

Genome 540 Discussion

Jan. 27th 2026



HW3 Intro

Weight Matrix – 3' Splice Sites (*C. elegans*)

SITE FREQUENCIES:

A	0.400	0.429	0.282	0.058	0.008	0.092	0.029	1.000	0.000	0.410	0.293	0.307
C	0.118	0.079	0.081	0.029	0.016	0.135	0.834	0.000	0.000	0.156	0.187	0.225
G	0.072	0.070	0.063	0.018	0.005	0.073	0.001	0.000	1.000	0.310	0.159	0.191
T	0.409	0.422	0.574	0.896	0.971	0.700	0.135	0.000	0.000	0.124	0.361	0.276

BACKGROUND FREQUENCIES:

A	0.321	0.321	0.321	0.321	0.321	0.321	0.321	0.321	0.321	0.321	0.321	0.321
C	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179
G	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179	0.179
T	0.321	0.321	0.321	0.321	0.321	0.321	0.321	0.321	0.321	0.321	0.321	0.321

WEIGHTS:

A	0.32	0.42	-0.18	-2.46	-5.29	-1.79	-3.45	1.64	-99.00	0.36	-0.13	-0.06
C	-0.60	-1.18	-1.15	-2.64	-3.51	-0.41	2.22	-99.00	-99.00	-0.20	0.06	0.33
G	-1.31	-1.35	-1.51	-3.35	-5.23	-1.30	-6.93	-99.00	2.48	0.79	-0.17	0.10
T	0.35	0.39	0.84	1.48	1.60	1.12	-1.24	-99.00	-99.00	-1.37	0.17	-0.22

General Idea:

Using known translation start sites, generate a weight matrix.

Use that matrix to score every known TSS, and then *every nucleotide* in the genome!

(Is there evidence of unannotated translation start sites?)

How do we get a hold of known TSSs?

From a GBFF (Genbank Flat File)

<https://www.insdc.org/submitting-standards/feature-table/>

(International Nucleotide Sequence Database Collaboration)

GBFF 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)
  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="G0 process: G0:0009088 - threonine biosynthetic
                        process"
                        /codon_start=1
                        /transl_table=11
                        /product="thr operon leader peptide"
                        /protein_id="AAC73112.1"
                        /db_xref="ASAP:ABE-00000006"
                        /db_xref="UniProtKB/Swiss-Prot:P0AD86"
                        /db_xref="EcoGene:Eg11277"
                        /translation="MKRISTITITTTITITGNGAG"
```

Sequence

```
ORIGIN
1   agcttttcat tctgactgca acgggcaata tgtctctgtg tggattaaaa aaagagtgtc
61  tgatagcagc ttctgaactg gttacctgcc gtgagtaaat taaaatttta ttgacttagg
121 tcactaaata ctttaaccaa tataggcata gcgcacagac agataaaaaa tacagagtac
181 acaacatcca tgaaacgcat tagcaccacc attaccacca ccataccatc taccacaggt
241 aacgggtcgg gctgacgcgt acaggaaaca cagaaaaaag cccgcacctg acagtgcggg
301 cttttttttt cgaccaaagg taacgaggta acaacatgac gagtgttgaa gticggcggt
361 acatcagttg caaatgcaga acgttttctg cgtgtgtccg atattctgga aagcaatgcc
421 aggcaggggc aggtggccac cgtcctctct gccccgccca aaatcaccaa ccacctggtg
481 gcgatgattg aaaaaacatc tagcggccag gatgctttac ccaatatcag cgatgccgaa
541 cgtatttttg ccgaactttt gacgggacct gccgccgccc agccggggtt cccgctggcg
601 caattgaaaa ctttcgtcga tcagggaattt gcccaataaa aacatgtcct gcattgcat
661 agtttgttgg ggcagtgccc ggatagcate aacgctgcgc tgatttgccg tggcgagaaa
721 atgctgatcg ccattatggc cggcgatta gaagcgcgcg gtcacaacgt tactgttatc
781 gatccggtcg aaaaactgct ggcagtgggg cattacctcg aatctaccgt cgattattgt
841 gagtccaccc gccgtattgc ggcaagccgc attccggctg atcacatggt gctgatggca
901 ggtttcacgc ccggtaatga aaaagcgcaa cttgtgtgtc ttggacgcaa cggttccgac
961 tactctgctg cgtgtctggc tgcctgttta cgcgccgatt gtgctgagat ttggacggac
1021 gttgacgggg tctataacct cgaccgcgct cagggtgccc atgcgaggtt gttgaagtcg
1081 atgtcttacc aggaagcgat ggagctttcc tactctggcg ctaaaattct tcacccccgc
1141 accattaccc ccacgcacca gttccagatc ccttgctgta ttaaaatac cggaatcct
1201 caagcaccag gtacgctcat tggtgccagc cgtgatgaag acgaattacc ggtaagggac
1261 atttccaatc tgaataacat ggcaatgttc agcgtttctg gtccgggagt gaaagggatg
1321 gtccgcatatg cggcgccgct ctttgacgac atgtcacgac cccgtatttc cgtgtgtctg
1381 attacgcaat catcttcgca atacagcatc agttttctgc ttccacaaaag cgactgtgtg
1441 cgagctgaac gggcaatgca ggaagagttc tactctggaac tgaagaagg cttactggag
1501 ccgctggcag tgacggaacg gctggccatt atctcggtg taggtgatgg tatgcgcacc
1561 ttgcgtggga tctcggcgaa attctttgcc gcactggccc gcgcaatat caacattgtc
```

Cliff Rostomily & Joe Min

GBFF file format

```
FEATURES
  source
    Location/Qualifiers
      1..1746380
      /organism="Streptococcus pyogenes"
      /mol_type="genomic DNA"
      /strain="NCTC12064"
      /serovar="Lancefield Group A"
      /isolation_source="not available: to be reported later"
      /culture_collection="NCTC:12064"
      /db_xref="taxon:1314"
      /chromosome="1"
      /country="United Kingdom: Telford"
      /collection_date="1900/1982"
  gene
    1..1356
    /gene="dnaA"
    /locus_tag="DQM35_RS00005"
    /old_locus_tag="NCTC12064_00001"
    /db_xref="GeneID:69899953"
  CDS
    1..1356
    /gene="dnaA"
    /locus_tag="DQM35_RS00005"
    /old_locus_tag="NCTC12064_00001"
    /inference="COORDINATES: similar to AA
    sequence:RefSeq:WP_012657571.1"
    /GO_function="GO:0003677 - DNA binding [Evidence IEA]"
    /GO_function="GO:0003688 - DNA replication origin binding
    [Evidence IEA]"
    /GO_function="GO:0005524 - ATP binding [Evidence IEA]"
    /GO_process="GO:0006270 - DNA replication initiation
    [Evidence IEA]"
    /GO_process="GO:0006275 - regulation of DNA replication
    [Evidence IEA]"
    /note="Derived by automated computational analysis using
    gene prediction method: Protein Homology."
    /codon_start=1
    /transl_table=11
    /product="chromosomal replication initiator protein DnaA"
    /protein_id="WP_002987659.1"
    /db_xref="GeneID:69899953"
    /translation="MTENEQIFWNRVLELAQSQLKQATYEFFVHDARLLKVKDHIATI
    YLDQMKELFWEKNLKDVLITAGFEVYNAQISVDYVFEEELMIEQNQTKINQPKQQAAL
    NSLPTVTSDLNSKYSFENFIQDENRWAVAASIAVANTPGTTYNPLFIWGGPGLGKTH
    LLNATIGNSVLLNPNAIKYITAENFINEFVIHRLDTMDLKEKFRNLDLLLIDDIQ
    SLAKKTLSGTQEEFNFTNALHNNKQIVLTSDRTPDHLNLEDRLVTRFKWGLTVNI
    TPDFETRVAILTNKIQEYNFIFPQDTIEYLAGQFDSNVRDLEGALKDISLVANFKQI
    DTITVDIAAEAIRARKQDGPKMTVIPIEEIQAQVGKFGYGVTVKEIKATKRTQNIIVLAR
    QVAMFLAREMTDNLSPKIGKEFGGRDHSTVLHAYNKIKNMISQDESRLIEIETIKNKI
```

Gene: includes
exons, introns,
UTRs

CDS: respects
strand, excludes
introns.
(use this! first
annotated
nucleotide = TSS)

GBFF file format

```
gene      complement(1233609..1234707)
          /gene="prfB"
          /locus_tag="DQM35_RS06355"
          /old_locus_tag="NCTC12064_01257"
          /old_locus_tag="NCTC12064_01257"
CDS       complement(join(1233609..1234634,1234636..1234707))
          /gene="prfB"
          /locus_tag="DQM35_RS06355"
          /old_locus_tag="NCTC12064_01257"
          /inference="COORDINATES: similar to AA
          sequence:RefSeq:WP_011889043.1"
          /ribosomal_slippage
          /GO_function="GO:0003747 - translation release factor
          activity [Evidence IEA]"
          /GO_process="GO:0006415 - translational termination
          [Evidence IEA]"
          /note="programmed frameshift; Derived by automated
          computational analysis using gene prediction method:
          Protein Homology."
          /codon_start=1
          /transl_table=11
          /product="peptide chain release factor 2"
          /protein_id="WP_096466706.1"
          /db_xref="GeneID:69901161"
          /translation="MEVAEIRQKIVENKEKLTFRRLDLDRLEEEIALLENHMTPEP
          FWNNDNIAAQKTSQELNELKGKYDTFHNMQELSDTELLLEMLDEDDSLKEELEENLMQ
          LDKIMGAYEMTLLSEPYDHNNAILIHPGSGGTEAQDWGDLRLMYTRFGNANGFKV
          EVLDYQAGDEAGIKSVTSLFEGPNAYGLLKSEMGVHRLVRI SPFDSAKRRHTSFASVE
          VMPELDDTIEVEVRDDDIKMDTFRSGGAGGQNVNKVSTGVRLTHIPTGIVVSSTVDRT
          QYGNRDRAMKMLQAKLYQLEQEKKAQEVDAKKGDKKEITWGSQIRSYVFTPYTMVKDH
          RTNFELAQVDKVMGDGEINGFIDAYLKWRIED"
```

1233609..1234634

means

“the region (inclusive) between 1233609
and 1234634”

join(r1start..r1end, r2start..r2end) means
append the two regions

complement(start..end) means take the
reverse complement of the region

Examples:

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

15..20,25..35



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



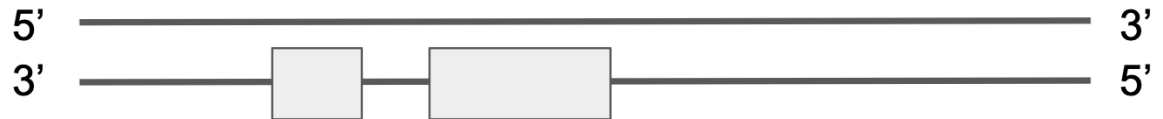
TSS is at 15

Examples:

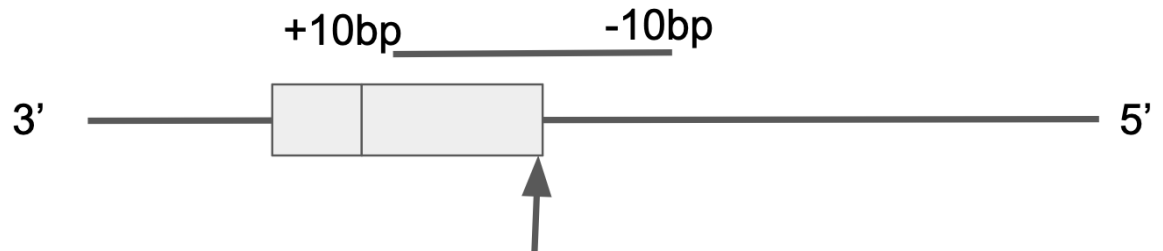
Example: `complement(join(15..20,25..35))`

15..20,25..35

- Coordinates on + strand
- But take sequence on reverse complement



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



TSS is at pos 35

Other Gotchas

Positions with a < or > represent unknown (but bounded) annotated features, just skip CDS's with ambiguous ends.

Consider each annotated CDS once (don't try to remove duplicates)

Building the Weights Matrix

1. Get background nucleotide counts & frequencies (whole genome, + and – strands!)
2. Get count matrices around each TSS (from -10 bp to +10 bp – 21 bp in total.)
3. Get TSS frequency matrix (normalized count matrix)
4. Get weight matrix:
 - $\text{weight} = \log_2(\text{nt freq in site model} / \text{nt freq in bg model})$
 - set weight = -99 for spots where the freq matrix = 0

Scoring a Candidate 3' Splice Site

Weights:

A	0.32	0.42	-0.18	-2.46	-5.29	-1.79	-3.45	1.64	-99.00	0.36	-0.13	-0.06
C	-0.60	-1.18	-1.15	-2.64	-3.51	-0.41	2.22	-99.00	-99.00	-0.20	0.06	0.33
G	-1.31	-1.35	-1.51	-3.35	-5.23	-1.30	-6.93	-99.00	2.48	0.79	-0.17	0.10
T	0.35	0.39	0.84	1.48	1.60	1.12	-1.24	-99.00	-99.00	-1.37	0.17	-0.22

T T C T T A C A G A A T

$$0.35 + 0.39 + -1.15 + 1.48 + 1.60 + -1.79 + 2.22 + 1.64 + 2.48 + 0.36 + -0.13 + -0.22 = 7.23$$

Scoring nucleotides



1. Compute scores for *every* nucleotide in the genome
 - Scores are associated with nucleotide at the center of the window!
 - Don't need to extend past the genome ends (i.e. start at index 10, end at index genome_length - 10)
 - forward and reverse strands

This HW is deterministic -- you can match the test output exactly!