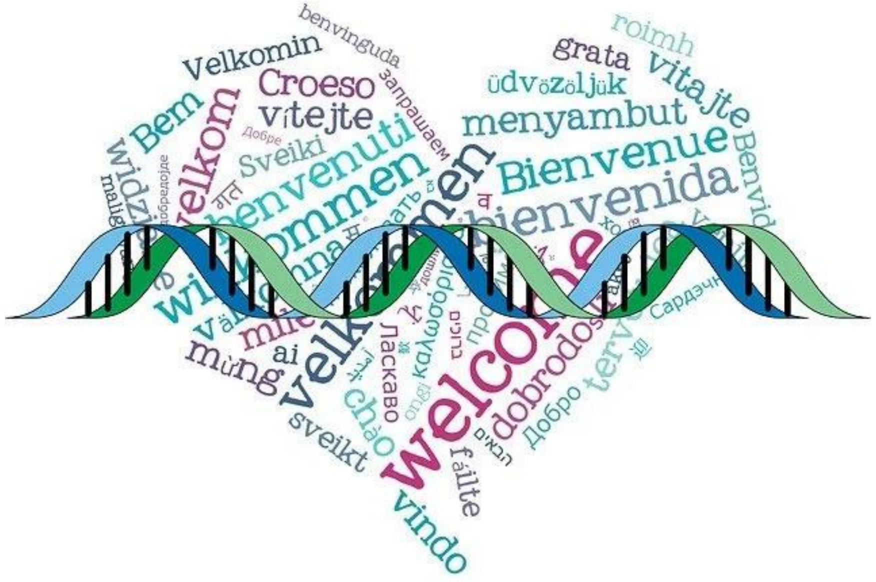
The Language of DNA

Can Machines Understand DNA Like English?



Justin Jia

Brinkman Lab

Sept 21, 2022

Natural Language

- Any language that has evolved naturally in humans through use and repetition without conscious planning or premeditation

Alphabet

A B C D
E F G H
I J K L
M N O P
Q R S T
U V W X
Y Z

Vowels

A, E, I, O, U

Consonant

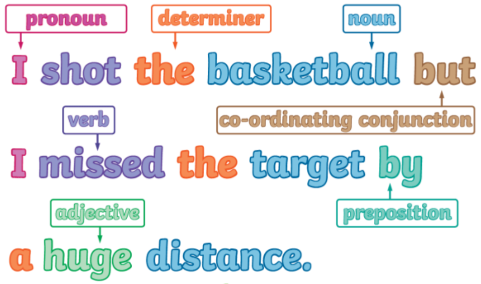
B, C, D, F, G, H, J, K, L, M, N, P, Q, R, S, T, V, W, X, Y, Z

Vocabulary

- Smile
- Skillful
- Skilled
- Simple
- Seemly
- Secure
- Perfect
- Paradise
- Lucky
- Lucid
- Lovely
- Light
- Honored
- Honorable
- Honest
- Heavenly
- Hearty
- Healthy

verb
Verbs are doing or action words. (This includes 'to be' and 'to have').
run
jump
talk
noun
dice
girl
pirate
ronoun
that
it
ne
she
these
you
preposition
Prepositions link nouns, pronouns and phrases.
across
of
on
against
throughout
adverb
verbs, adjectives, or other verbs.
he ate slowly
junction
link words, phrases and clauses.
yet
nor
or
and

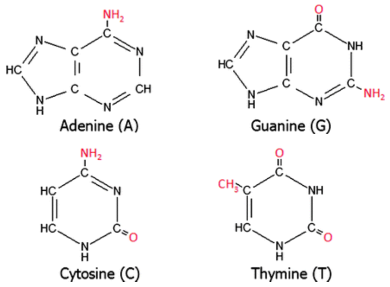
Sentence



I shot the basketball but I missed the target by a huge distance.

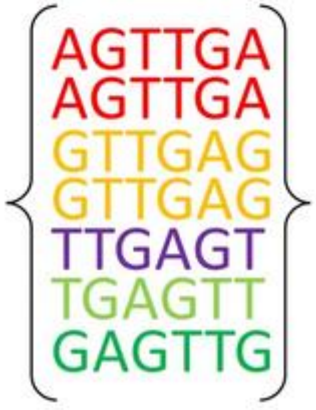
The Language of DNA

Alphabet



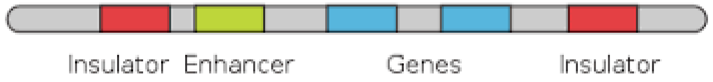
Purines/Pyrimidines

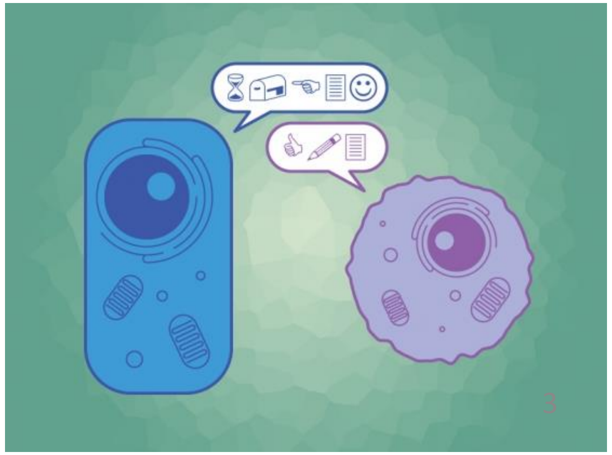
Vocabulary



Recognition sites
Promoters
Start codons
Etc…

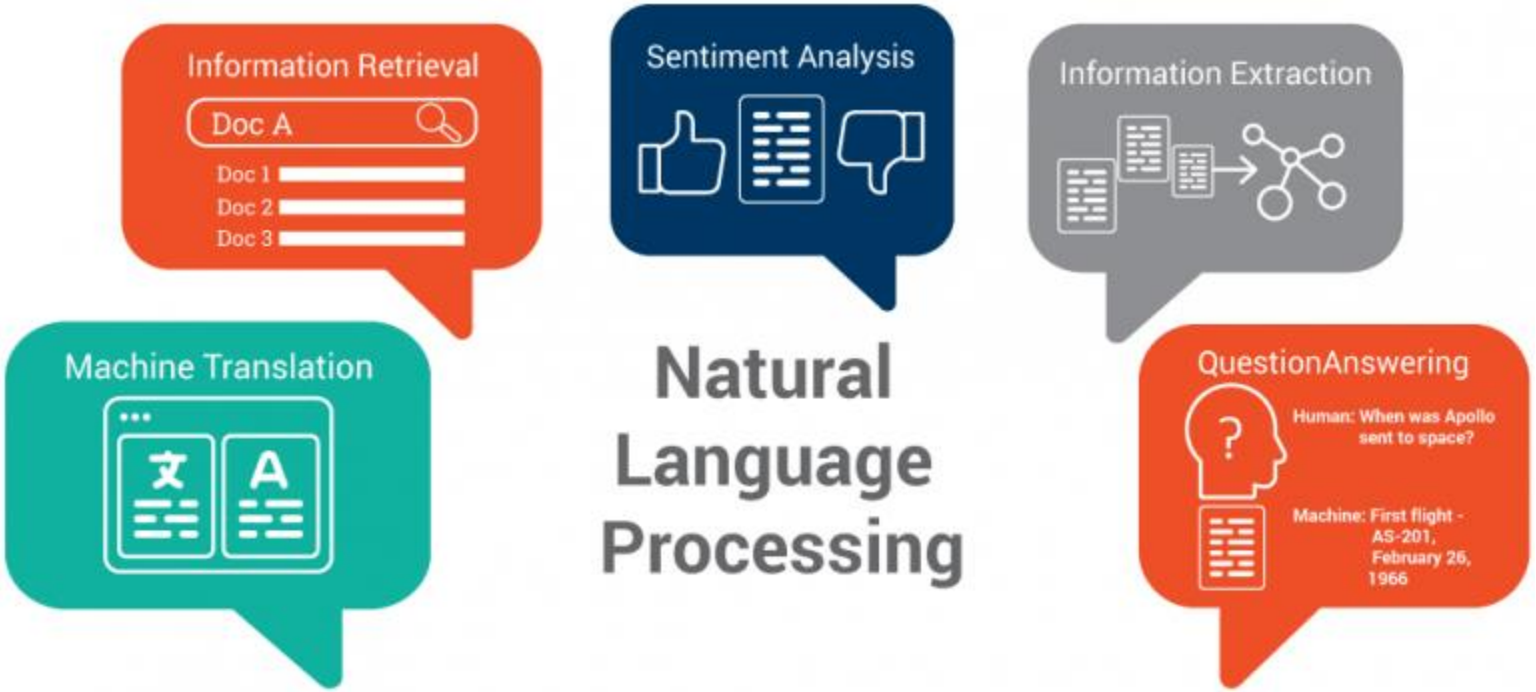
Sentence



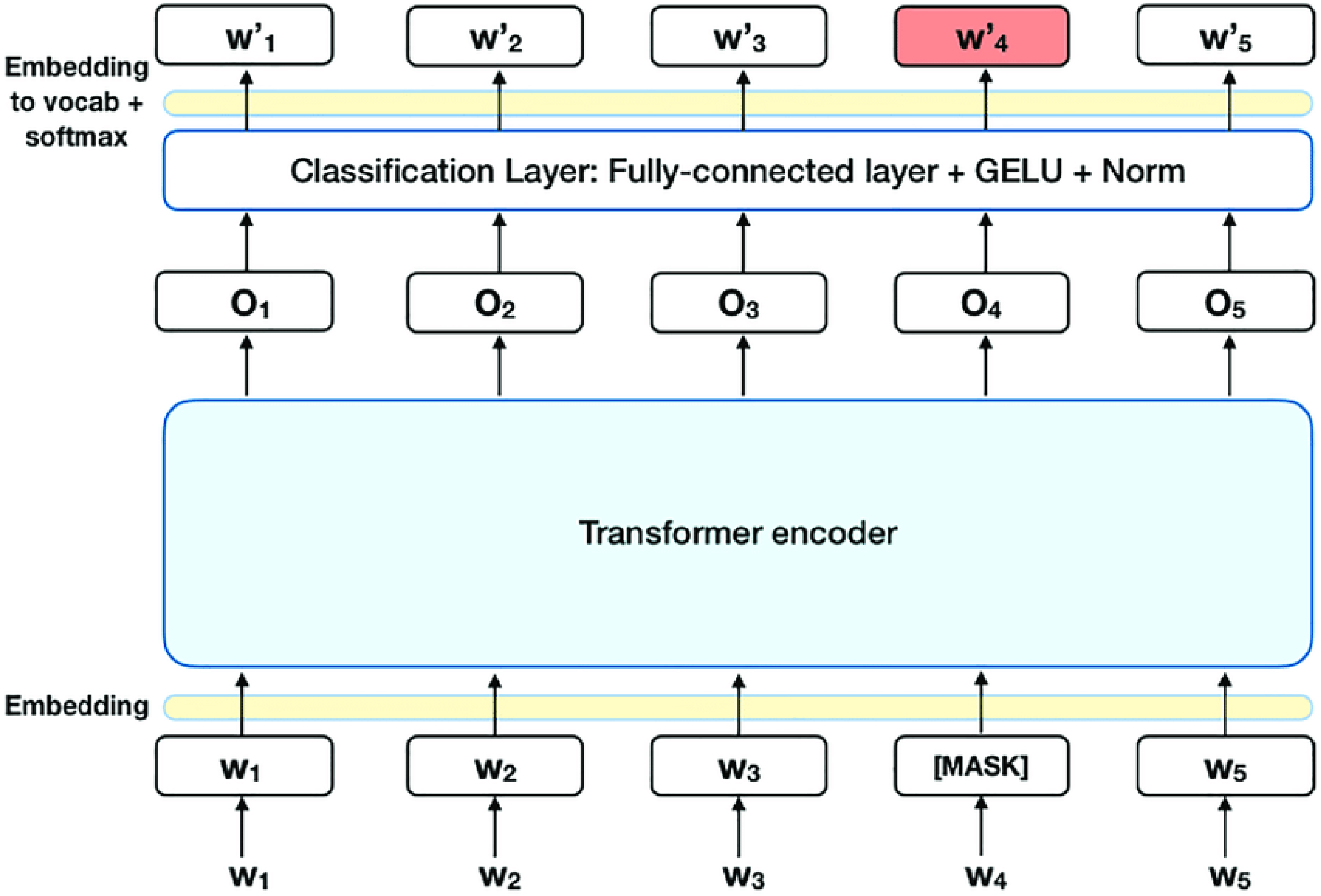


Natural Language Processing (NLP)

- Programming computers to process and analyze large amounts of natural language data.
- The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them.



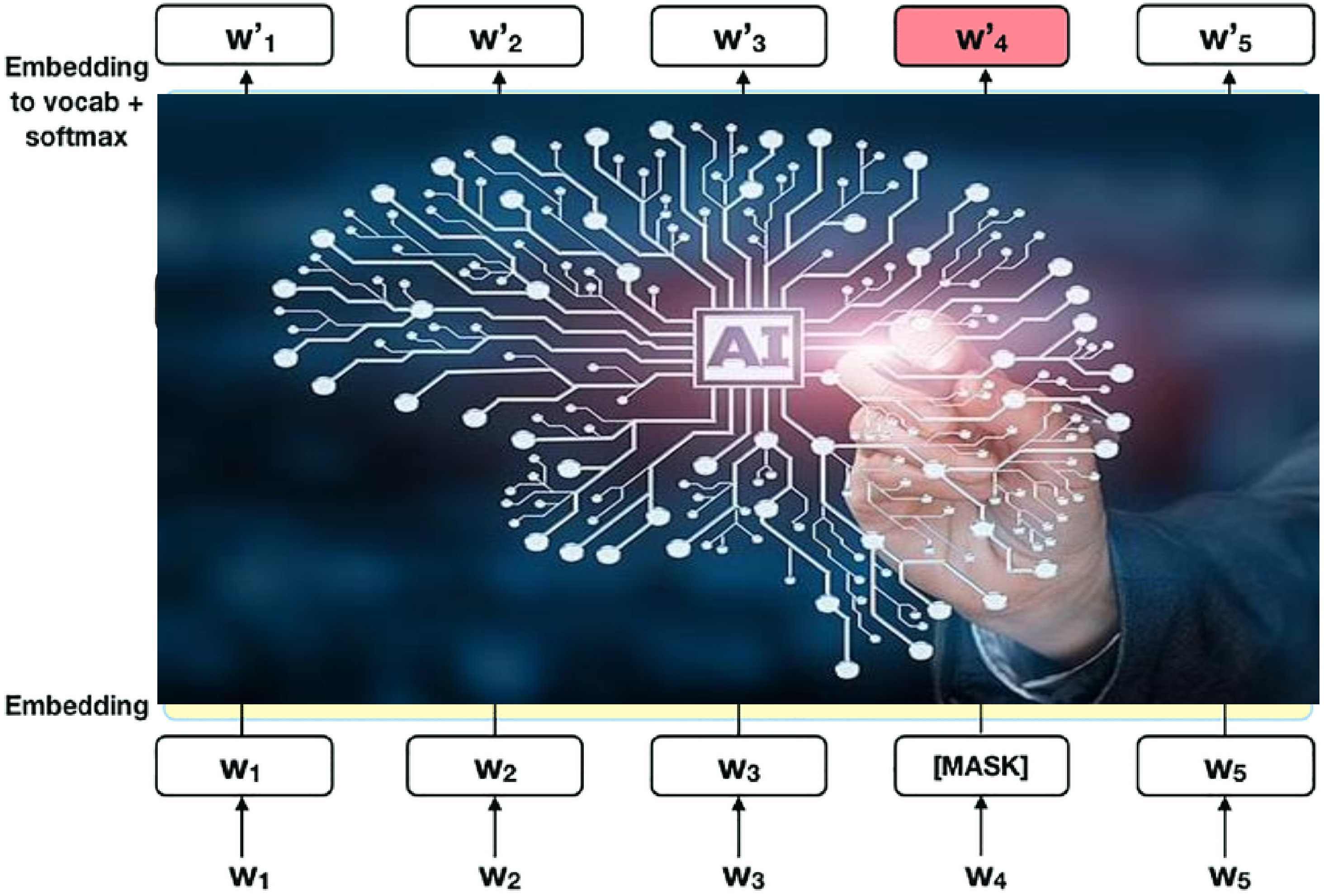
BERT: Bidirectional Encoder Representation from Transformers


w'1
w'2
w'3
w'4
w'5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
w1
w2
w3
[MASK]
w5
w1
w2
w3
w4
w5

- 2019 Google's research on natural language processing

- Self-learning of natural languages
- Learns contextual relations between words in a text.

BERT: Bidirectional Encoder Representation from Transformers



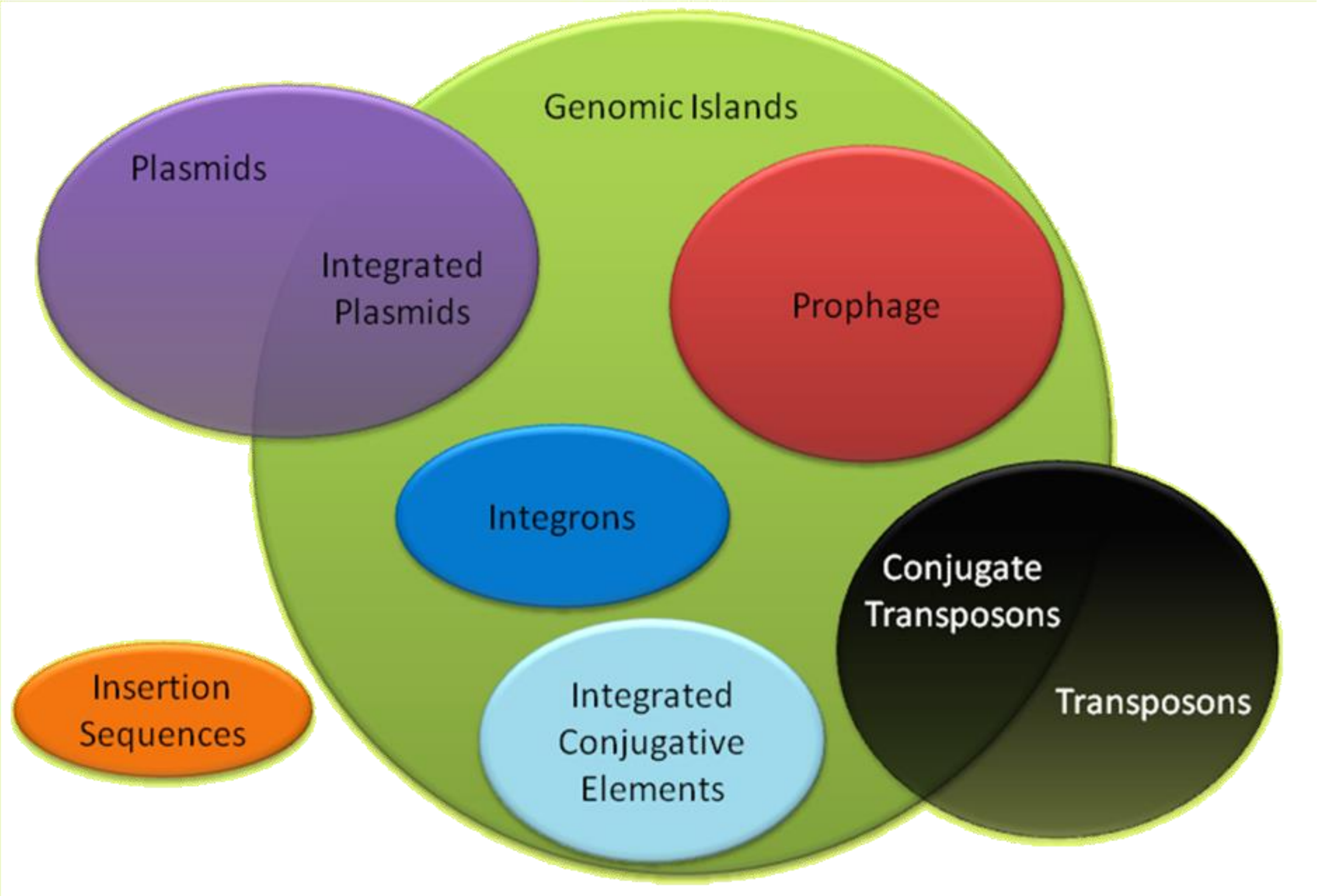
- 2019 Google's research on natural language processing

- Self-learning of natural languages
- Learns contextual relations between words in a text.

If machines can utilize NLP for human languages, can the same process be used for DNA?

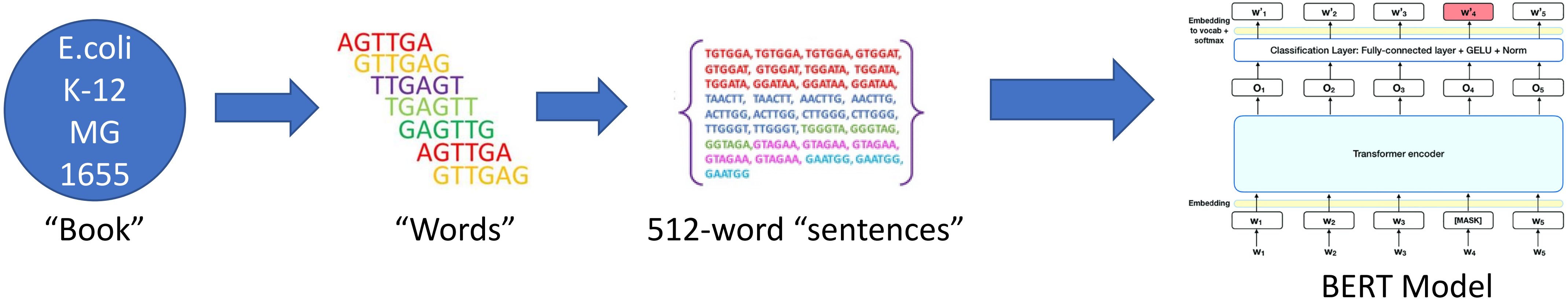
Genomic Islands (GIs)

- Large segments (>8KB) of a genome that has evidence of horizontal origins



Teaching the BERT Model the DNA Language

Step 1: Generalized self-learning of Bacterial DNA “words”:



“Book”

“Words”

TGTGGA, TGTGGA, TGTGGA, GTGGAT, GTGGAT, GTGGAT, TGGATA, TGGATA, TGGATA, GGATAA, GGATAA, GGATAA, TAACTT, TAACTT, AACTTG, AACTTG, ACTTGG, ACTTGG, CTTGGG, CTTGGG, TTGGGT, TTGGGT, TGGGTA, GGGTAG, GGTAGA,GTAGAA, GTAGAA, GTAGAA, GTAGAA, GTAGAA, GAATGG, GAATGG, GAATGG

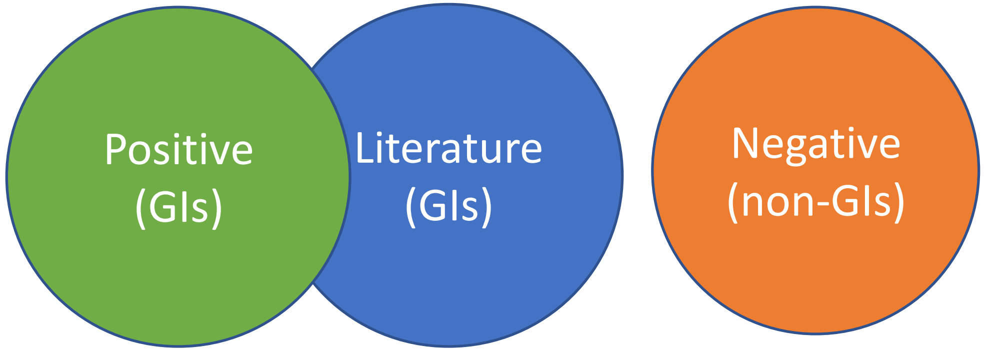
512-word “sentences”

w’1
w’2
w’3
w’4
w’5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
w1
w2
w3
[MASK]
w5
w1
w2
w3
w4
w5

BERT Model

Predicting Genomic Islands

Bertelli et al. 2018. Bioinformatics.

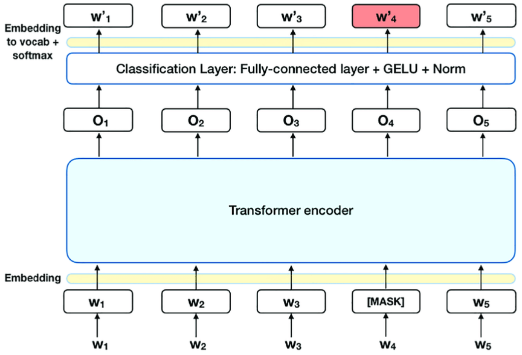


Step 2: fine-tuning to predict GI:

Training:
80 GIs (all) from Literature Dataset +
80 non-GIs from Negative Dataset (Same Accession)

TGTGGA, TGTGGA, TGTGGA, GTGGAT, GTGGAT, GTGGAT, TGGATA, TGGATA, TGGATA, GGATAA, GGATAA, GGATAA, TAACTT, TAACTT, AACTTG, AACTTG, ACTTGG, ACTTGG, CTTGGG, CTTGGG, TTGGGT, TTGGGT, TGGGTA, GGGTAG, GGTAGA, GTAGAA, GTAGAA, GTAGAA, GTAGAA, GTAGAA, GAATGG, GAATGG, GAATGG

3003-word “sentences”

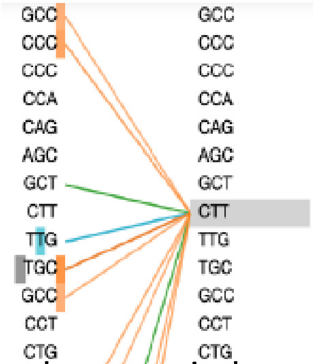

Embedding to vocab + softmax
w'₁
w'₂
w'₃
w'₄
w'₅
Classification Layer: Fully-connected layer + GELU + Norm
O₁
O₂
O₃
O₄
O₅
Transformer encoder
Embedding
w₁
w₂
w₃
[MASK]
w₅
w₁
w₂
w₃
w₄
w₅

BERT Model from Step 1

Evaluation:
9000 x “3003-word sentences”
Randomly chosen from Positive + Negative dataset (non training accessions)

Never seen to model*

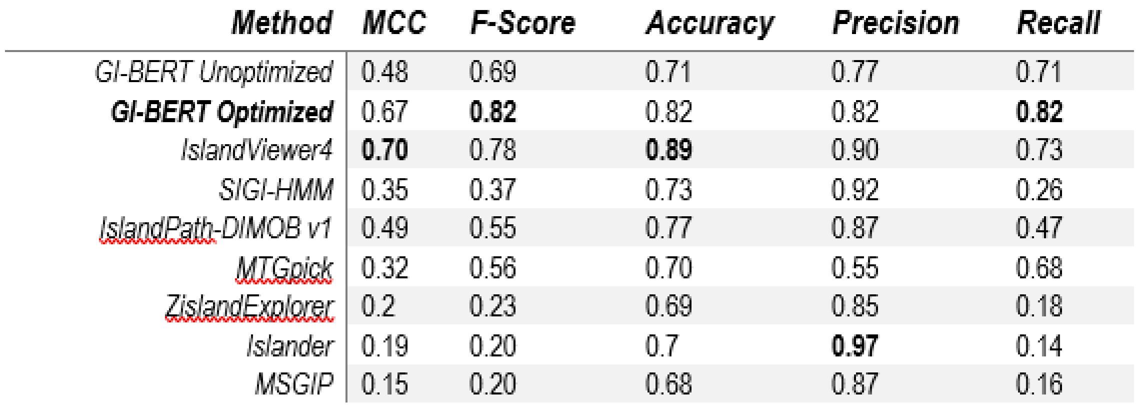
Visualize



Example

Performance - Preliminary

Yes. NLP works on DNA



Method	*MCC*	*F-Score*	*Accuracy*	*Precision*	*Recall*
GI-BERT Unoptimized	0.48	0.69	0.71	0.77	0.71
GI-BERT Optimized	0.67	**0.82**	0.82	0.82	**0.82**
IslandViewer4	**0.70**	0.78	**0.89**	0.90	0.73
SIGI-HMM	0.35	0.37	0.73	0.92	0.26
IslandPath-DIMOB v1	0.49	0.55	0.77	0.87	0.47
MTGpick	0.32	0.56	0.70	0.55	0.68
ZislandExplorer	0.2	0.23	0.69	0.85	0.18
Islander	0.19	0.20	0.7	**0.97**	0.14
MSGIP	0.15	0.20	0.68	0.87	0.16

Why does it work?

- Mostly picking up patterns
  - E.g. repetitive sequences that flank GIs
  - E.g. presence of certain gene sequences in a GI.

- Need in depth evaluation.

Acknowledgements

- Brinkman lab

- Omar Nassif

Fisheries and Oceans
Canada

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

NSERC
CRSNG

GenomeCanada

Genome
BritishColumbia

GenomeAtlantic

SFU SIMON FRASER
UNIVERSITY

Graduate Studies in BIOINFORMATICS at UBC AND SFU