective of **output word being in the context of the input codon**.

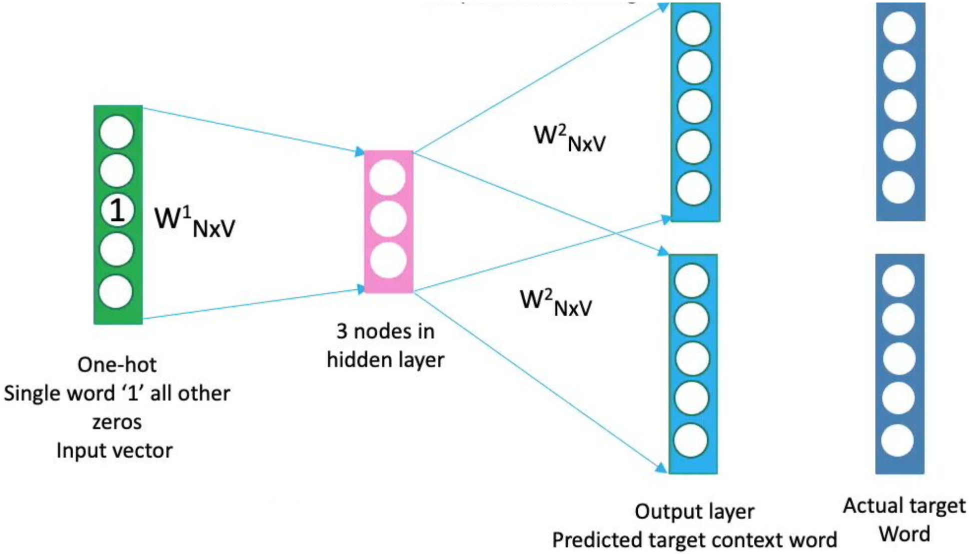


Training

UUCAACCACG AUU GCGCCGCUUU



Codon Embedding



Cosine similarity

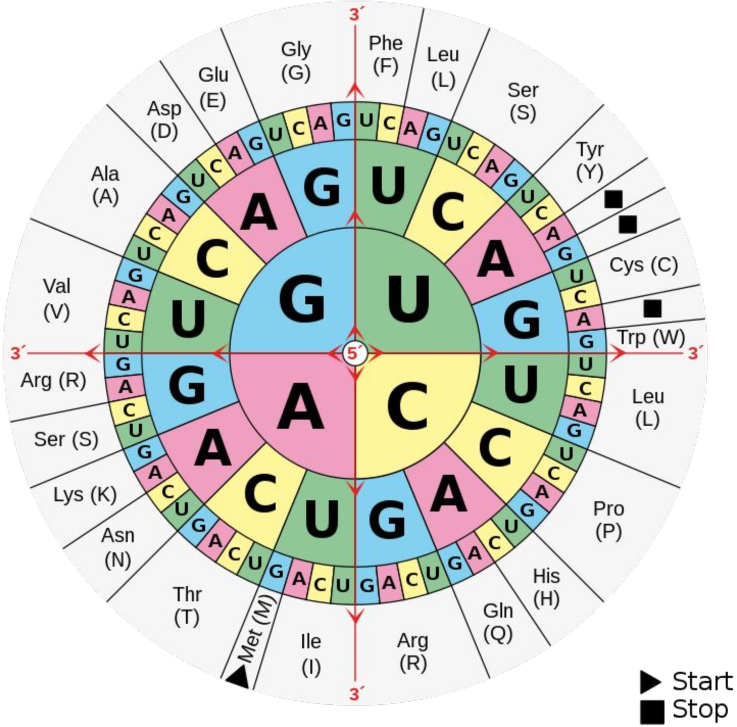
is a measure between vectorised codons.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2} \sqrt{\sum\limits_{i=1}^{n} B_i^2}}$$

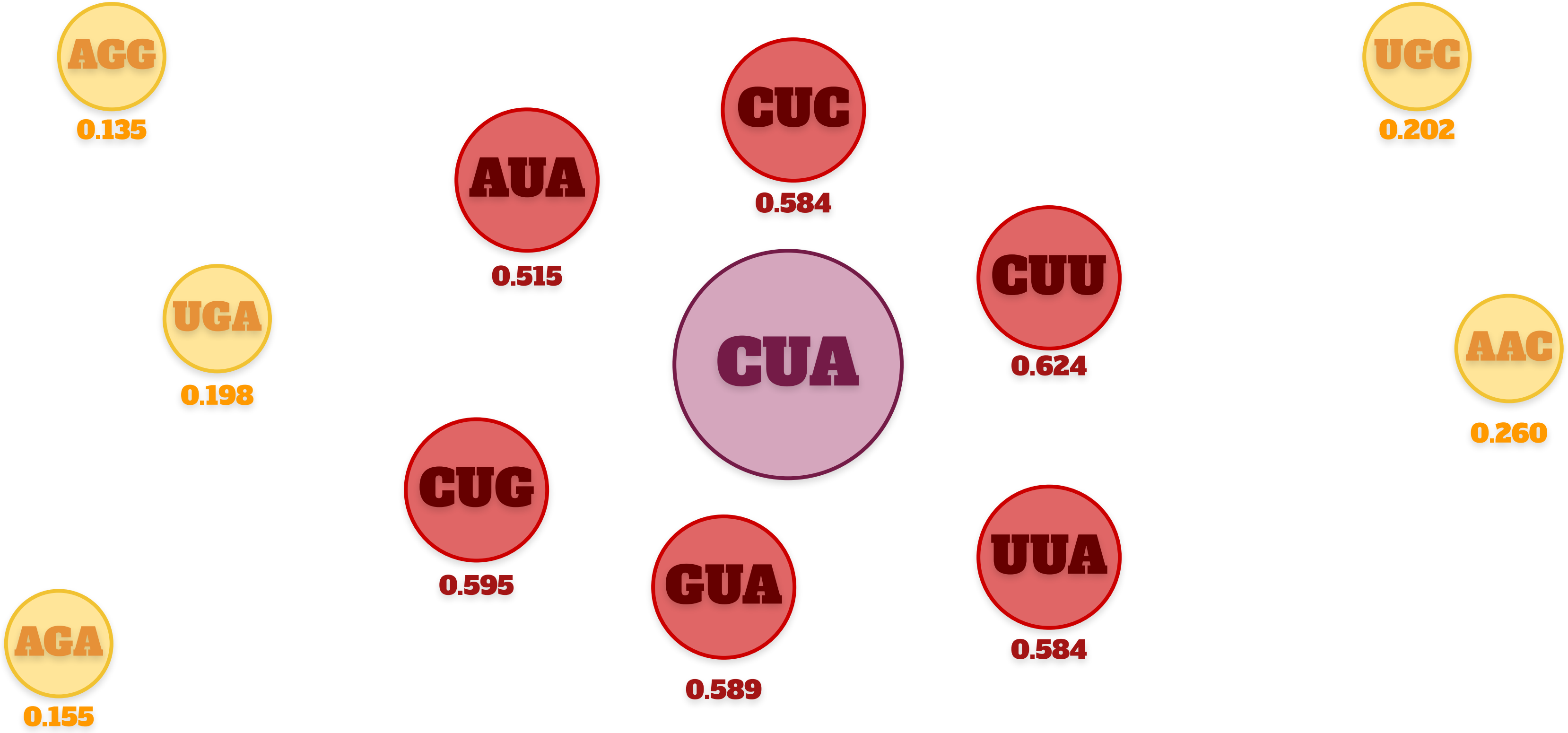
Cosine similarity

is a measure between vectorised codons.

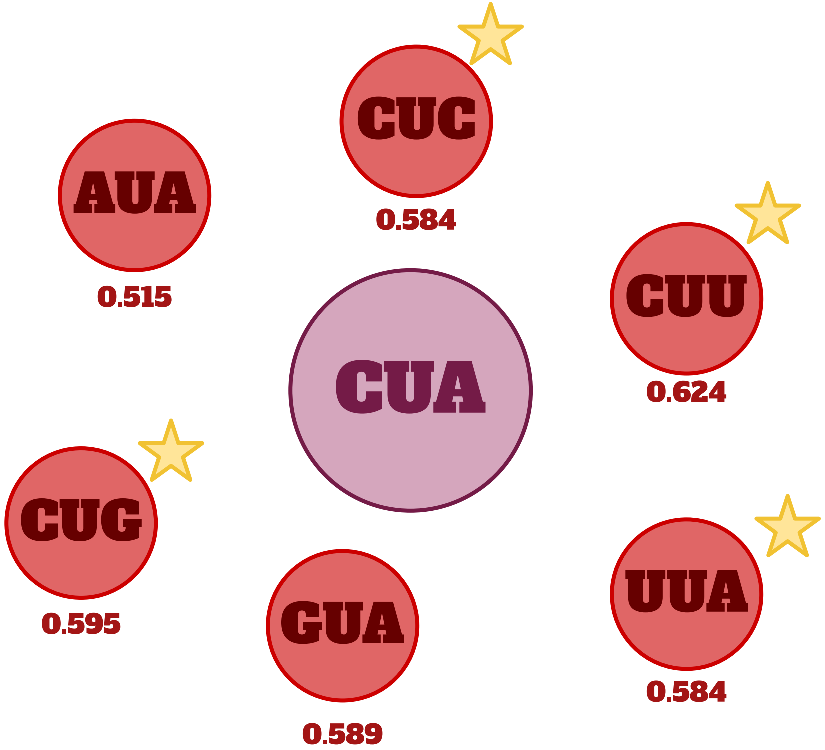
$$\cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}}$$



Score of the trained model?



$$\textbf{Recall} = \frac{\text{\# codons from the same group in top 6 neighbors}}{\text{total number of codons in the group}}$$



CUC ★ 0.584

AUA 0.515

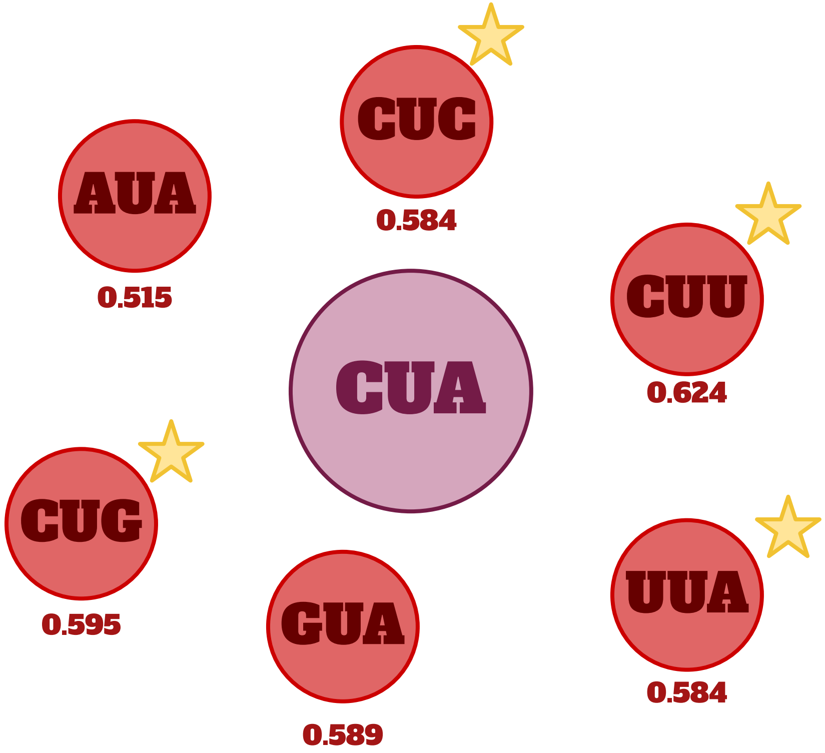
CUU ★ 0.624

CUA

CUG ★ 0.595

GUA 0.589

UUA ★ 0.584

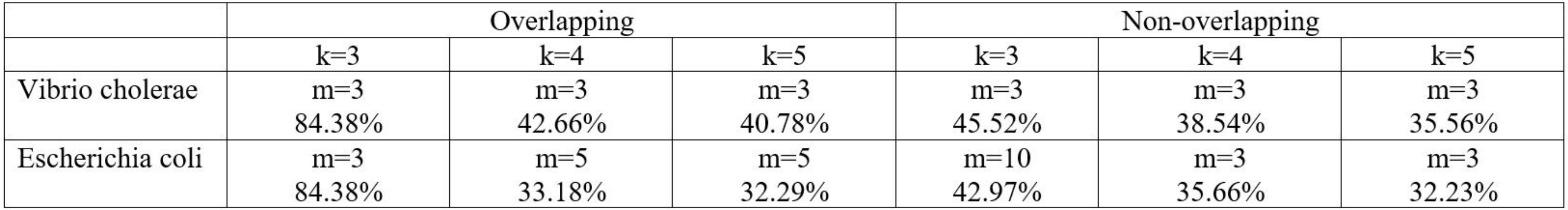


$$\text{Recall ('CUA')} = \frac{4}{5} = 80\%$$

1. Find recall for each codon
2. Calculate the average recall for each model
3. Compare the results

Results

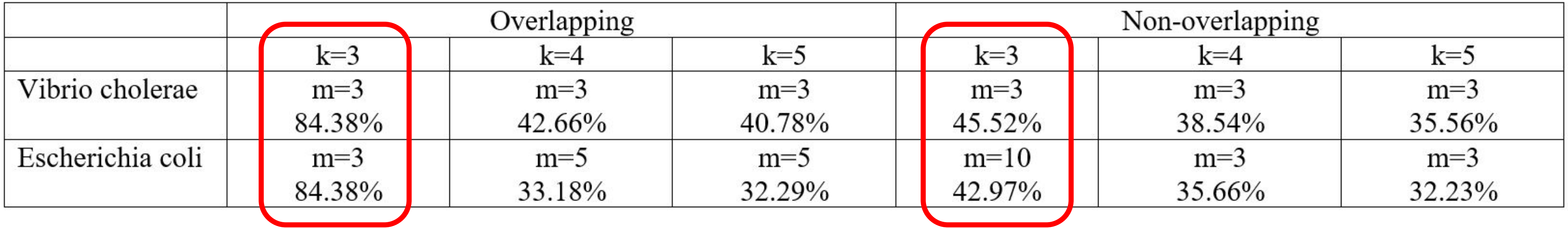
the best combinations of hyperparameters



	Overlapping			Non-overlapping		
	k=3	k=4	k=5	k=3	k=4	k=5
Vibrio cholerae	m=3 84.38%	m=3 42.66%	m=3 40.78%	m=3 45.52%	m=3 38.54%	m=3 35.56%
Escherichia coli	m=3 84.38%	m=5 33.18%	m=5 32.29%	m=10 42.97%	m=3 35.66%	m=3 32.23%

Results

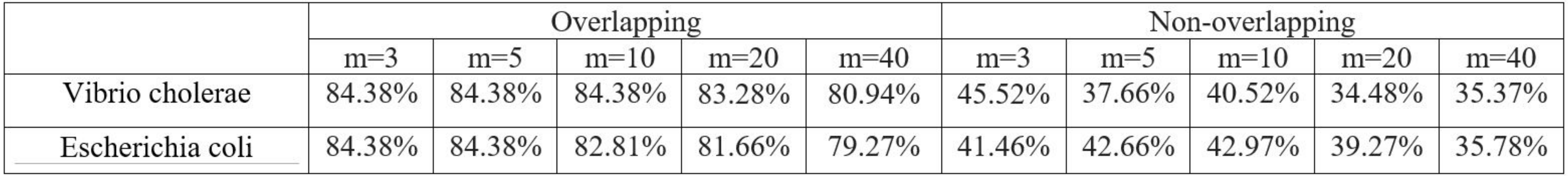
the best combinations of hyperparameters



	Overlapping			Non-overlapping		
	k=3	k=4	k=5	k=3	k=4	k=5
Vibrio cholerae	m=3 84.38%	m=3 42.66%	m=3 40.78%	m=3 45.52%	m=3 38.54%	m=3 35.56%
Escherichia coli	m=3 84.38%	m=5 33.18%	m=5 32.29%	m=10 42.97%	m=3 35.66%	m=3 32.23%

Results

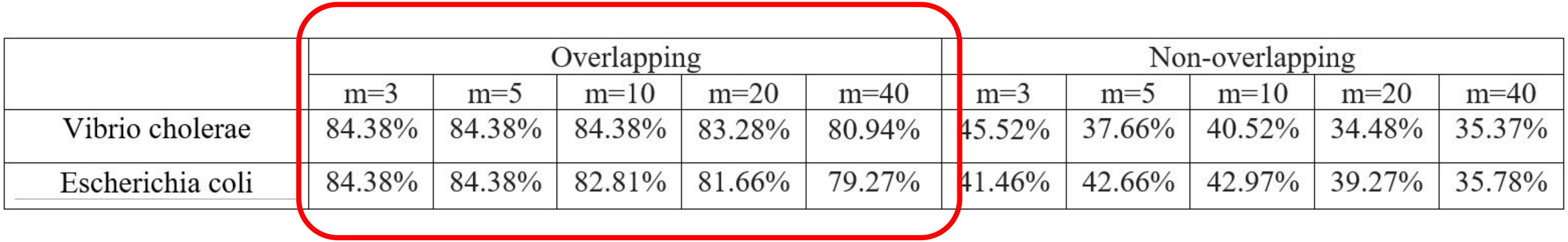
3-mers as neighbors



	Overlapping					Non-overlapping				
	m=3	m=5	m=10	m=20	m=40	m=3	m=5	m=10	m=20	m=40
Vibrio cholerae	84.38%	84.38%	84.38%	83.28%	80.94%	45.52%	37.66%	40.52%	34.48%	35.37%
Escherichia coli	84.38%	84.38%	82.81%	81.66%	79.27%	41.46%	42.66%	42.97%	39.27%	35.78%

Results

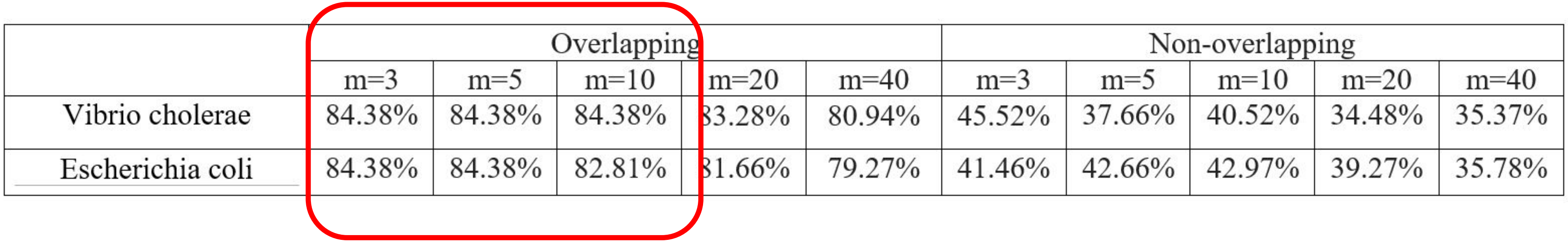
3-mers as neighbors, overlapping window



	Overlapping					Non-overlapping				
	m=3	m=5	m=10	m=20	m=40	m=3	m=5	m=10	m=20	m=40
Vibrio cholerae	84.38%	84.38%	84.38%	83.28%	80.94%	45.52%	37.66%	40.52%	34.48%	35.37%
Escherichia coli	84.38%	84.38%	82.81%	81.66%	79.27%	41.46%	42.66%	42.97%	39.27%	35.78%

Results

3-mers as neighbors, overlapping window and smaller contect size



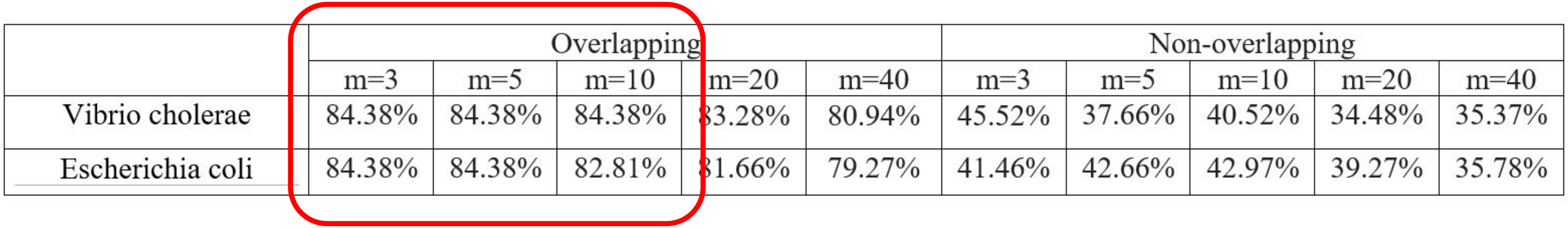
	Overlapping					Non-overlapping				
	m=3	m=5	m=10	m=20	m=40	m=3	m=5	m=10	m=20	m=40
Vibrio cholerae	84.38%	84.38%	84.38%	83.28%	80.94%	45.52%	37.66%	40.52%	34.48%	35.37%
Escherichia coli	84.38%	84.38%	82.81%	81.66%	79.27%	41.46%	42.66%	42.97%	39.27%	35.78%

Conclusion

Using:

- 3-mers as neighbors
- overlapping windows
- about 5 neighbors

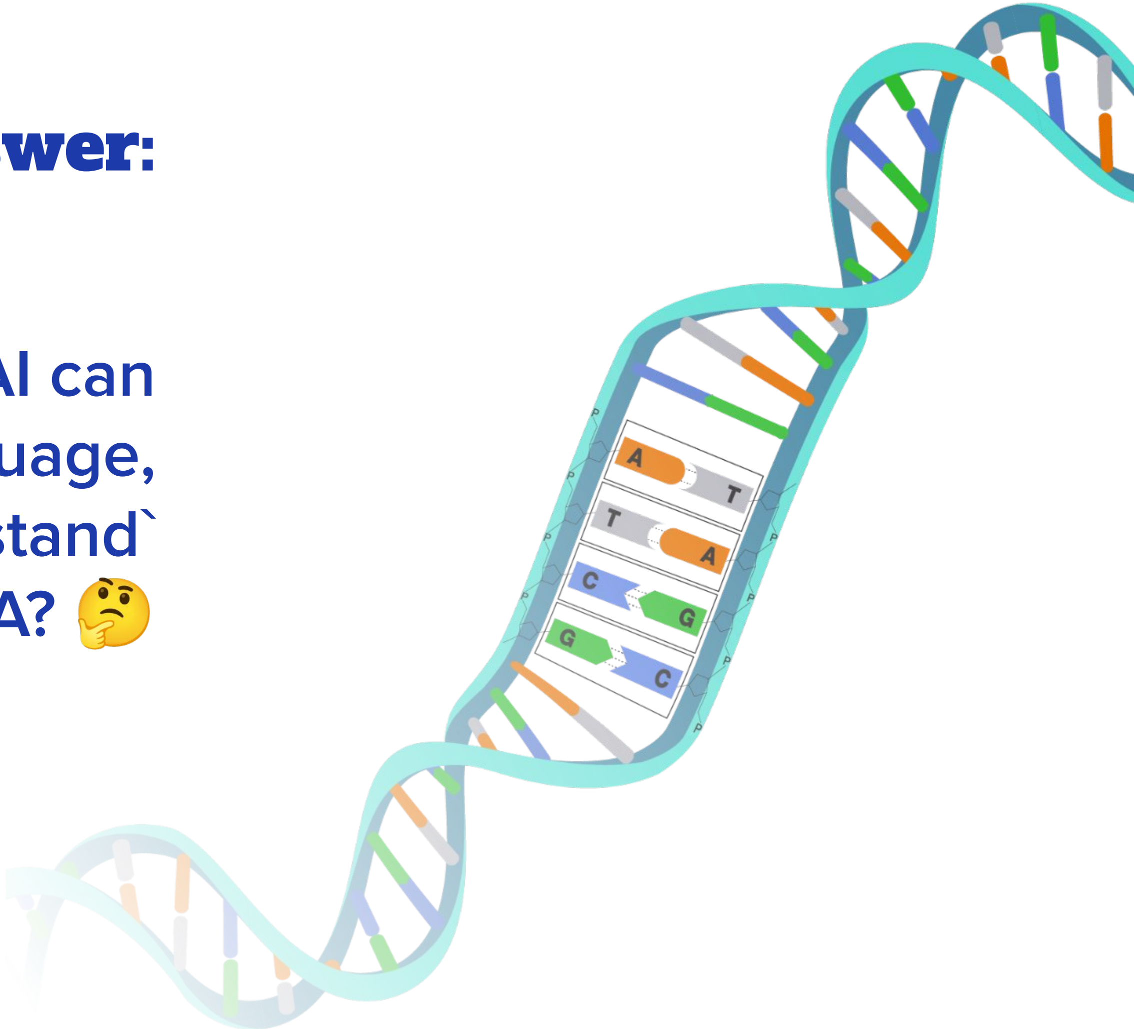
gave the highest similarity among vectorized codons from the same group.



	Overlapping					Non-overlapping				
	m=3	m=5	m=10	m=20	m=40	m=3	m=5	m=10	m=20	m=40
Vibrio cholerae	84.38%	84.38%	84.38%	83.28%	80.94%	45.52%	37.66%	40.52%	34.48%	35.37%
Escherichia coli	84.38%	84.38%	82.81%	81.66%	79.27%	41.46%	42.66%	42.97%	39.27%	35.78%

Answer:

AI can
`understand` language,
can it `understand`
DNA? 🤔



Conclusion

AI (word2vec) can learn the code written in DNA! 🎉

Future work

🧬 Train models on different organism’s genome

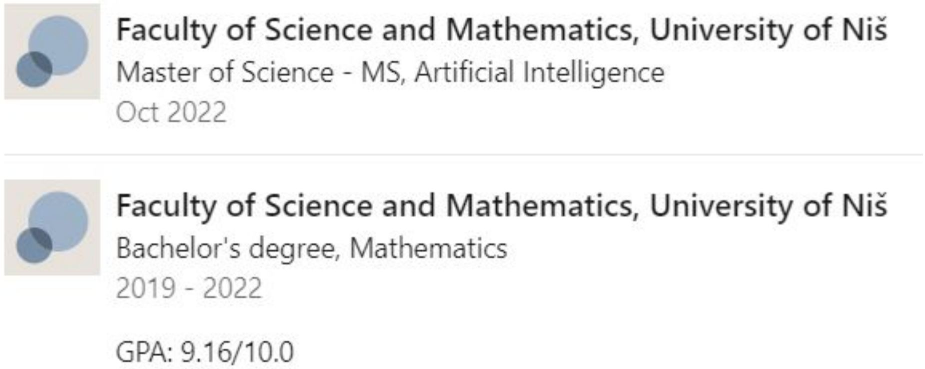
🧬 The best context may vary between organisms

Potential applications

🧬Gene function prediction

🧬 Detection of start and end of a gene

Potential collaboration?



Faculty of Science and Mathematics, University of Niš

Master of Science - MS, Artificial Intelligence

Oct 2022

Faculty of Science and Mathematics, University of Niš

Bachelor's degree, Mathematics

2019 - 2022

GPA: 9.16/10.0

Thank you

Question, comment?

📧 andjadenic@gmail.com