Genome 540 Discussion

Conor Camplisson

January 19th, 2023

Genome Sciences
UNIVERSITY OF WASHINGTON

Outline

- Homework 3 overview
- Homework 2 & 3 questions

Outline

- Homework 3 overview
- Homework 2 & 3 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
```

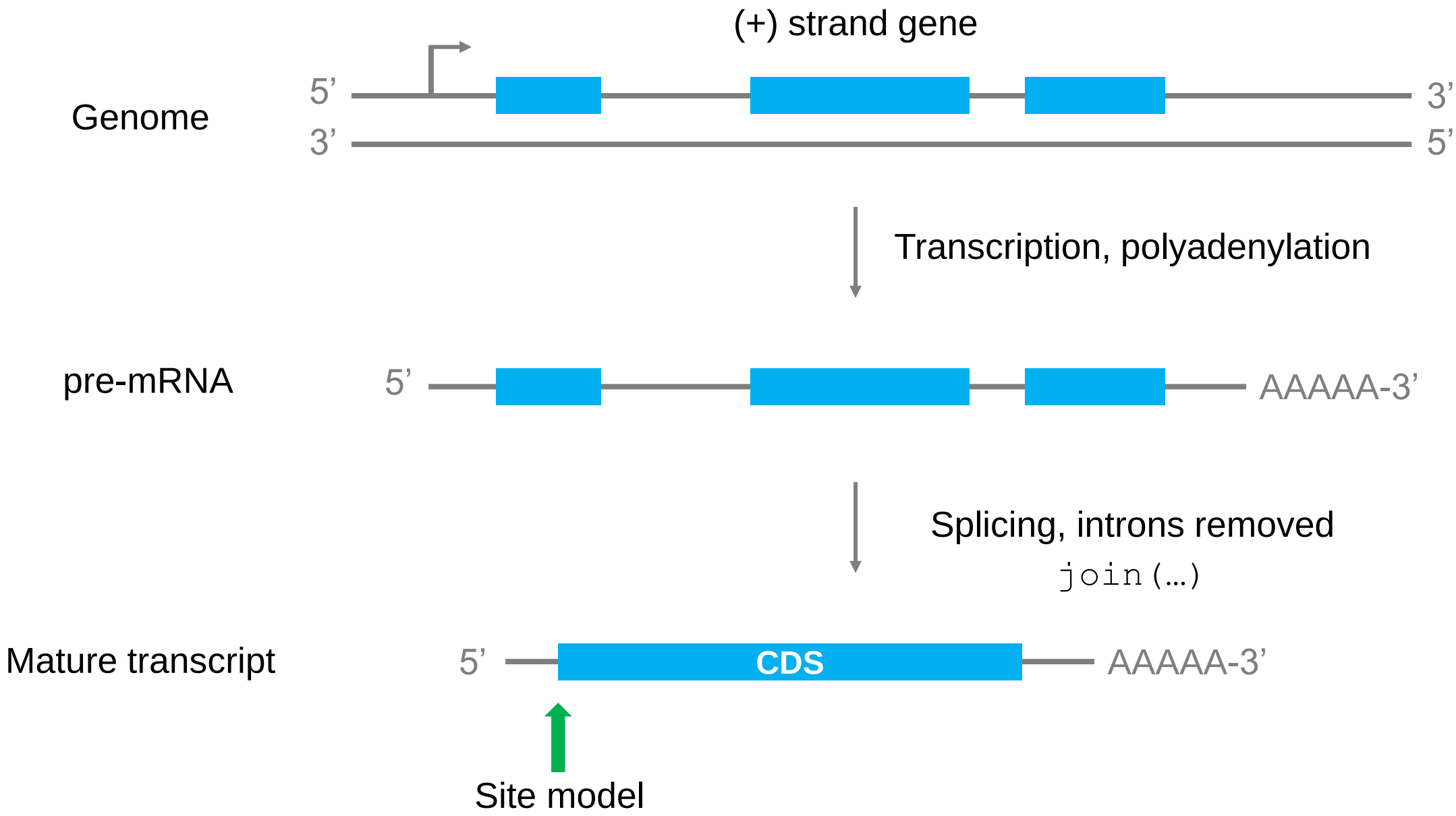
Other coord. string examples:

```
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))
```

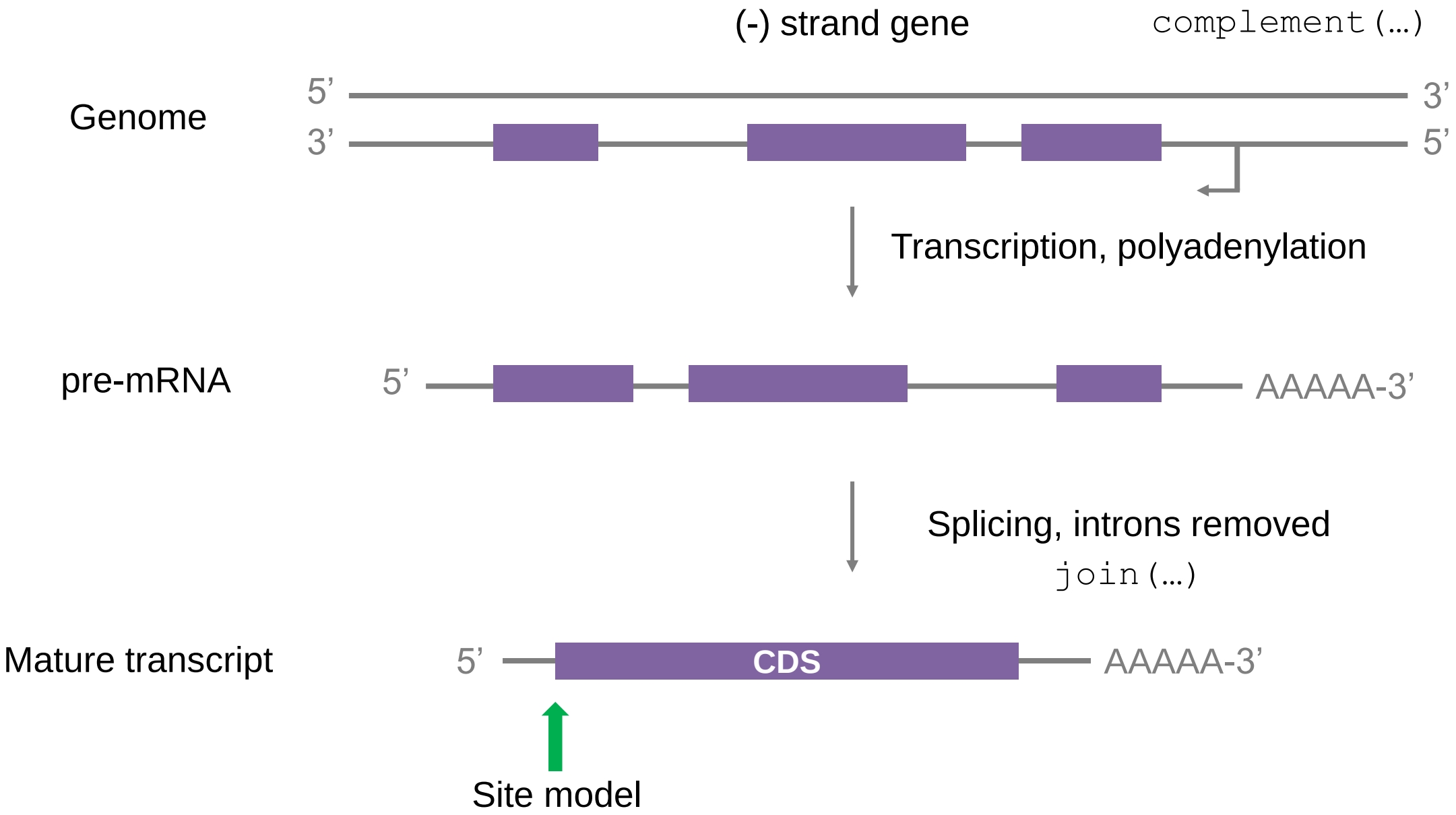
Genbank File Format - Sequence

```
                     /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
```

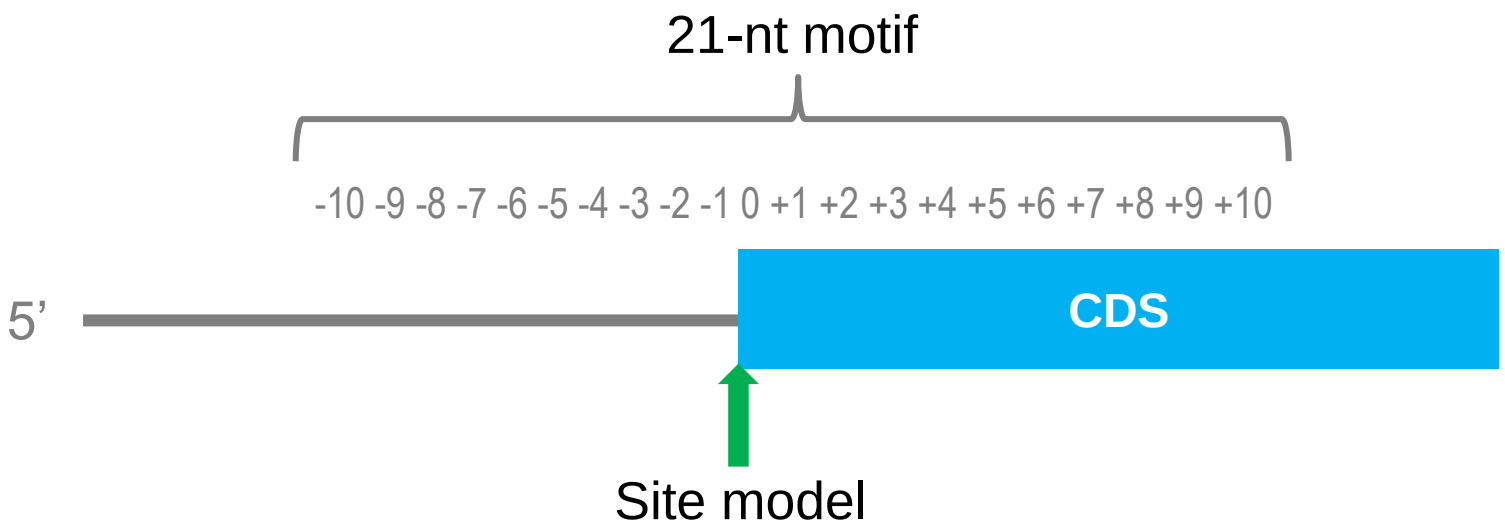
Homework 3 Overview



Homework 3 Overview



Homework 3 Overview




```
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 Overview

```
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

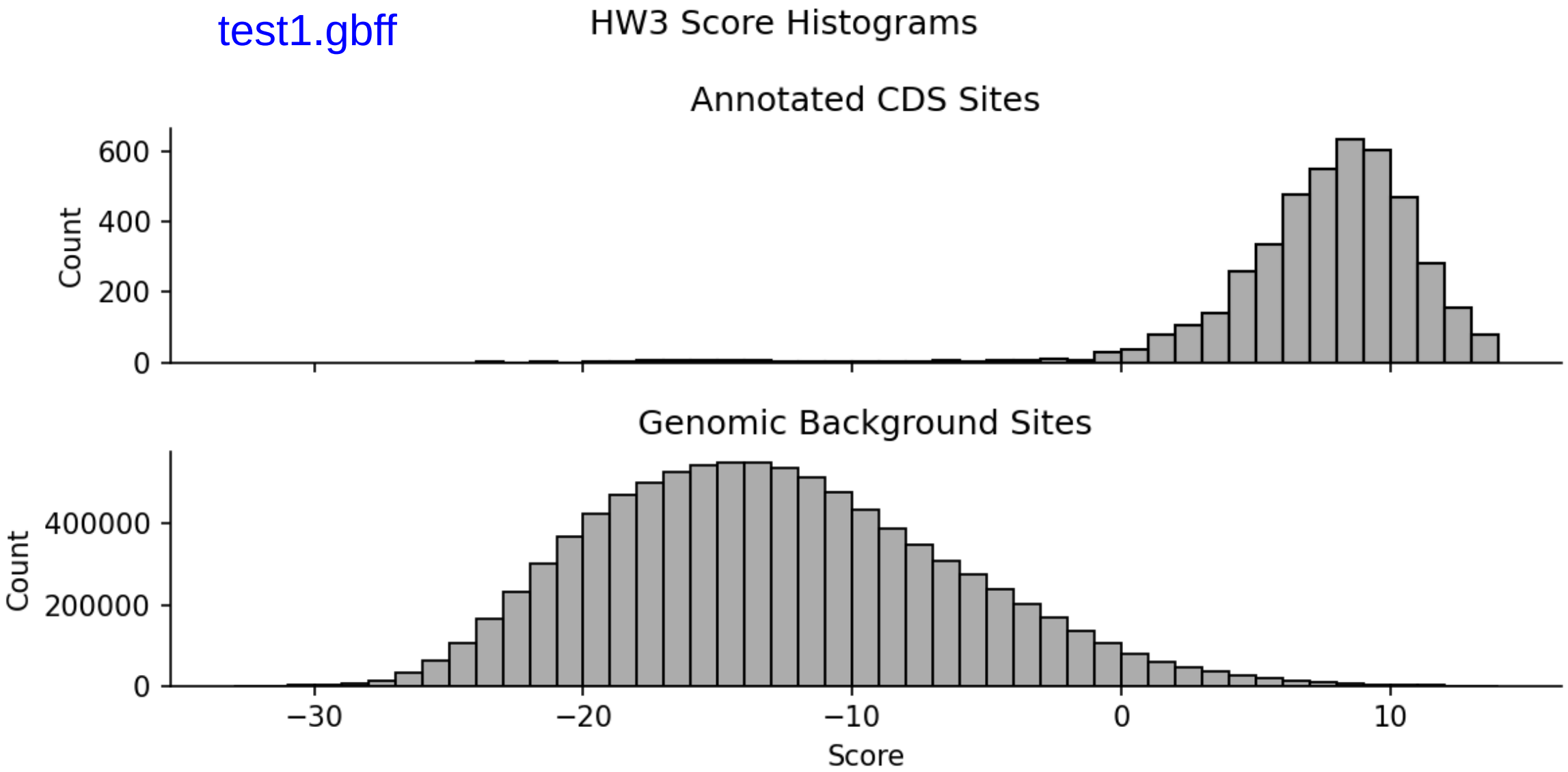
Answer for test sequence 2

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 Overview

test1.gbff



Outline

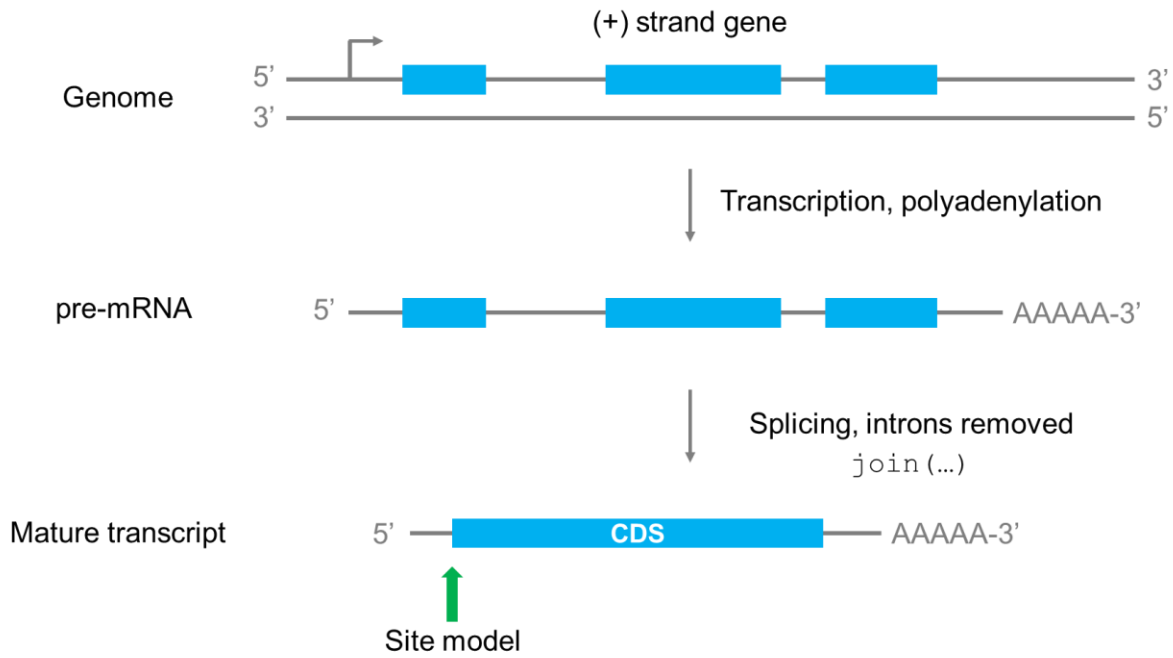
- Homework 3 overview

- Homework 2 & 3 questions

Homework 2 & 3 Questions ?

Observed Dinuc Freqs

	A	C	G	T
A	0.099	0.051	0.069	0.078
C	0.073	0.052	0.011	0.069
G	0.059	0.043	0.052	0.050
T	0.066	0.059	0.072	0.098



Reminders

- Homework 2 due this Sunday Jan. 22, 11:59 pm
  - Single text file, compressed with `gzip`
  - name in the file: `camplisson_hw2.txt.gz`

- Homework 3 due next Sunday Jan. 29, 11:59 pm

. . .