Genome 540 discussion

January 28th, 2025

Joe Min

Genome Sciences
UNIVERSITY OF WASHINGTON

Agenda

Homework 3

Information theory in Wordle

Homework 3

Overview

1. Parse a genbank file (.gbff) and…
    a. Extract all CDS features
    b. Read in the sequence
2. Build a site model for translation start sites (TSS)
    a. Use CDS features to get nucleotide frequencies +/- 10bp around all TSS (21bp total including TSS)
    b. Use sequence to get nucleotide frequencies throughout the genome *on both strands*
    c. Compute the weights using the log2 ratios of the frequencies
3. Use the site model to compute scores at
    a. Every annotated TSS
    b. The entire genome (21bp window) on both strands

Genbank Flat File

Header

```
LOCUS       U00096               4641652 bp    DNA     circular BCT 01-AUG-2014
DEFINITION  Escherichia coli str. K-12 substr. MG1655, complete genome.
ACCESSION   U00096
VERSION     U00096.3
DBLINK      BioProject: PRJNA225
            BioSample: SAMN02604091
KEYWORDS    .
SOURCE      Escherichia coli str. K-12 substr. MG1655
  ORGANISM  Escherichia coli str. K-12 substr. MG1655
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (bases 1 to 4641652)
  AUTHORS   Blattner,F.R., Plunkett,G. III, Bloch,C.A., Perna,N.T., Burland,V.,
            Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F.,
            Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J.,
            Mau,B. and Shao,Y.
  TITLE     The complete genome sequence of Escherichia coli K-12
  JOURNAL   Science 277 (5331), 1453-1462 (1997)
   PUBMED   9278503
REFERENCE   2  (bases 1 to 4641652)
  AUTHORS   Hayashi,K., Morooka,N., Yamamoto,Y., Fujita,K., Isono,K., Choi,S.,
            Ohtsubo,E., Baba,T., Wanner,B.L., Mori,H. and Horiuchi,T.
  TITLE     Highly accurate genome sequences of Escherichia coli K-12 strains
            MG1655 and W3110
  JOURNAL   Mol. Syst. Biol. 2, 2006 (2006)
   PUBMED   16738553
REFERENCE   3  (bases 1 to 4641652)
  AUTHORS   Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R.,
            Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T.,
            Mori,H., Perna,N.T., Plunkett,G. III, Rudd,K.E., Serres,M.H.,
            Thomas,G.H., Thomson,N.R., Wishart,D. and Wanner,B.L.
```

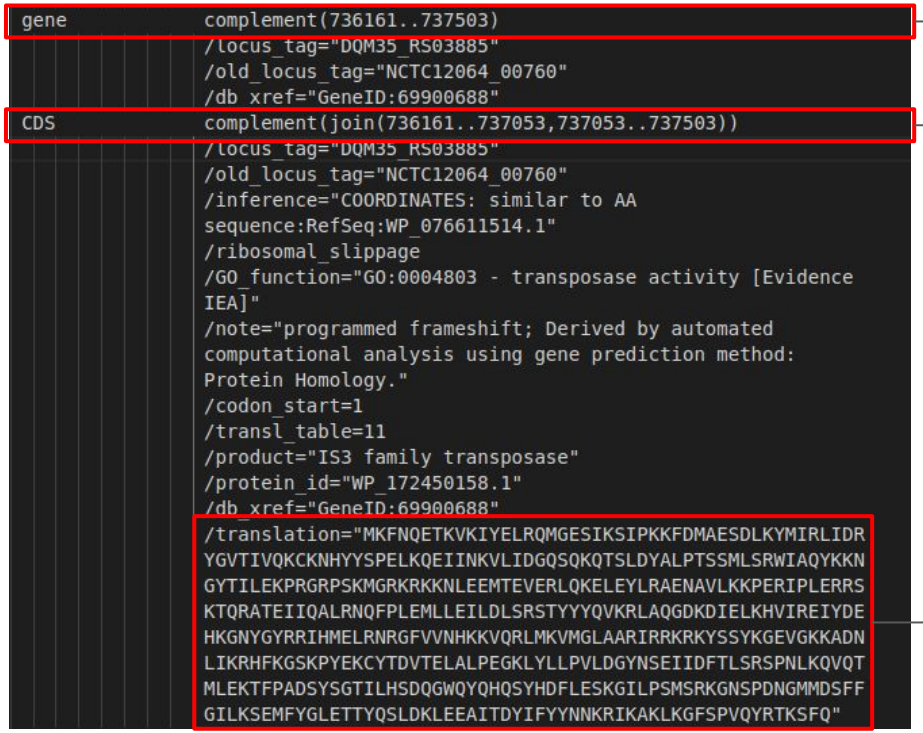
Features

```
FEATURES             Location/Qualifiers
     source          1..4641652
                     /organism="Escherichia coli str. K-12 substr. MG1655"
                     /mol_type="genomic DNA"
                     /strain="K-12"
                     /sub_strain="MG1655"
                     /db_xref="taxon:511145"
     gene            190..255
                     /gene="thrL"
                     /locus_tag="b0001"
                     /gene_synonym="ECK0001"
                     /gene_synonym="JW4367"
                     /db_xref="EcoGene:EG11277"
     CDS             190..255
                     /gene="thrL"
                     /locus_tag="b0001"
                     /gene_synonym="ECK0001"
                     /gene_synonym="JW4367"
                     /function="leader; Amino acid biosynthesis: Threonine"
                     /note="GO_process: GO:0009088 - threonine biosynthetic
                     process"
                     /codon_start=1
                     /transl_table=11
                     /product="thr operon leader peptide"
                     /protein_id="AAC73112.1"
                     /db_xref="ASAP:ABE-0000006"
                     /db_xref="UniProtKB/Swiss-Prot:P0AD86"
                     /db_xref="EcoGene:EG11277"
                     /translation="MKRISTTITTTITITTGNGAG"
```

Sequence

```
ORIGIN
        1 agcttttcat tctgactgca acgggcaata tgtctctgtg tggattaaaa aaagagtgtc
       61 tgatagcagc ttctgaactg gttacctgcc gtgagtaaat taaaatttta ttgacttagg
      121 tcactaaata ctttaaccaa tataggcata gcgcacagac agataaaaat tacagagtac
      181 acaacatcca tgaaacgcat tagcaccacc attaccacca ccatcaccat taccacaggt
      241 aacggtgcgg gctgacgcgt acaggaaaca cagaaaaaag cccgcacctg acagtgcggg
      301 ctttttttt cgaccaaagg taacgaggta acaaccatgc gagtgttgaa gttcggcggt
      361 acatcagtgg caaatgcaga acgttttctg cgtgttgccg atattctgga aagcaatgcc
      421 aggcaggggc aggtggccac cgtcctctct gcccccgcca aaatcaccaa ccacctggtg
      481 gcgatgattg aaaaaaccat tagcggccag gatgctttac ccaatatcag cgatgccgaa
      541 cgtatttttg ccgaactttt gacgggactc gccgccgccc agccggggtt cccgctggcg
      601 caattgaaaa ctttcgtcga tcaggaattt gcccaaataa aacatgtcct gcatggcatt
      661 agtttgttgg ggcagtgccc ggatagcatc aacgctgcgc tgatttgccg tggcgagaaa
      721 atgtcgatcg ccattatggc cggcgtatta gaagcgcgcg gtcacaacgt tactgttatc
      781 gatccggtcg aaaaactgct ggcagtgggg cattacctcg aatctaccgt cgatattgct
      841 gagtccaccc gccgtattgc ggcaagccgc attccggctg atcacatggt gctgatggca
      901 ggtttcaccg ccggtaatga aaaaggcgaa ctggtggtgc ttggacgcaa cggttccgac
      961 tactctgctg cggtgctggc tgcctgttta cgcgccgatt gttgcgagat ttggacggac
     1021 gttgacgggg tctatacctg cgacccgcgt caggtgcccg atgcgaggtt gttgaagtcg
     1081 atgtcctacc aggaagcgat ggagctttcc tacttcggcg ctaaagttct tcacccccgc
     1141 accattaccc ccatcgcccca gttccagatc ccttgcctga ttaaaaatac cggaaatcct
     1201 caagcaccag gtacgctcat tggtgccagc cgtgatgaag acgaattacc ggtcaagggc
     1261 atttccaatc tgaataacat ggcaatgttc agcgtttctg gtccggggat gaaagggatg
     1321 gtcggcatgg cggcgcgcgt ctttgcagcg atgtcacgcg cccgtatttc cgtggtgctg
     1381 attacgcaat catcttccga atacagcatc agtttctgcg ttccacaaag cgactgtgtg
     1441 cgagctgaac gggcaatgca ggaagagttc tacctggaac tgaaagaagg cttactggag
     1501 ccgctggcag tgacggaacg gctggccatt atctcggtgg taggtgatgg tatgcgcacc
     1561 ttgcgtgggа tctcggcgaa attctttgcc gcactggccc gcgccaatat caacattgtc
```

gbff Features



Gene + introns

Strand + exons

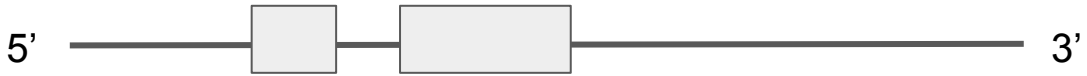
- 736161..737053
  - Specifies a coding region
- join(...)
  - Join coding sequences
- complement(...)
  - Take the reverse complement

Peptide product
Warning: may not match sequence

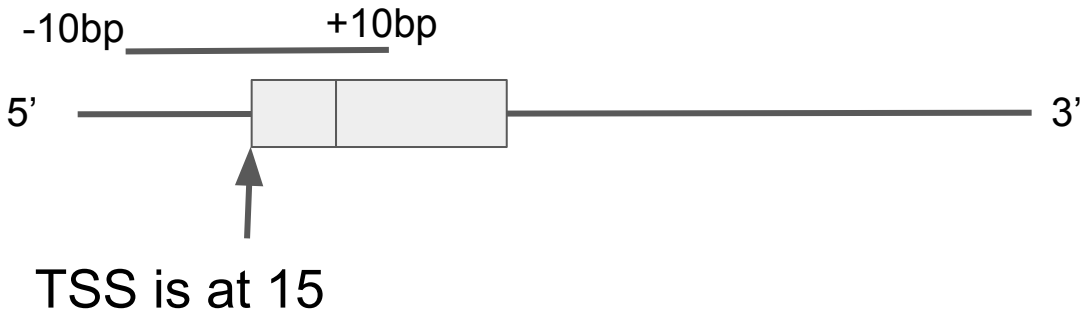
join(...) example

Example: join(15..20,25..35)

15..20,25..35



join(15..20,25..35)

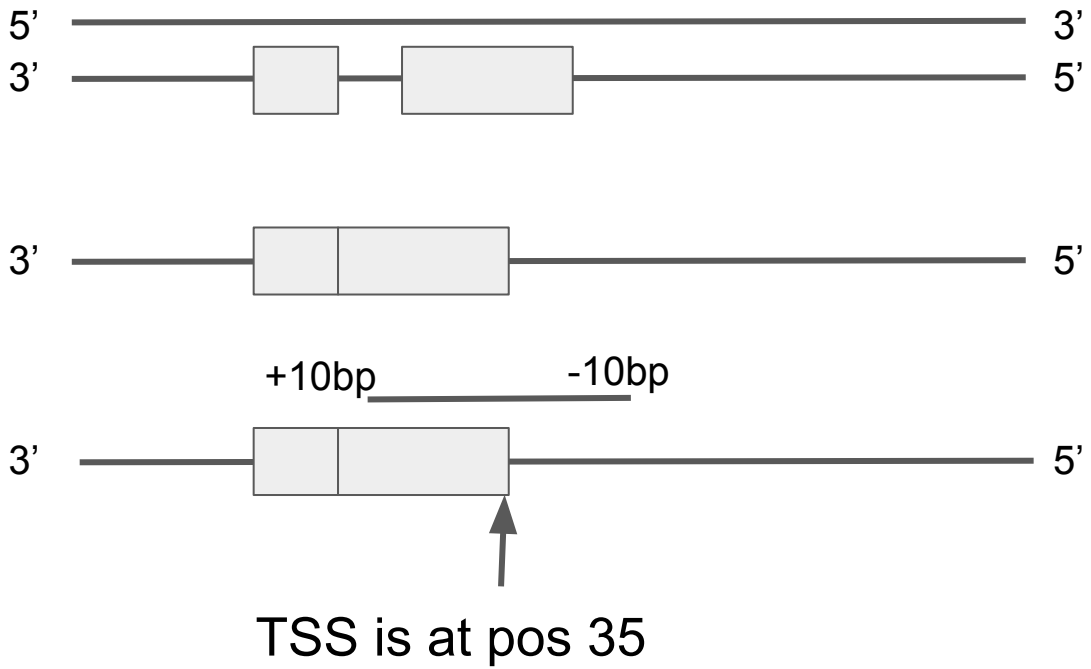


complement(join(...)) example

Example: complement(join(15..20,25..35))

15..20,25..35
- Coordinates on + strand
- But take sequence on reverse complement

join(15..20,25..35)



Other gotchas

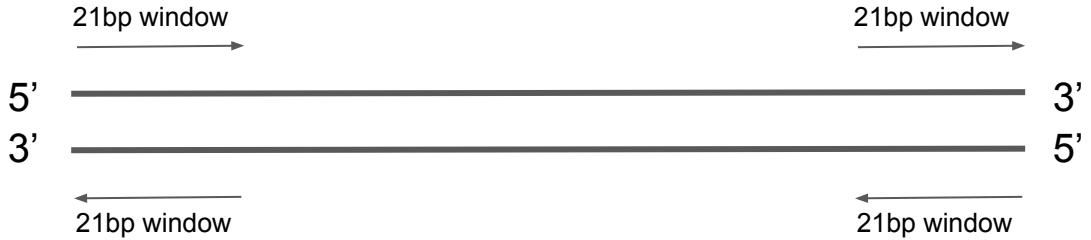
- What if the window is outside of the sequence (e.g. 1..100)?
- “>” and “<” characters
  - If a CDS contains these the position is uncertain and you can skip that CDS

Building the weight matrix

Steps:

1. Compute the background nucleotide frequencies
    a. Forward and reverse strands
2. Count matrix
    a. Compute the nucleotide counts around every TSS
3. Frequency Matrix
    a. Compute the proportion of times a nucleotide occurs at each position
4. Weight matrix
    a. Weight = log2([nt freq at motif position] / [background nt freq])
    b. If a nt has a frequency = 0, assign it a weight of -99.0

Computing site scores



- Use weight matrix to compute site scores at **all** positions in the genome
  - Score = sum of weights for nucleotide present at each position
  - Scores should be associated with motif **centered** on that position
  - Don’t extend window beyond the genome
  - Run on forward and reverse strands

Matching the template

Remember you can use diff or an online version (e.g., https://text-compare.com/)

You will be able to match it exactly because the output is deterministic

Information theory in Wordle

What is information?

What is "Information"?

Information, *n*. Data communicated or received that resolves uncertainty about a particular fact or circumstance.

Example: you receive some data about a card drawn at random from a 52-card deck. Which of the following data conveys the most information? The least?

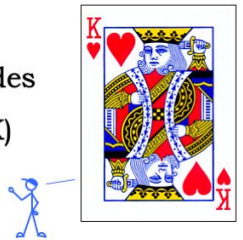
# of possibilities remaining

13 A. The card is a heart

51 B. The card is not the Ace of spades

12 C. The card is a face card (J, Q, K)

1 D. The card is the "suicide king"



https://computationstructures.org/lectures/info/slides/Slide02.png

How is information measured?

Shannon entropy, H

- Measures uncertainty

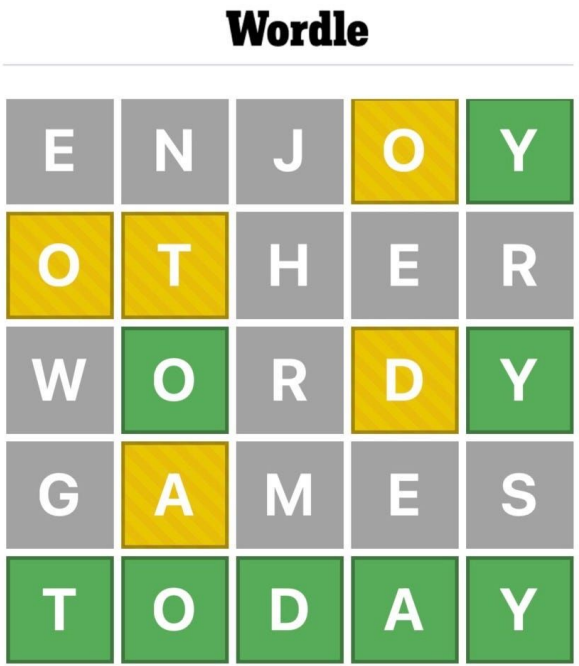
$$H = -\sum p(x) \log p(x)$$

Information is -log(p(x)), or roughly, how many times an observation, x, cuts the space of possibilities in half (because this log is base 2)

- Measured in bits

https://upload.wikimedia.org/wikipedia/commons/thumb/8/85/LexA_gram_positive_bacteria_sequence_logo.png/500px-LexA_gram_positive_bacteria_sequence_logo.png

What is Wordle?



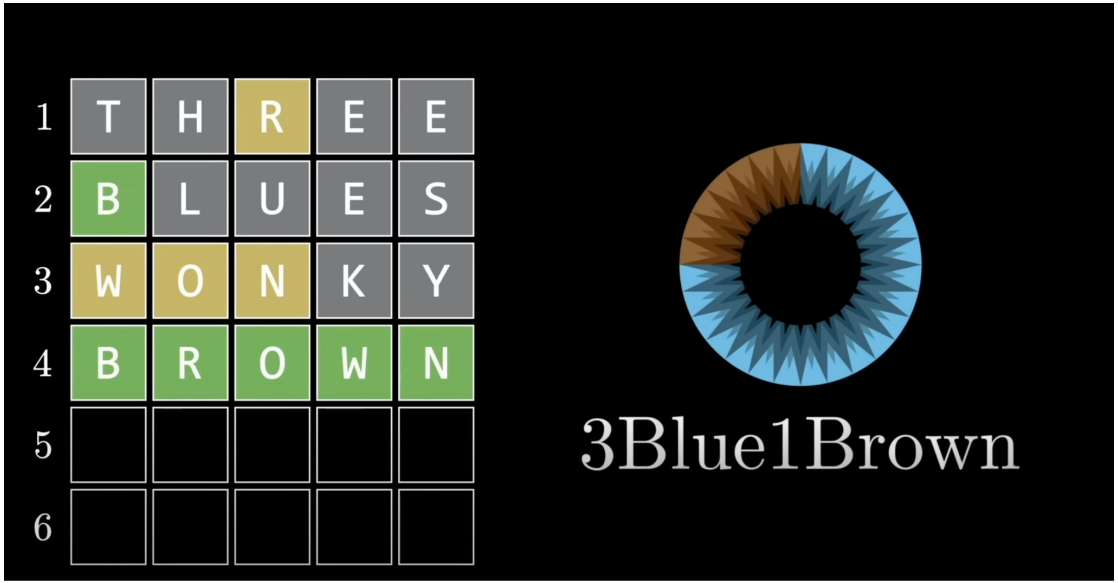
Word guessing game (similar to hangman)

Each word is 5 letters long

Each guess reveals information

Using information theory in Wordle

Adapted from 3blue1brown's video (linked below)



https://www.youtube.com/watch?v=v68zYyaEmEA

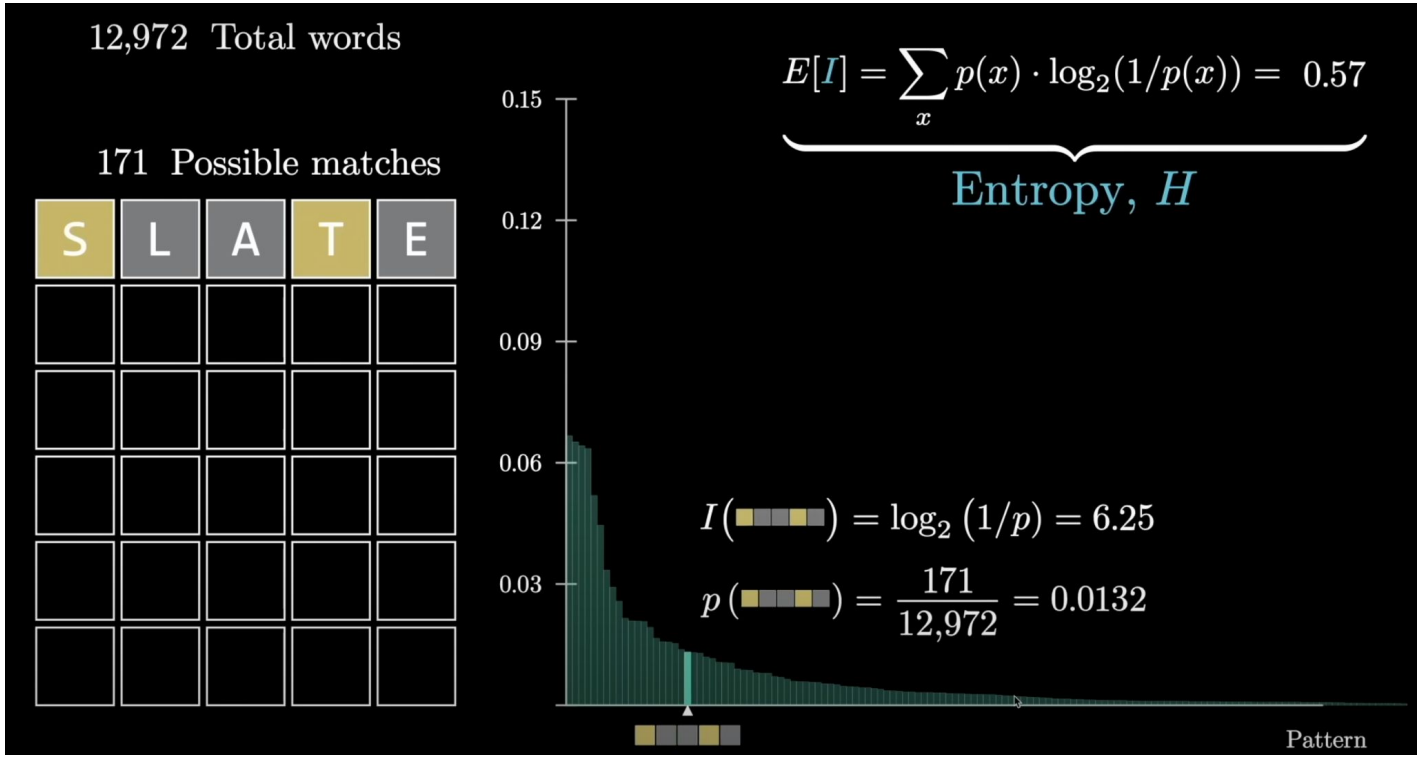
Defining the space of Wordle

12,972 possible guesses

2,315 potential answers

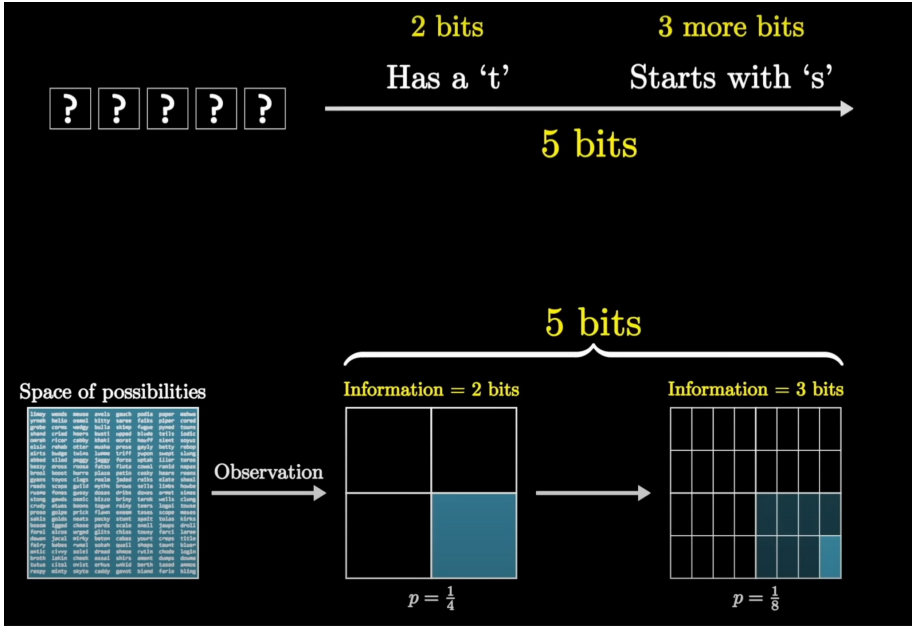
What first guess has the most information?

Best guess maximizes average information



Information is additive

In normal mode for Wordle, we can make unrelated guesses and combine information from those guesses



So… what’s the best first guess??

Office hours

Good work on homework 2!

Homework 3 is due this Sunday, February 2nd at 11:59pm!