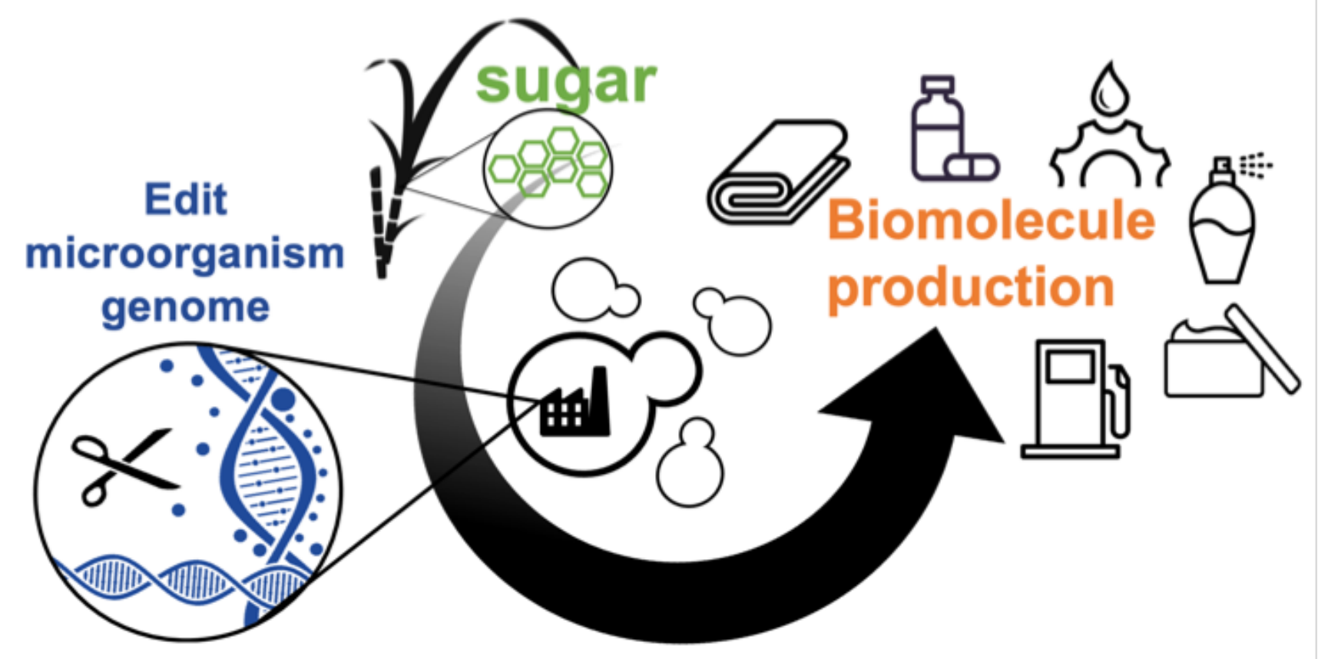
Applying NLP to Nature's Language (DNA!)

An examination of perplexity in the genetic grammar of microorganisms

Erin H. Wilson
The Paul G. Allen School of Computer Science and Engineering at the University of Washington
CSE 517 – Winter 2019

Motivation: sustainable molecule production



Globally, human societies are consuming finite resources at unsustainable rates. Transitioning away from our dependencies on non-renewable resources and towards a cyclical, sustainable use of natural products is critical for preserving Earth's most threatened ecosystems. An alternative way to source many natural products is to engineer microorganisms into **biological molecule factories**[1].

How?

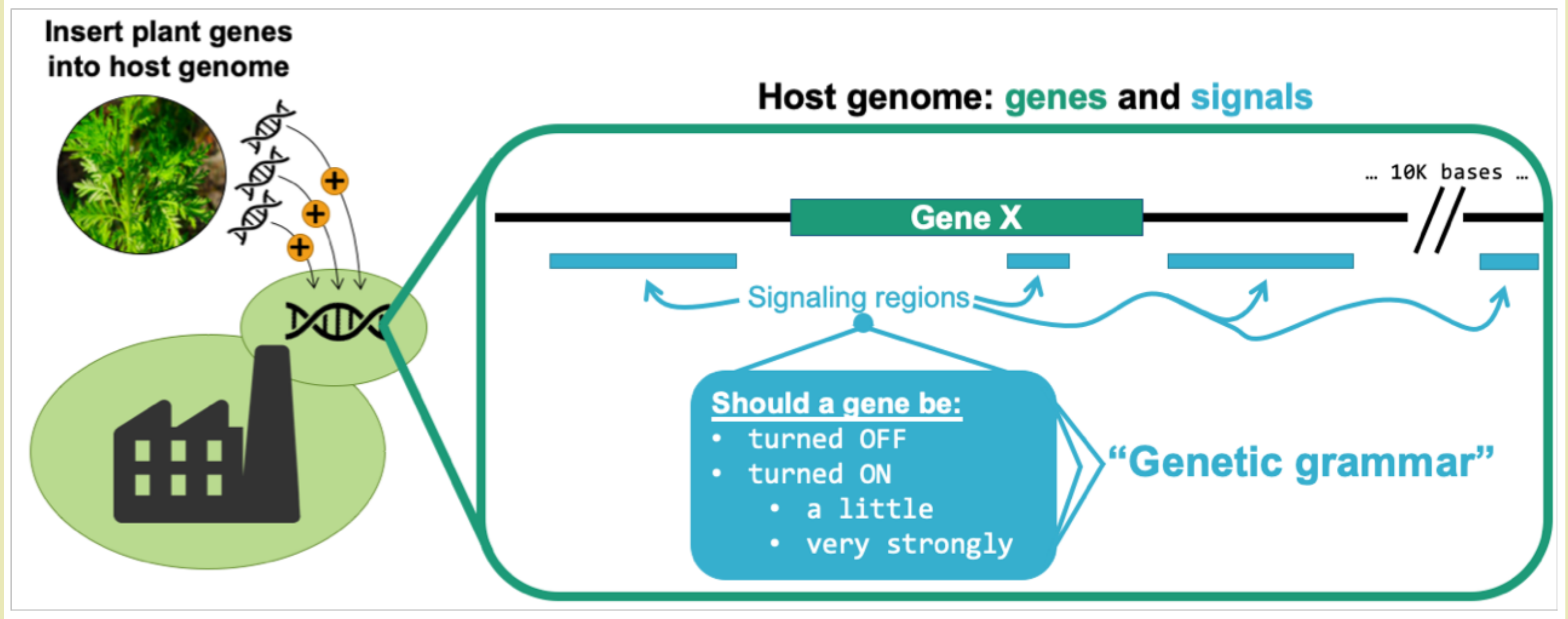
- Introduce foreign plant genes into microorganism
- Organism consumes renewable sugarcane
- Reroute carbon to new target molecule (naturally produced in plant)

The Challenge

- Every organism has a distinct "genetic grammar"
- It is difficult to express foreign genes with host microorganism's grammar
- Low gene expression leads to inefficient molecule production

Gene regulation: a cell's genetic grammar

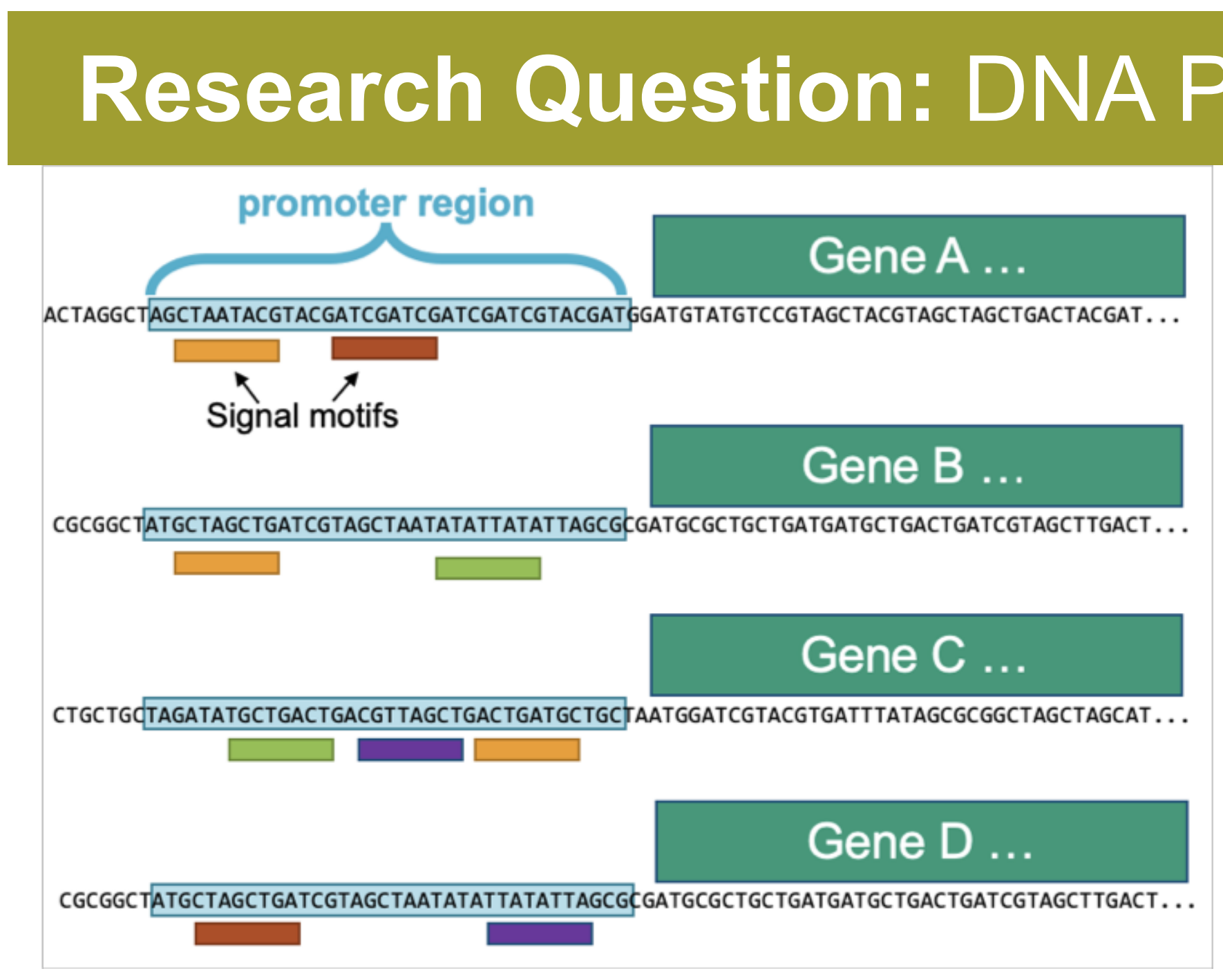
- In addition to genes that code for proteins, genomes contain **signaling regions**
- Signaling regions help **control** which genes are turned ON or OFF ("expressed")
- These regions contain a sort of "grammar": patterns in the DNA sequence that guide a cell to turn specific genes on in specific situations ("gene regulation")


Insert plant genes into host genome
Host genome: genes and signals
… 10K bases …
Gene X
Signaling regions
"Genetic grammar"

Should a gene be:

- turned OFF
- turned ON
  - a little
  - very strongly

Research Question: DNA Perplexity?



```
ACTAGGCTAGCTAATACGTACGATCGATCGATCGATCGTACGATGGATGTATGTCCGTAGCTACGTAGCTAGCTGACTACGAT...
CGCGGCTATGCTAGCTGATCGTAGCTAATATATTATATTAGCGCGATGCGCTGCTGATGATGCTGACTGATCGTAGCTTGACT...
CTGCTGCTAGATATGCTGACTGACGTTAGCTGACTGATGCTGCTAATGGATCGTACGTGATTTATAGCGCGGCTAGCTAGCAT...
CGCGGCTATGCTAGCTGATCGTAGCTAATATATTATATTAGCGCGATGCGCTGCTGATGATGCTGACTGATCGTAGCTTGACT...
```

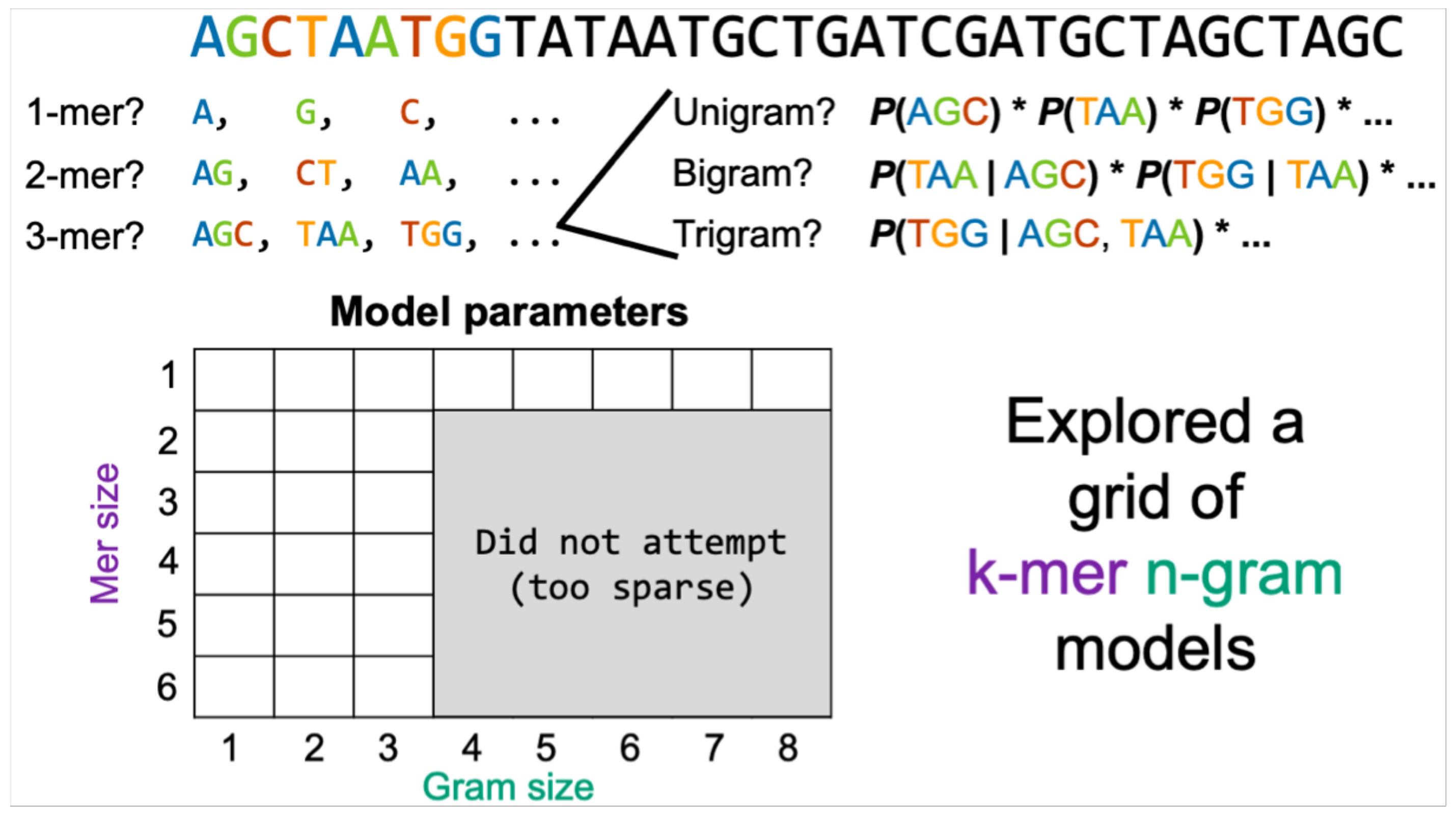
- We know generally where signaling regions are, but the exact signals and what they mean are still not well understood
- The promoter region (the area right before a gene starts) contains shorter **signal motifs**
- Motifs can be combined and rearranged in many configurations
- Modeling the **regulatory grammar** is very complex

Can **NLP language modeling** capture regulatory patterns in DNA sequences?

Do we observe **lower perplexity** in signaling regions?

Convert DNA into "words"

- Regulatory DNA sequences don't have a natural word boundary
- Test various model parameters for how to break a DNA sequence into words



AGCTAATGGTATAATGCTGATCGATGCTAGCTAGC

1-mer? A, G, C, ...
2-mer? AG, CT, AA, ...
3-mer? AGC, TAA, TGG, ...

Unigram? $P(AGC) * P(TAA) * P(TGG) * ...$
Bigram? $P(TAA \mid AGC) * P(TGG \mid TAA) * ...$
Trigram? $P(TGG \mid AGC, TAA) * ...$

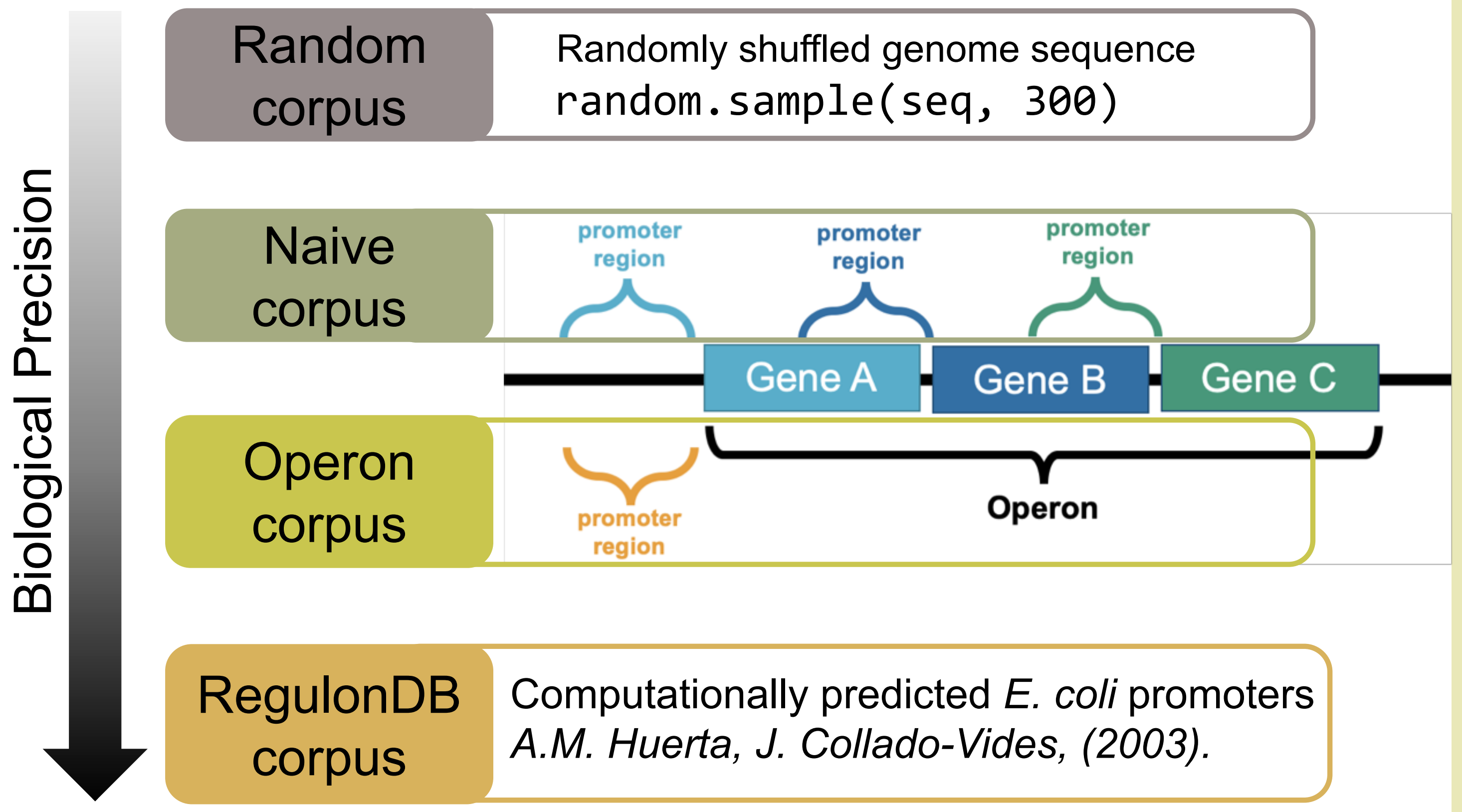
Model parameters

Mer size: 1, 2, 3, 4, 5, 6
Gram size: 1, 2, 3, 4, 5, 6, 7, 8

Did not attempt (too sparse)

Explored a grid of k-mer n-gram models

Generate a "corpus" of promoters



Random corpus: Randomly shuffled genome sequence `random.sample(seq, 300)`

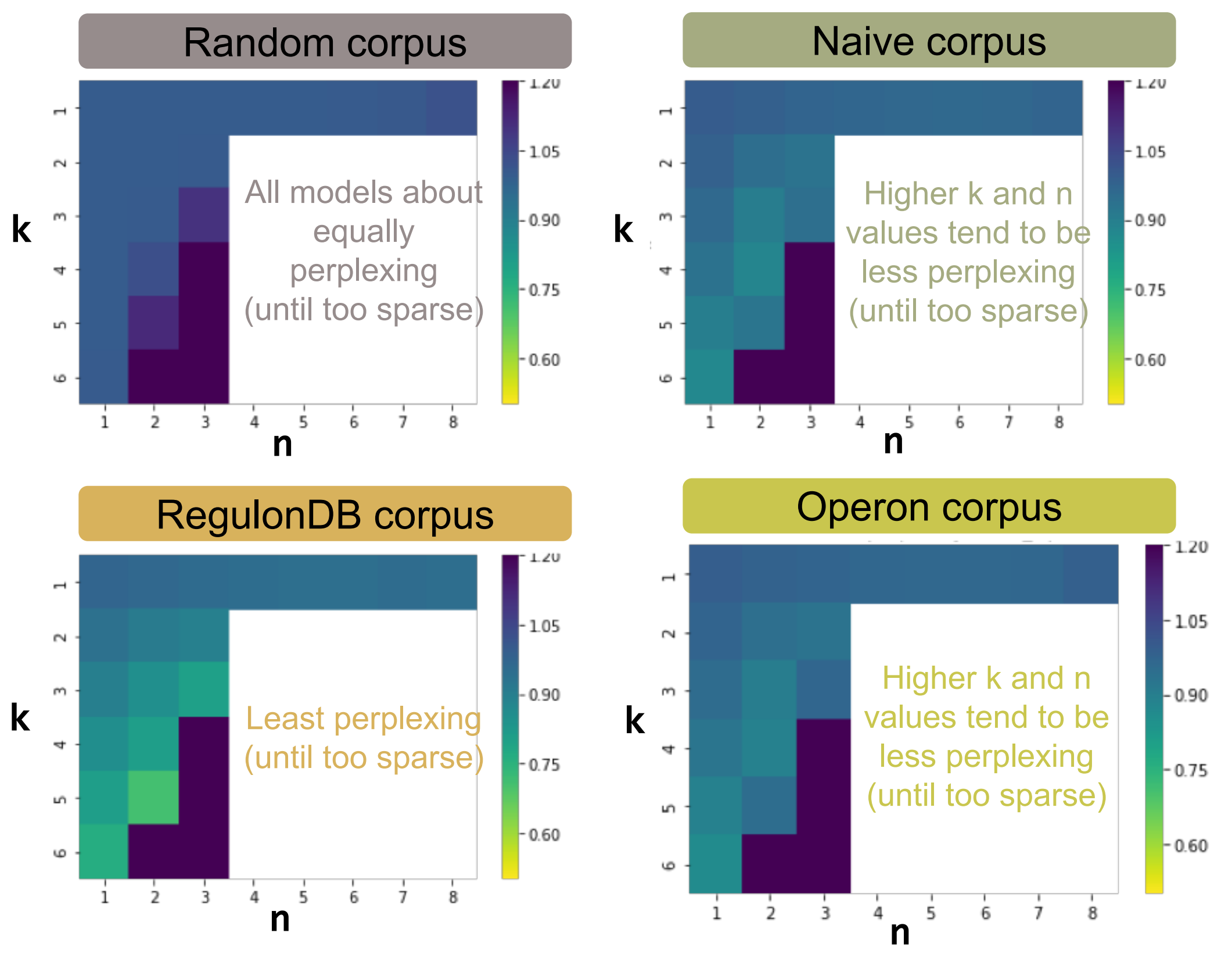
Naive corpus

Operon corpus

RegulonDB corpus: Computationally predicted *E. coli* promoters *A.M. Huerta, J. Collado-Vides, (2003).*

Results: Perplexity decreases with sequence prior context and biological precision

Normalized Perplexity across k-mer n-gram grid



Random corpus: All models about equally perplexing (until too sparse)

Naive corpus: Higher k and n values tend to be less perplexing (until too sparse)

RegulonDB corpus: Least perplexing (until too sparse)

Operon corpus: Higher k and n values tend to be less perplexing (until too sparse)

Perplexity:

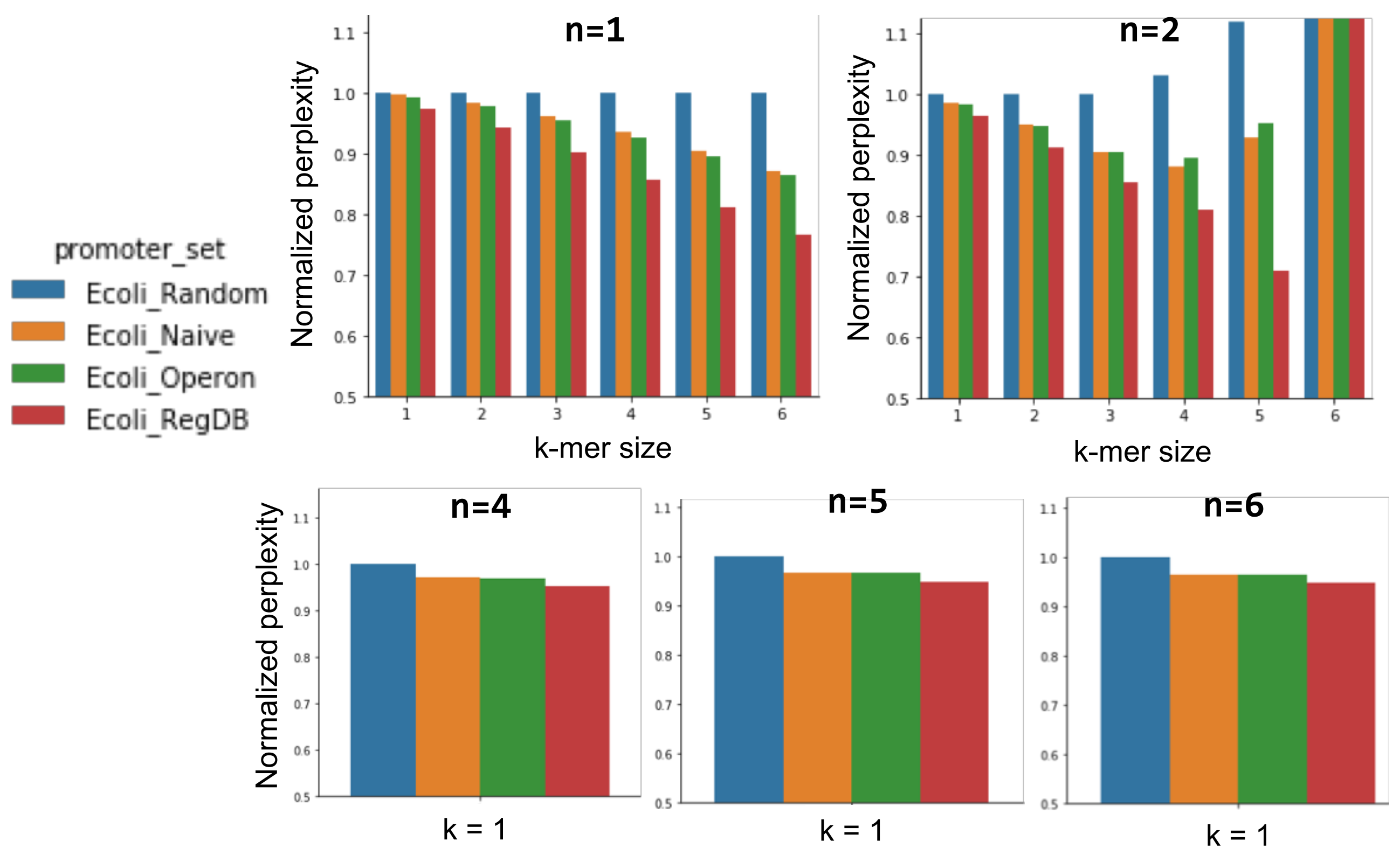
$$2^{-l} \text{ where } l = \frac{1}{M}\sum_{i=1}^{m}\log p(s_i)$$

Normalized Perplexity:

$$\frac{\texttt{test\_pplex}}{\texttt{worst\_possible\_pplex}}$$

- close to 1: as perplexed as can be
- 1.2: infinite perplexity

Normalized Perplexity between corpora

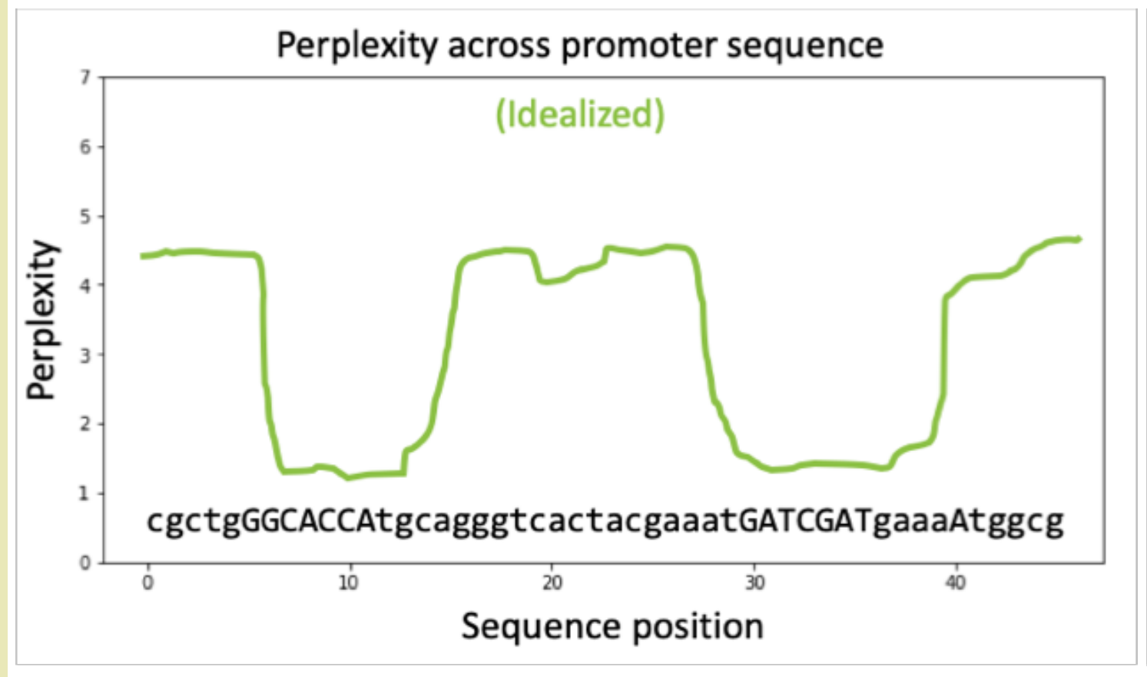


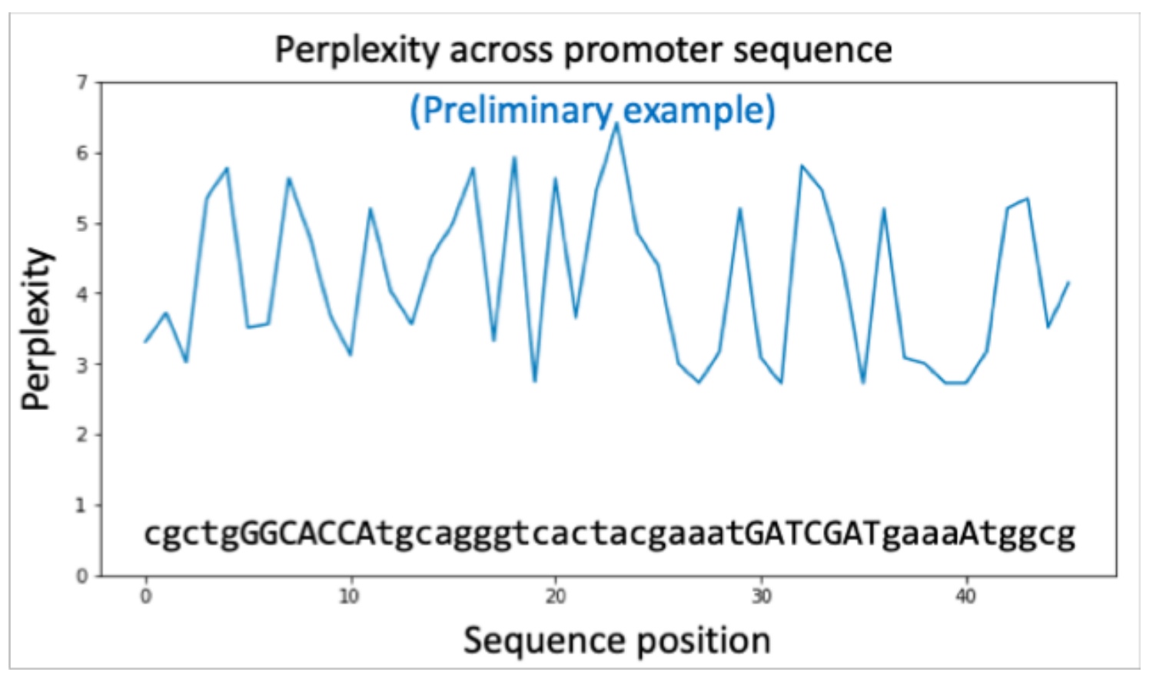
Discussion: promoters are less perplexing

- The data show that **random** sequences are all about **equally perplexing**, regardless of the choice of k-mer or n-grams.
- Corpora of actual promoter sequences show **lower perplexity than random** and that models which use **more biological context** (higher values of k and n) decrease perplexity until the counts become too sparse
- Corpora of actual promoter sequences generally show that **perplexity decreases with increased biological precision** of the promoter set
- This all suggests that language modeling can indeed **capture some underlying pattern** in promoter regions!
- However, the discrepancy between model performance on the operon corpus and the RegulonDB corpus suggests that this approach **needs more refinement** before applying to other microorganisms without a "ground truth database" available.

Future work: perplexity across a sequence

- If the language model has captured the grammar well, there hypothetically should be low perplexity regions in the sequence which represent the "true promoter signal"
- Can we **map low perplexity regions** to important regulatory signals?





References

- [1] A. Meadows et al. (2016) "Rewriting yeast central carbon metabolism for industrial isoprenoid production." *Nature*.
- [2] A.M. Huerta, J. Collado-Vides. (2003) "Computational Prediction of Promoters in the Escherichia coli genome." *J Mol Biol.*
- [3] J. T. Cuperus et al. (2017) "Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences." *Genome Research.*

`