Genome 540 Discussion

Conor Camplisson

January 24th, 2023

Genome Sciences
UNIVERSITY OF WASHINGTON

Outline

- Homework 2 wrap-up
- Homework 3 overview & questions

Outline

- Homework 2 wrap-up
- Homework 3 overview & questions

Summary

- Homework 1: find the longest match
  - Orthologous 10-Mb regions of Human and Mouse genome

- Homework 2: how significant is the result?
  - Run HW1 program on simulated mouse genomes
  - Collect stats on spontaneous match lengths
  - Use results to conclude match length significance

Homework 2 Wrap-up

```
print(seq)
for suffix in sorted(suffix_array, key=lambda suffix: suffix.seq):
    print(suffix.start_pos, suffix.seq)
```

```
cagtcatgAAAAAAAAAatgcgcaCCCCCCCgctacgctagAAAAAAAAAAAAAAAtctcatCCCCCCCCCcgtcaagca
42 AAAAAAAAAAAAAAAtctcatCCCCCCCCCcgtcaagca
43 AAAAAAAAAAAAAAtctcatCCCCCCCCCcgtcaagca
44 AAAAAAAAAAAAAtctcatCCCCCCCCCcgtcaagca
45 AAAAAAAAAAAAtctcatCCCCCCCCCcgtcaagca
46 AAAAAAAAAAAtctcatCCCCCCCCCcgtcaagca
47 AAAAAAAAAAtctcatCCCCCCCCCcgtcaagca
9 AAAAAAAAAatgcgcaCCCCCCCgctacgctagAAAAAAAAAAAAAAAtctcatCCCCCCCCCcgtcaagca
48 AAAAAAAAAtctcatCCCCCCCCCcgtcaagca
10 AAAAAAAAatgcgcaCCCCCCCgctacgctagAAAAAAAAAAAAAAAtctcatCCCCCCCCCcgtcaagca
49 AAAAAAAAtctcatCCCCCCCCCcgtcaagca
11 AAAAAAAatgcgcaCCCCCCCgctacgctagAAAAAAAAAAAAAAAtctcatCCCCCCCCCcgtcaagca
50 AAAAAAAtctcatCCCCCCCCCcgtcaagca
12 AAAAAAatgcgcaCCCCCCCgctacgctagAAAAAAAAAAAAAAAtctcatCCCCCCCCCcgtcaagca
51 AAAAAAtctcatCCCCCCCCCcgtcaagca
13 AAAAAatgcgcaCCCCCCCgctacgctagAAAAAAAAAAAAAAAtctcatCCCCCCCCCcgtcaagca
52 AAAAAtctcatCCCCCCCCCcgtcaagca
14 AAAAatgcgcaCCCCCCCgctacgctagAAAAAAAAAAAAAAAtctcatCCCCCCCCCcgtcaagca
53 AAAAtctcatCCCCCCCCCcgtcaagca
15 AAAatgcgcaCCCCCCCgctacgctagAAAAAAAAAAAAAAAtctcatCCCCCCCCCcgtcaagca
54 AAAtctcatCCCCCCCCCcgtcaagca
16 AAatgcgcaCCCCCCCgctacgctagAAAAAAAAAAAAAAAtctcatCCCCCCCCCcgtcaagca
55 AAtctcatCCCCCCCCCcgtcaagca
17 AatgcgcaCCCCCCCgctacgctagAAAAAAAAAAAAAAAtctcatCCCCCCCCCcgtcaagca
56 AtctcatCCCCCCCCCcgtcaagca
63 CCCCCCCCCcgtcaagca
64 CCCCCCCCcgtcaagca
65 CCCCCCCcgtcaagca
25 CCCCCCCgctacgctagAAAAAAAAAAAAAAAtctcatCCCCCCCCCcgtcaagca
66 CCCCCCcgtcaagca
26 CCCCCCgctacgctagAAAAAAAAAAAAAAAtctcatCCCCCCCCCcgtcaagca
67 CCCCCcgtcaagca
```

Suffix array algorithm

Note: when longest match N seq found, also find N-1 seq, N-2 seq, etc.

Single match result produces redundant histogram entries

Homework 2 Wrap-up

Hypothetical Match Results

4 significant match seqs:

{53-mer, 59-mer, 63-mer, 65-mer}

Match Length Histogram

match_len	count
50	4
51	4
52	4
53	4
54	3
55	3
56	3
57	3
58	3
59	3
60	2
61	2
62	2
63	2
64	1
65	1

HW2: Simulating genomes to assess match significance

Fasta 2: simulated_equal_freq.fa
Non-alphabetic characters: 0
>simulated_equal_freq
*=816373
A=204496
C=203794
G=204159
T=203924
N=0

Equal nucleotide frequency

Nucleotide Frequencies:
A=0.2505
C=0.2496
G=0.2501
T=0.2498

Dinucleotide Count Matrix:
A=51122 51212 51114 51048
C=50998 50966 50784 51045
G=51138 50840 51232 50949
T=51238 50775 51029 50882

Dinucleotide Frequency Matrix:
A=0.0626 0.0627 0.0626 0.0625
C=0.0625 0.0624 0.0622 0.0625
G=0.0626 0.0623 0.0628 0.0624
T=0.0628 0.0622 0.0625 0.0623

Conditional Frequency Matrix:
A=0.2500 0.2504 0.2500 0.2496
C=0.2502 0.2501 0.2492 0.2505
G=0.2505 0.2490 0.2509 0.2496
T=0.2513 0.2490 0.2502 0.2495

Fasta 3: simulated_markov_0.fa
Non-alphabetic characters: 0
>simulated_markov_0
*=816373
A=248748
C=162830
G=164026
T=240769
N=0

Zero-order Markov

Nucleotide Frequencies:
A=0.3047
C=0.1995
G=0.2009
T=0.2949

Dinucleotide Count Matrix:
A=75320 49910 50241 73276
C=49626 32300 32789 48115
G=50322 32668 32614 48422
T=73480 47952 48382 70955

Dinucleotide Frequency Matrix:
A=0.0923 0.0611 0.0615 0.0898
C=0.0608 0.0396 0.0402 0.0589
G=0.0616 0.0400 0.0399 0.0593
T=0.0900 0.0587 0.0593 0.0869

Conditional Frequency Matrix:
A=0.3028 0.2006 0.2020 0.2946
C=0.3048 0.1984 0.2014 0.2955
G=0.3068 0.1992 0.1988 0.2952
T=0.3052 0.1992 0.2009 0.2947

Fasta 4: simulated_markov_1.fa
Non-alphabetic characters: 0
>simulated_markov_1
*=816373
A=249747
C=162576
G=163642
T=240408
N=0

First-order Markov

Nucleotide Frequencies:
A=0.3059
C=0.1991
G=0.2005
T=0.2945

Dinucleotide Count Matrix:
A=99263 50746 47775 51963
C=52971 36546 26807 46252
G=40731 37256 36898 48757
T=56782 38028 52161 93436

Dinucleotide Frequency Matrix:
A=0.1216 0.0622 0.0585 0.0637
C=0.0649 0.0448 0.0328 0.0567
G=0.0499 0.0456 0.0452 0.0597
T=0.0696 0.0466 0.0639 0.1145

Conditional Frequency Matrix:
A=0.3975 0.2032 0.1913 0.2081
C=0.3258 0.2248 0.1649 0.2845
G=0.2489 0.2277 0.2255 0.2979
T=0.2362 0.1582 0.2170 0.3887

Outline

- Homework 2 wrap-up
- Homework 3 overview & questions

Homework 3 Overview

- Part one: parse genbank (.gbff) file with *S. pyogenes* genome
  - Extract all CDS features, coordinates (note: join(), complement())
  - Extract genomic sequence, compute reverse complement

- Part two: build a site model for translation start sites (TSS)
  - Compute nucleotide frequencies at annotated TSS positions
  - Compute nucleotide frequencies throughout genome (both strands!)
  - Compute weights using $\log_2$ ratio of the appropriate frequencies

- Part three: compute scores using site model
  - Scores for every annotated TSS (21-nt window centered on)
  - Scores for every 21-nt window in the genome (both strands!)

Genbank File Format - 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)
```

Genbank File Format - Features

```
  TITLE     Direct Submission
  JOURNAL   Submitted (30-JUL-2014) Laboratory of Genetics, University of
            Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
  REMARK    Protein update by submitter
COMMENT     On Sep 26, 2013 this sequence version replaced U00096.2.
            Current U00096 annotation updates are derived from EcoGene
            http://ecogene.org. Suggestions for updates can be sent to Dr.
            Kenneth Rudd (krudd@miami.edu). These updates are being generated
            from a collaboration that also includes ASAP/ERIC, the Coli Genetic
            Stock Center, EcoliHub, EcoCyc, RegulonDB and UniProtKB/Swiss-Prot.
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"
     gene            337..2799
```

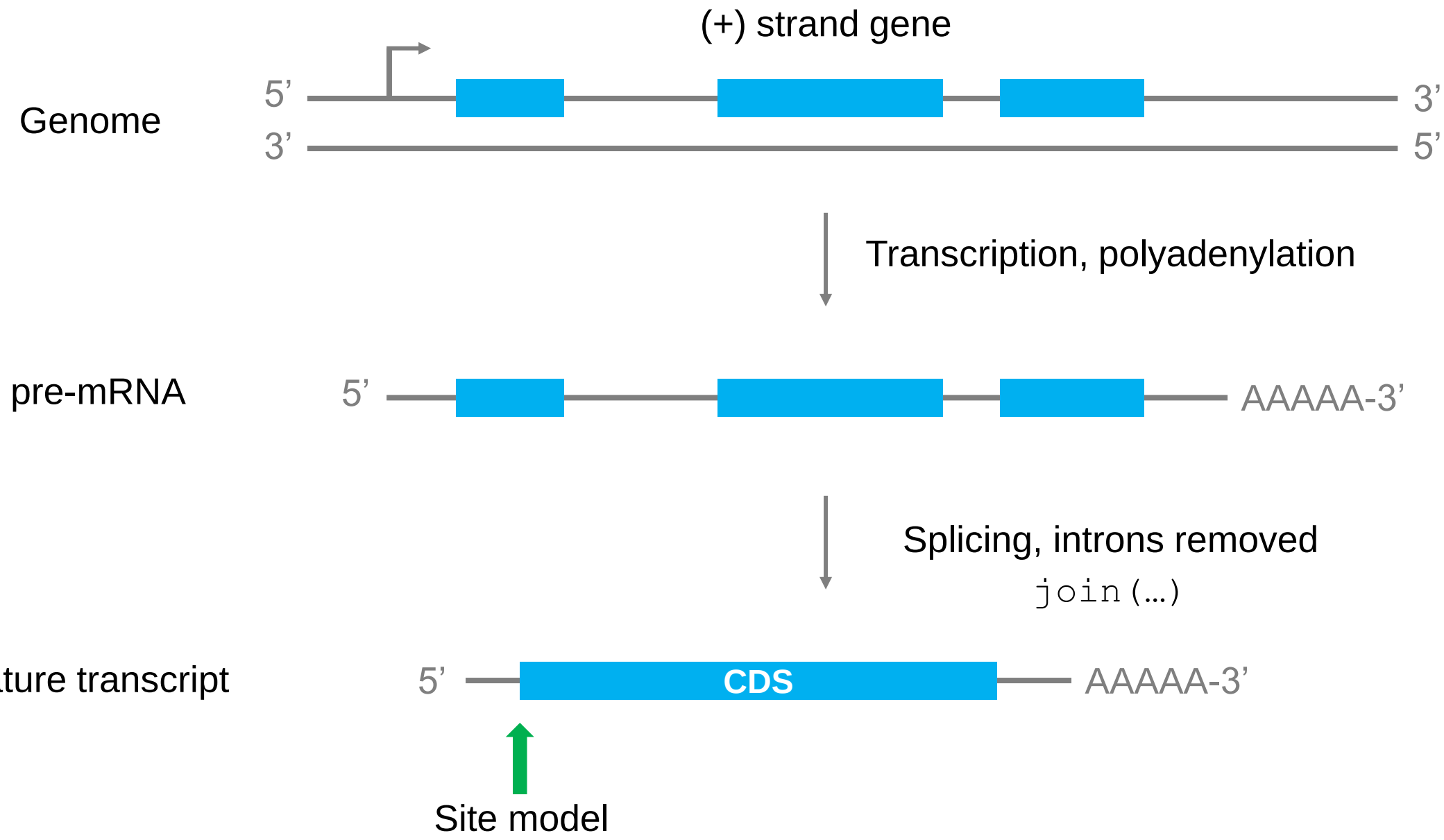
Other coord. string examples:

```
17489..18655
18715..19620
complement(19811..20314)
complement(20233..20508)
complement(20815..21078)
21181..21399
21407..22348
join(1465392..1467904,1469241..1469293,1470517..1474013)
complement(join(1489713..1489964,1489964..1490713))
join(1530586..1531323,1531325..1531639)
complement(join(1544384..1544764,1544764..1545714))
complement(join(1590334..1590426,1590426..1590536))
complement(join(1592665..1594125,1594127..1597987))
```

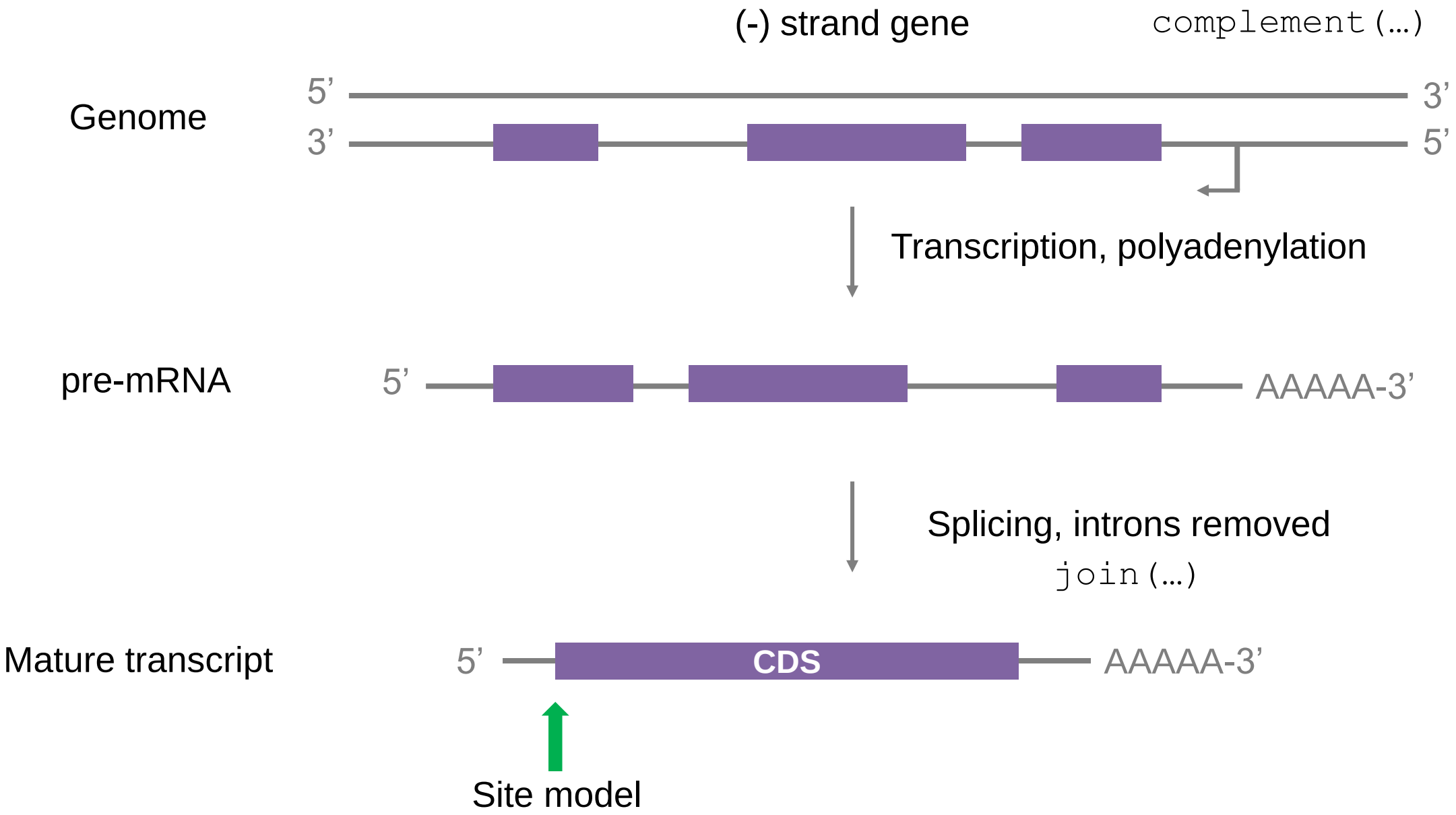
Genbank File Format - Sequence

```
                     /protein_id="AAC77356.1"
                     /db_xref="ASAP:ABE-0014442"
                     /db_xref="UniProtKB/Swiss-Prot:P37005"
                     /db_xref="EcoGene:EG12309"
                     /translation="MRITIILVAPARAENIGAAARAMKTMGFSDLRIVDSQAHLEPAT
                     RWVAHGSGDIIDNIKVFPTLAESLHDVDFTVATTARSRAKYHYYATPVELVPLLEEKS
                     SWMSHAALVFGREDSGLTNEELALADVLTGVPMVADYPSLNLGQAVMVYCYQLATLIQ
                     QPAKSDATADQHQLQALRERAMTLLTTLAVADDIKLVDWLQQRLGLLEQRDTAMLHRL
                     LHDIEKNITK"
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
```

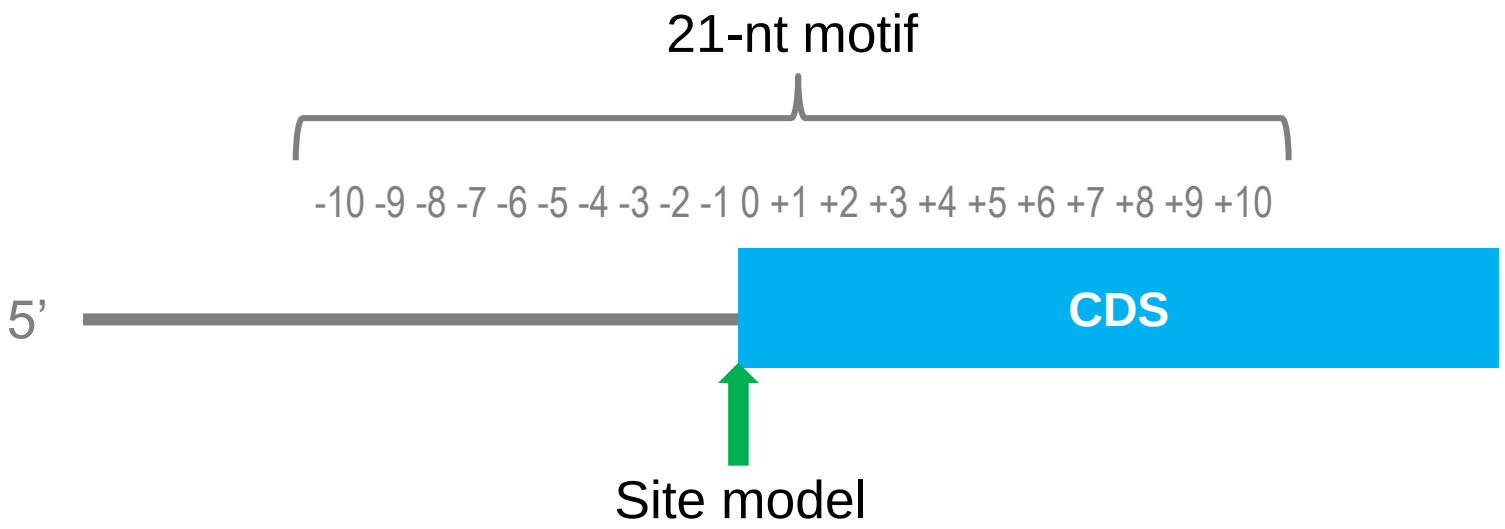
Homework 3 – CDS Structure



Homework 3 – CDS Structure



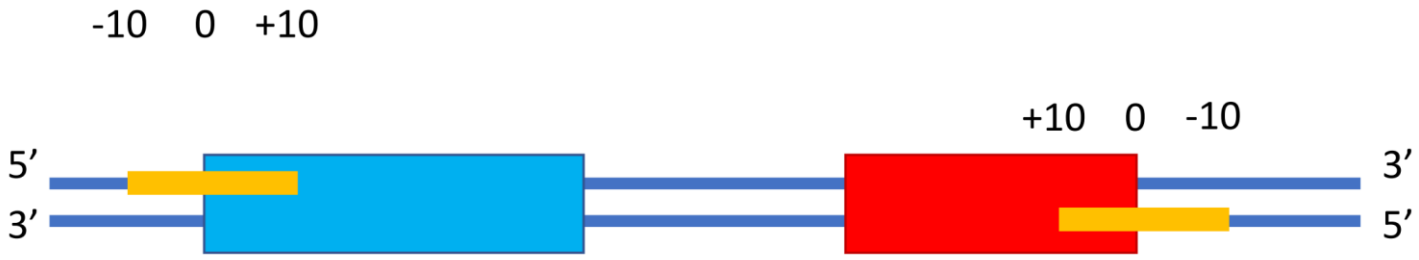
Homework 3 – Site Model




```
Weight Matrix:
-10 0.4344 -1.8751 0.9260 -1.0877
-9 0.1992 -2.1333 1.1363 -1.2787
-8 0.5695 -1.6520 0.7448 -0.9371
-7 0.5922 -1.1832 0.3613 -0.3652
-6 0.4622 -0.7999 0.0858 -0.0023
-5 0.4464 -0.5142 -0.2966 0.1873
-4 0.5277 -0.2149 -0.5142 0.0153
-3 0.7707 -0.4807 -0.1522 -0.5358
-2 0.0809 0.0733 -0.8775 0.4424
-1 0.1376 0.0532 -0.7133 0.3378
0 1.8624 -6.2918 -1.6562 -3.5143
1 -5.4685 -6.3987 -6.3987 2.0062
2 -6.2461 -5.5142 1.9572 -5.5941
3 1.0011 -0.6813 -0.3326 -0.7383
4 0.5795 0.1594 -0.6173 -0.4260
5 0.4182 -0.3952 -0.6458 0.3601
6 0.7643 -0.2521 -0.1728 -0.7774
7 0.7810 -0.3935 -0.9545 0.0393
8 0.3291 -0.4004 -0.2789 0.2331
9 0.5211 -0.0991 -0.3526 -0.2222
10 0.3150 -0.2303 -1.0011 0.5107
```

(A, C, G, T)

Homework 3 – Site Model



Step 0: Compute background nucleotide frequencies (genome + reverse complement).

Step 1: Count matrix – record the number of times each nucleotide shows up at each motif position (-10 to +10).

Step 2: Frequency matrix – proportion of times each nucleotide shows up at each motif position (-10 to +10).

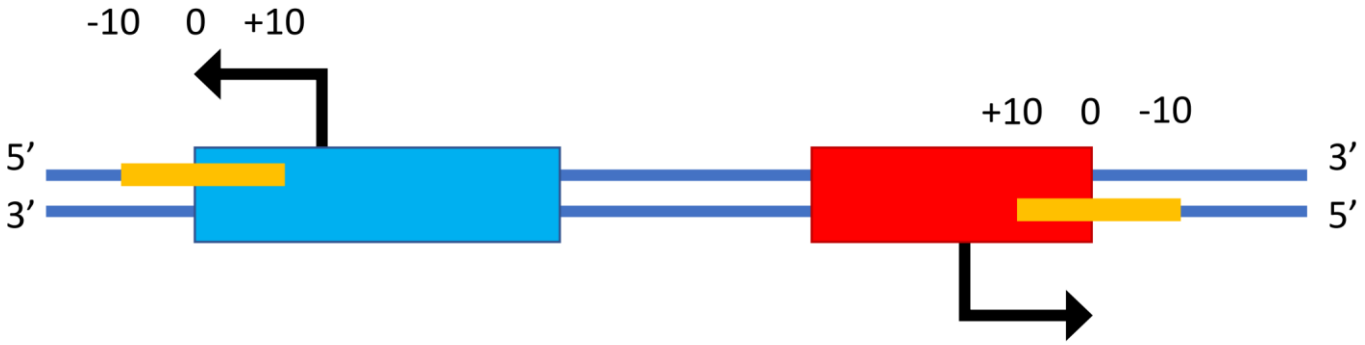
Step 3: Weight matrix

- $$\text{weight} = \log_2\left(\frac{\text{nt frequency at motif position}}{\text{nt background frequency}}\right)$$
- If a nt has frequency zero, assign a weight of -99.0 ($2^{-99} = 1.6\text{x}10^{-30} \approx 0$)

```
Weight Matrix:
-10 0.4344 -1.8751 0.9260 -1.0877
-9 0.1992 -2.1333 1.1363 -1.2787
-8 0.5695 -1.6520 0.7448 -0.9371
-7 0.5922 -1.1832 0.3613 -0.3652
-6 0.4622 -0.7999 0.0858 -0.0023
-5 0.4464 -0.5142 -0.2966 0.1873
-4 0.5277 -0.2149 -0.5142 0.0153
-3 0.7707 -0.4807 -0.1522 -0.5358
-2 0.0809 0.0733 -0.8775 0.4424
-1 0.1376 0.0532 -0.7133 0.3378
0 1.8624 -6.2918 -1.6562 -3.5143
1 -5.4685 -6.3987 -6.3987 2.0062
2 -6.2461 -5.5142 1.9572 -5.5941
3 1.0011 -0.6813 -0.3326 -0.7383
4 0.5795 0.1594 -0.6173 -0.4260
5 0.4182 -0.3952 -0.6458 0.3601
6 0.7643 -0.2521 -0.1728 -0.7774
7 0.7810 -0.3935 -0.9545 0.0393
8 0.3291 -0.4004 -0.2789 0.2331
9 0.5211 -0.0991 -0.3526 -0.2222
10 0.3150 -0.2303 -1.0011 0.5107
```

(A, C, G, T)

Homework 3 – Site Model



- Score for a position = sum of the weights for each nucleotide in the 21bp motif *centered at* that position
- Scores for a position are strand-specific (different for forward vs. reverse)
- Compute scores for *all* possible positions (both strands)

```
Weight Matrix:
-10 0.4344 -1.8751 0.9260 -1.0877
-9 0.1992 -2.1333 1.1363 -1.2787
-8 0.5695 -1.6520 0.7448 -0.9371
-7 0.5922 -1.1832 0.3613 -0.3652
-6 0.4622 -0.7999 0.0858 -0.0023
-5 0.4464 -0.5142 -0.2966 0.1873
-4 0.5277 -0.2149 -0.5142 0.0153
-3 0.7707 -0.4807 -0.1522 -0.5358
-2 0.0809 0.0733 -0.8775 0.4424
-1 0.1376 0.0532 -0.7133 0.3378
0 1.8624 -6.2918 -1.6562 -3.5143
1 -5.4685 -6.3987 -6.3987 2.0062
2 -6.2461 -5.5142 1.9572 -5.5941
3 1.0011 -0.6813 -0.3326 -0.7383
4 0.5795 0.1594 -0.6173 -0.4260
5 0.4182 -0.3952 -0.6458 0.3601
6 0.7643 -0.2521 -0.1728 -0.7774
7 0.7810 -0.3935 -0.9545 0.0393
8 0.3291 -0.4004 -0.2789 0.2331
9 0.5211 -0.0991 -0.3526 -0.2222
10 0.3150 -0.2303 -1.0011 0.5107
```

(A, C, G, T)

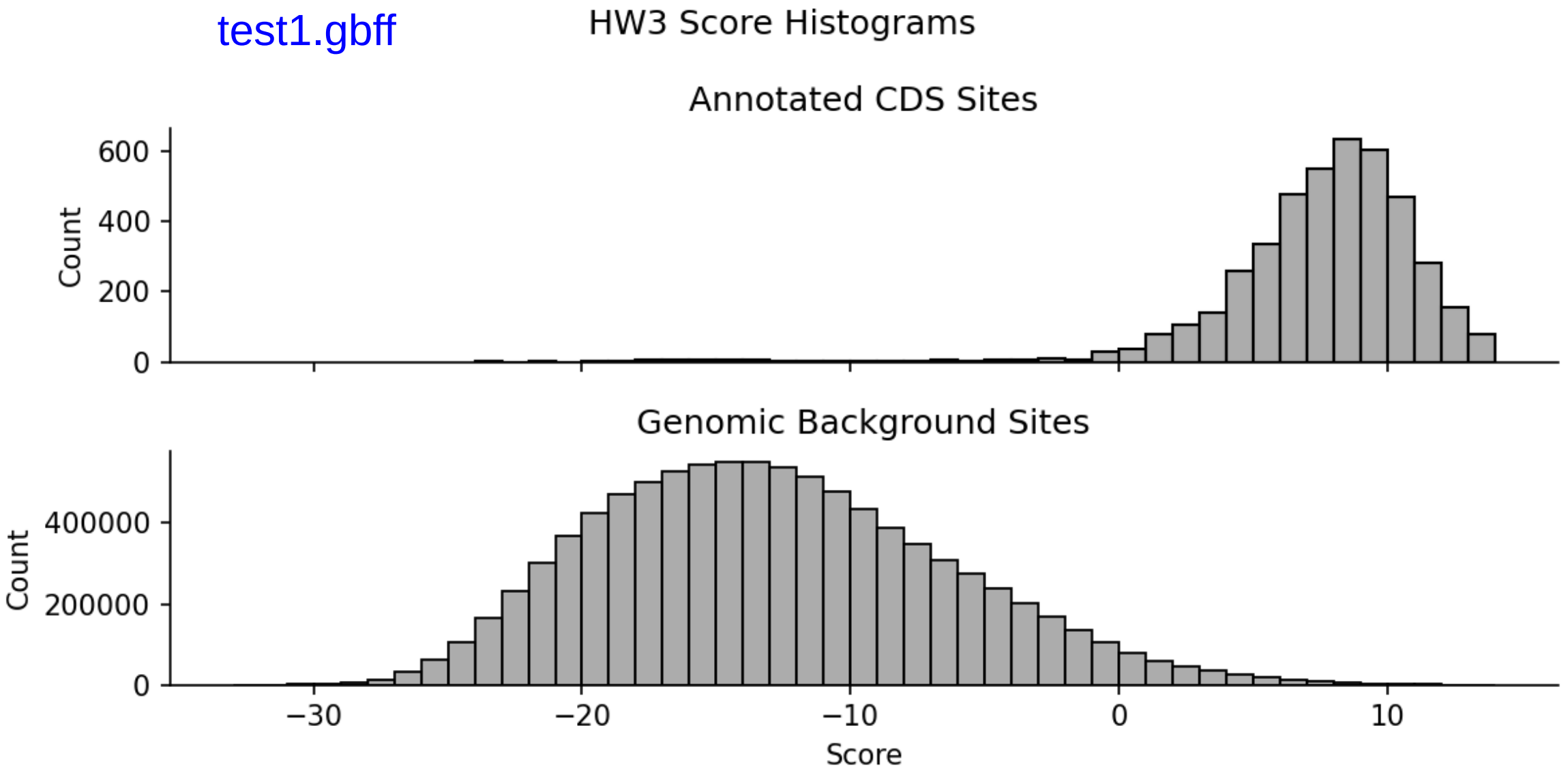
Homework 3 – Score Histograms

- Two histograms:
  - All genomic positions
  - Positions that are annotated CDS TSSs
- Group scores into bins of size 1 (round down to nearest integer)
- Format – two columns:
  - Score bin
  - Number of sites with that score
- Print all bins with at least one count
- Put all scores less than -50 into one bin

```
Score Histogram All:
-5 101880
-4 76413
-3 54704
-2 38081
-1 27202
0 21440
1 18671
2 18825
3 19072
4 18675
5 17308
6 14429
7 10595
8 6915
9 3886
10 1850
11 699
12 225
13 46
14 4
lt-50 6132782
```

Homework 3 – Test Case Results

test1.gbff



Homework 3 – Position List

- List of *non-CDS* positions with a motif score >= 10
- Format – three columns:
  - 1-indexed genome position (on forward strand)
  - Strand indicator (0 for forward, 1 for reverse)
  - Score (to four decimal places)

```
Position List:
1899 0 10.1167
2274 0 10.1923
2502 0 10.1098
4646 0 10.5886
5252 0 10.5534
6127 0 11.0669
7250 1 10.0453
11016 1 10.1616
...
```

Homework 3 – Output Summary

- Nucleotide histogram
- Background nt frequencies (based on both strands)
- Count matrix (-10 to +10 nucleotides)
- Frequency matrix (-10 to +10 nucleotides)
- Weight matrix (-10 to +10 nucleotides)
- Maximum score (Max score possible given scoring matrix, not max observed score)
- Score histogram for annotated CDS TSSs
- Score histogram for all positions
- List of non-CDS positions with score >=10

Homework 3 – Notes

```
Assignment: GS 540 HW3
Name: Conor Camplisson
Email: concamp@uw.edu
Language: Python
Runtime: 4.285 seconds

Nucleotide Histogram:
A=1142742
C=1180091
G=1177437
T=1141382
N=0

Background Frequency:
A=0.2460
C=0.2540
G=0.2540
T=0.2460
```

(slight difference from HW1, HW2)

Use 4 decimal places in outputs
(except 10 for max score)

Example sequences & solutions:

Test sequence 1

Answer for test sequence 1

Test sequence 2

Answer for test sequence 2

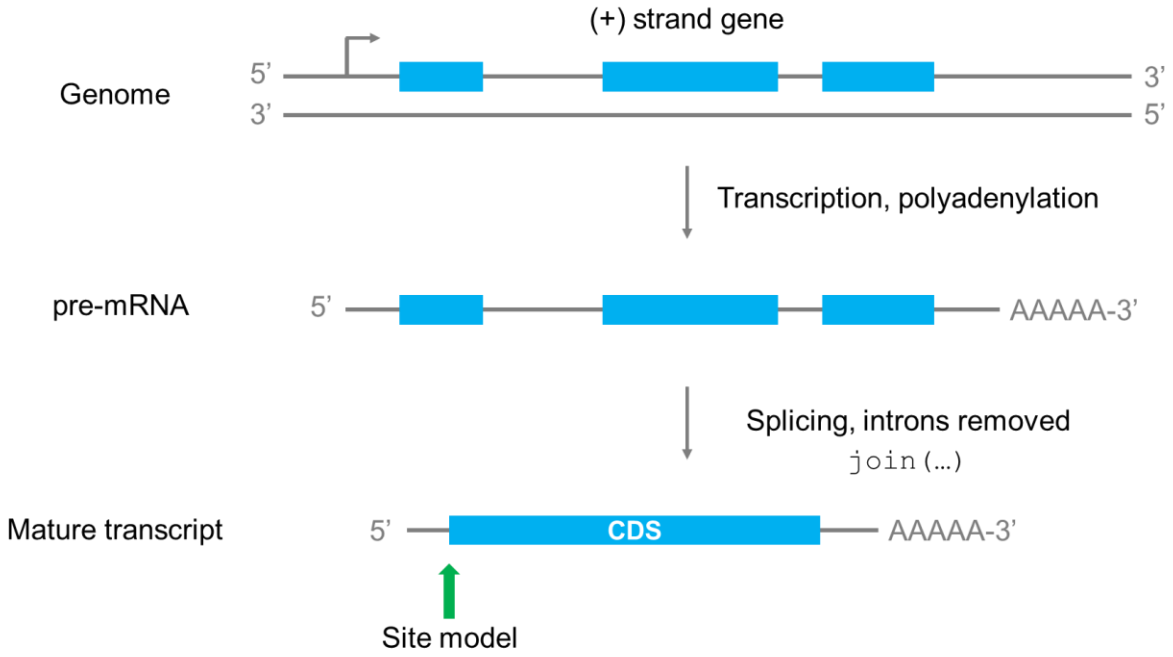
Test 1
does not contain N's
will not produce "lt-50" scores
does not require -99.0 handling

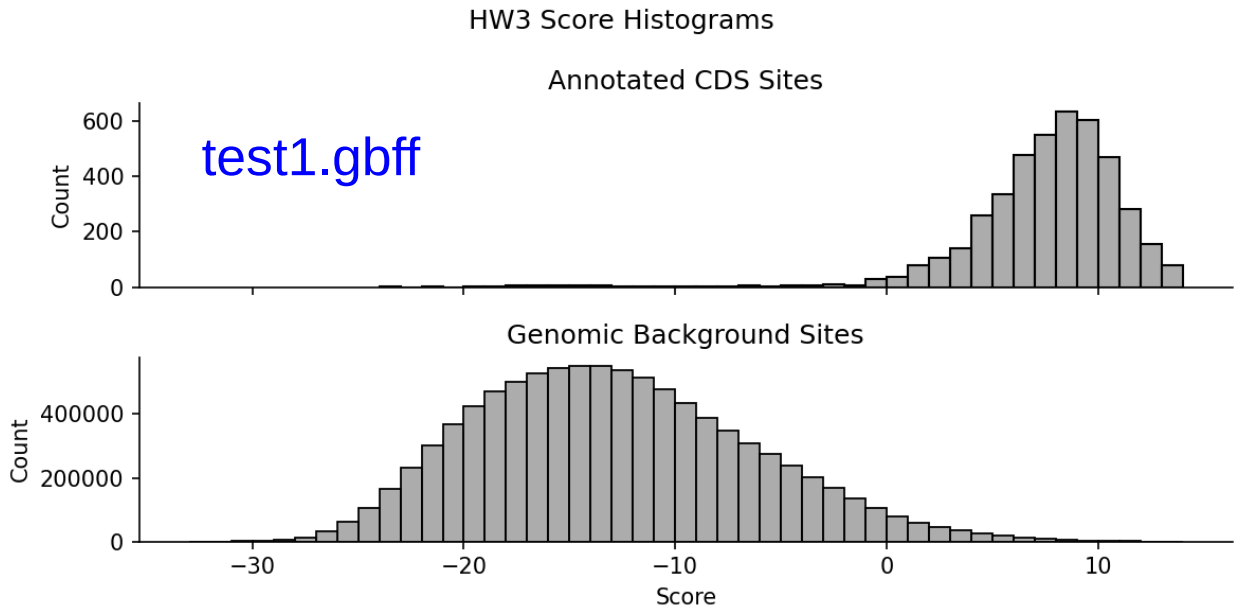
Test 2
does contain N's
will produce "lt-50" scores
does require -99.0 handling

Homework 3 – Tips

- Looking only for ‘CDS’ features
  - Only consider positions where location is certain (no < or >)
- Positions downstream of the translation start site could be noncontiguous
  - join(1000…1008, 1200…1500)
- Also watch out for multi-line joins
- Precision matters! (**use doubles over floats**)
- Make sure outputs make sense (frequencies sum to 1, etc.)

Homework 3 Questions ?

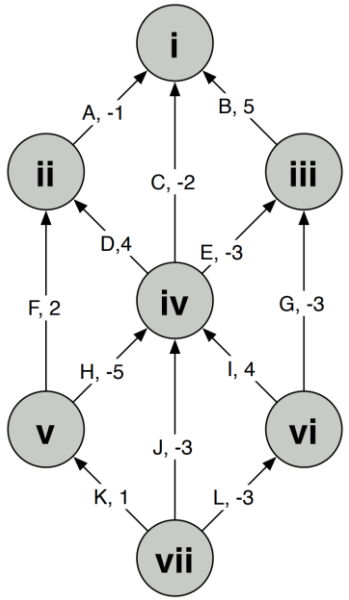




Reminders

- Homework 3 due this Sunday Jan. 29, 11:59 pm
  - Single text file, compressed with `gzip`
  - name in the file: `camplisson_hw3.txt.gz`

- Homework 4 will be posted tomorrow



. . .