

Genome 540 Discussion

January 15th, 2024

Clifford Rostomily

Agenda

- Assignment 2
- Assignment 3

Assignment 2 questions?

1. Using an input sequence generate 3 sequences of the same length using:
 - a. the length (equal frequency assumption)
 - b. nucleotide frequency (order 0-Markov)
 - c. dinucleotide frequency (order 1-Markov)
2. Run your program from HW1 on each of those sequences
 - a. Sequence 1 should always be the 10Mb human region from HW1,
 - b. Sequence 2 should be your simulated sequence



Assignment 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      U00996               4641652 bp    DNA     circular BCT 01-AUG-2014
DEFINITION Escherichia coli str. K-12 substr. MG1655, complete genome.
ACCESSION  U00996
VERSION   U00996.3
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; Enterobacteriales;
          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"
     gene              /sub_strain="MG1655"
                        /db_xref="taxon:511145"
                        190..255
                        /gene="thrL"
                        /locus_tag="b0001"
                        /gene_synonym="ECK0001"
     CDS                /gene_synonym="JW4367"
                        /db_xref="EcoGene:EG11277"
                        190..255
                        /gene="thrL"
                        /locus_tag="b0001"
                        /gene_synonym="ECK0001"
                        /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-0000006"
                        /db_xref="UniProtKB/Swiss-Prot:P0AD86"
                        /db_xref="EcoGene:EG11277"
                        /translation="MKRISITITITITITGGNGAG"
```

Sequence

```
ORIGIN
1   agcttttcat tctgactgca acgggcaata tgtctctgtg tggattaaaa aaagagtgtc
61  tgatagcagc ttctgaactg gttaccctgc gfgagtaaat taaaatttta ttgacttagg
121 tcactaaaaa ctttaaccaa tataggcata gcgcacagac agataaaaaa tacagagtag
181 acaacatcca tgaacgcat  tagcaccacc attaccacca ccatcaccat taccacaggt
241 aacggtgctg gctgacgcgt acaggaaaac cagaaaaaag cccgcacctg acagtgcggg
301 cttttttttt cgaccaaaag taacgaggta acaacatgc  gagtgttgaa gttcggcggg
361 acatcagttg caaatgcaga acgttttctg cgtgttcgcc atattctgga aagcaatgcc
421 aggcaggggc aggtggccac cgtcctctct gccccgccca aaatcaccaa ccactgggtg
481 cgcgatgatt aaaaaccat  tagcggccag gatcgtttac ccaatcacag cgatgccgaa
541 cgtatttttg ccgaactttt gacgggactc gccccgcccc agccgggggt cccgtggcgg
601 caattgaaaa ctttcgtcga tcaggaattt gcccaataaa aacatgtcct gcatggcatt
661 agtttgtttg ggcagtgccc ggatgatcat aacgctgcgc tgatttgccc tggcgagaaa
721 atgtcgcagc ccattatggc cggcgtatta gaagcgcggc gtcacaacgt tactgtttac
781 gatccggtgc aaaaactgct ggcagtgggg cattaacctg aatctaccgt cgatattgtc
841 gggttcaacc gcgctattgc ggcaagccgc attccggctg atcacatggt gctgatggca
901 ggtttcacgc cggctaatga aaaaagcgaa ctggtgtgtc ttggaccgaa cgtttccgac
961 tactctctgc cgggtctggc tgcctgttta gcgcgcgatt gttcgcagat gttgcagcagc
1021 gttgacgggg tctatactcg gaccgccggt cagggtccgc atgcgaggtt gttgaagtgc
1081 atgtcctacc aggaagcagt gtagctttcc tacttgcggc ctaaaagtct tcaccgccgc
1141 accattacc  ccacgcccca gttccagatc ccttgcttga ttaaaaatac cggaaatcct
1201 caagcaccag tgaactcctat tggctgcagg cgtgatgaag acgaaattacc ggtcaagggtc
1261 atttccaatc tgaataacat ggcaatgttc agcgtttctg gtcgggggat gaaggagtag
1321 gtccgatcgt cggcgccgct ctttgcagcg atgcacccgc cccgtatttc cgtgtgtgctg
1381 attarcgaat catcttcgca atacagcatc agtttctggt ttccacaagc cgaactgtgtg
1441 cgagctgaac gggcaatgca ggaagagttc tacctggaac tgaagaaggg cttactggag
1501 ccgtgctgac tgacggaaag gctggcattt atctcgggtg taggtgatgg tatgcaccac
1561 ttgcgtggga tctgcggcaa attctttgcc gcactggccc gcgccaatat caacattgtc
```

gbff Features

```
gene      complement(736161..737503)
         /locus_tag="DQM35_RS03885"
         /old_locus_tag="NCTC12064_00760"
         /db_xref="GeneID:69900688"
CDS       complement(join(736161..737053,737053..737503))
         /locus_tag="DQM35_RS03885"
         /old_locus_tag="NCTC12064_00760"
         /inference="COORDINATES: similar to AA
         sequence:RefSeq:WP_076611514.1"
         /ribosomal_slippage
         /GO_function="G0:0004803 - transposase activity [Evidence
         IEA]"
         /note="programmed frameshift; Derived by automated
         computational analysis using gene prediction method:
         Protein Homology."
         /codon_start=1
         /transl_table=11
         /product="IS3 family transposase"
         /protein_id="WP_172450158.1"
         /db_xref="GeneID:69900688"
         /translation="MKFNQETKVKVIYELRQMGESIKSIPKFKDMAESDLKYMIRLIDR
         YGVTIVQKCKNHYYSPELKQEIINKVLIDGQSQKQTSLDYALPTSSMLSRWIAQYKKN
         GYTILEKPRGRPSKMGKRKRKNLEEMTEVERLOKELEYLRAENAVLKKPERIPLERRS
         KTQRATEIIQALRNQFPLEMLLEILDLSRSTYYYQVKRLAQGDKDIELKHVIREIYDE
         HKGNYGYRRIHMLRNRRGFVNVHKKVQRLMKVMGLAARIRRRKRYSSYKGEVGKKADN
         LIKRHFVKGSKPYEKCYTDVTELALPEGKLYLLPVLDDGYNSEIIDFTLSRSPNLKQVQT
         MLEKTFPADSYSGTILHSDQGWQYQHQSYPHDFLESKGIPLPSMRKGNPNDGMMDSFF
         GILKSEMFYGLETTYQSLDKLEEAITDYIFYNNKRIKAKLKGFSVPVQYRTKSFQ"
```

Gene + introns

Strand + exons

- 736161..737053
 - Specifies a coding region
 - End position is 1 *greater* than actual end
- join(...)
 - Join coding sequences
- complement(...)
 - Take the reverse complement

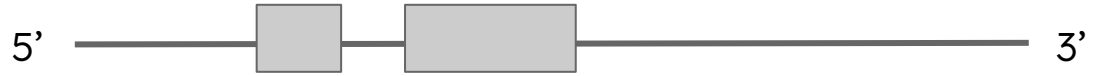
Peptide product

Warning: may not match sequence

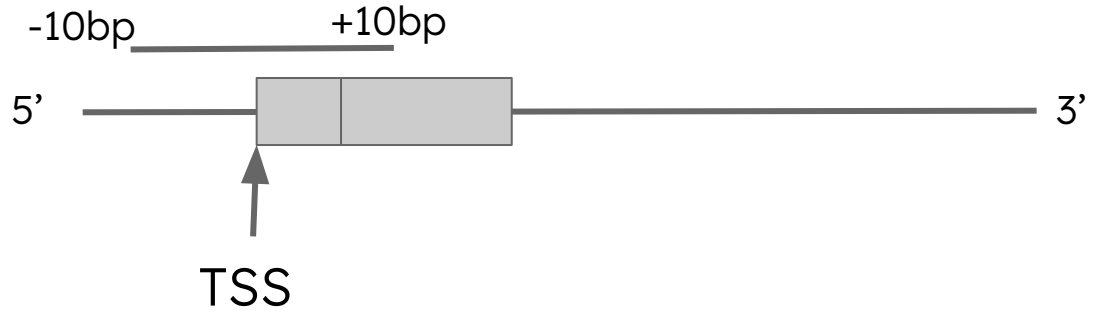
join(...) example

Example: join(15..20,25..30)

15..20,25..30



join(15..20,25..30)



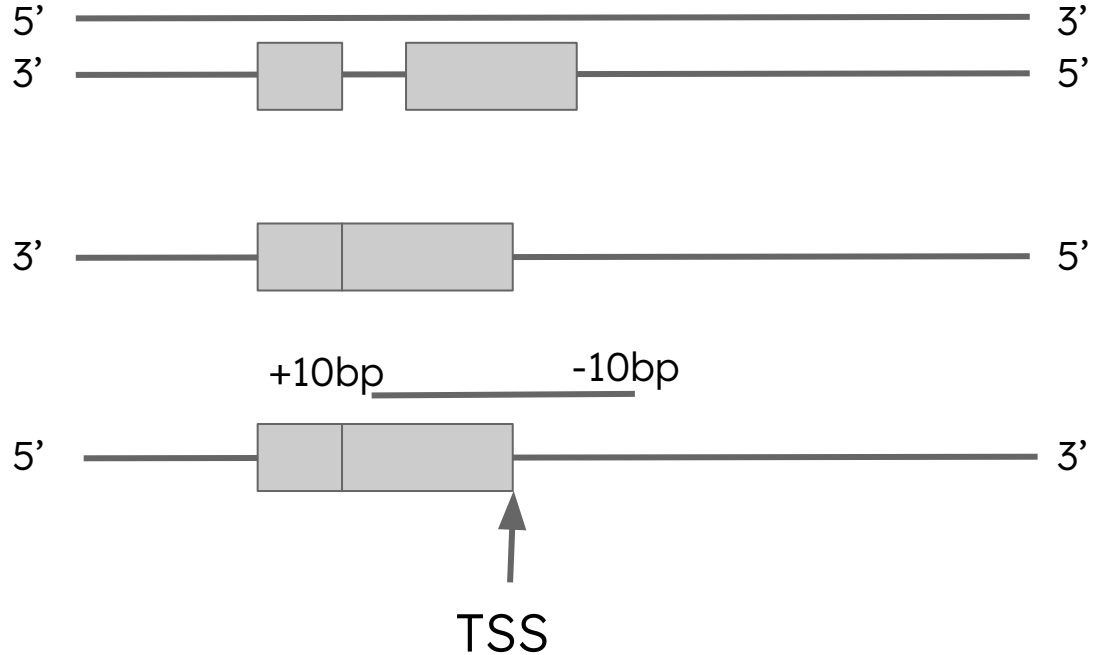
complement(join(...)) example

Example: complement(join(15..20,25..30))

15..20,25..30

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

join(15..20,25..30)



Duplicates

```
CDS
join(2265392..2265394,2266033..2266077,2266183..2266408,
2266762..2266904,2267059..2267170,2267600..2267727,
2267877..2267965,2268483..2268627,2268962..2269041,
2269532..2269640,2270513..2270677,2270818..2270921,
2271221..2271405,2271571..2271777,2272250..2272473,
2272625..2272751,2272946..2273025)
/gene="LOH11CR2A"
/note="Derived by automated computational analysis using
gene prediction method: Gnomon."
/codon_start=1
/product="von Willebrand factor A domain-containing
protein 5A isoform X4"
/protein_id="XP_004948513.1"
/db_xref="GeneID:419937"
/db_xref="CGNC:108"
/translation="MACSEDAKIKAVLQDETQQLYRGSTGEGENFDYLQYEVTSSEGV
FACFLGSLSPGKEMVVTLRVYQELSRKPDGAAQFMLPSTMHPYKTHYTCNCRTKGLHY
SLLLTASLQSPRGVADVQANCALTPLIYTAQDHSTAQVSLAGTPPNHLELLVYVREP
TAVSVVVEKGDVPVATAGSLLGDSLVLVTLAPNIHDAKPGQCKSGEFIFVLDSTSLEHA
QDPLFLLLKSLPLGCFYNIYCYGATPVGIYQPSVEYTDNLNEMQLISTTGSRLGDT
DLGLTLRTIYSTPRPCGHARQLFIMSELPPDTEAIAAEVCHHRNSHRCFSFCFSTDS
VSLATALARETDGEAVYVSSDNVIVQVLKCLKQALKPVAEGVSLWTLPSGLEVEVLG
GTPQFIFQGQHIFLYAQIHGKEQDMKEASGVMTLHFNLDGQDVTHKIQFPLCPQGDGR
MAGHHLAARHLEKLLPEVVRGSGDEPMQRAIEISLTSGIICPFTSYVGVRTSRRAP
WYHGPLALLSPRQSFVPCKILLLRGLSDTSTCFPKTIWPPRWHAVQESRIATKRLT
NGIANLLQHGAHKEAPEQPPPSIFSLKYVDSTRFVLCSQIFGPMWNEAIAECRELVAL
QNVDSWTLSSGLASVLQVEEAIEIKGMPGEVMEPSFWATVLAVTWLRQDNRRYHELCL
ELLEAKAVTWLCSRDSVQLDKCLEASNTLLGSSVSPSVFRL"
```

```
CDS
join(2265392..2265394,2266033..2266077,2266183..2266408,
2266762..2266904,2267059..2267170,2267600..2267727,
2267877..2267965,2268483..2268627,2268962..2269041,
2269532..2269640,2270513..2270677,2270818..2270921,
2271221..2271405,2271571..2271777,2272250..2272473,
2272625..2272751,2272946..2273025)
/gene="LOH11CR2A"
/note="Derived by automated computational analysis using
gene prediction method: Gnomon."
/codon_start=1
/product="von Willebrand factor A domain-containing
protein 5A isoform X4"
/protein_id="XP_024999836.1"
/db_xref="GeneID:419937"
/db_xref="CGNC:108"
/translation="MACSEDAKIKAVLQDETQQLYRGSTGEGENFDYLQYEVTSSEGV
FACFLGSLSPGKEMVVTLRVYQELSRKPDGAAQFMLPSTMHPYKTHYTCNCRTKGLHY
SLLLTASLQSPRGVADVQANCALTPLIYTAQDHSTAQVSLAGTPPNHLELLVYVREP
TAVSVVVEKGDVPVATAGSLLGDSLVLVTLAPNIHDAKPGQCKSGEFIFVLDSTSLEHA
QDPLFLLLKSLPLGCFYNIYCYGATPVGIYQPSVEYTDNLNEMQLISTTGSRLGDT
DLGLTLRTIYSTPRPCGHARQLFIMSELPPDTEAIAAEVCHHRNSHRCFSFCFSTDS
VSLATALARETDGEAVYVSSDNVIVQVLKCLKQALKPVAEGVSLWTLPSGLEVEVLG
GTPQFIFQGQHIFLYAQIHGKEQDMKEASGVMTLHFNLDGQDVTHKIQFPLCPQGDGR
MAGHHLAARHLEKLLPEVVRGSGDEPMQRAIEISLTSGIICPFTSYVGVRTSRRAP
WYHGPLALLSPRQSFVPCKILLLRGLSDTSTCFPKTIWPPRWHAVQESRIATKRLT
NGIANLLQHGAHKEAPEQPPPSIFSLKYVDSTRFVLCSQIFGPMWNEAIAECRELVAL
QNVDSWTLSSGLASVLQVEEAIEIKGMPGEVMEPSFWATVLAVTWLRQDNRRYHELCL
ELLEAKAVTWLCSRDSVQLDKCLEASNTLLGSSVSPSVFRL"
```

No special handling needed, just use each CDS entry once regardless of if it is a duplicate

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. $\text{Weight} = \log_2([\text{nt freq at motif position}] / [\text{background nt freq}])$
 - 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

Precision of floating point numbers

- Float
 - 32-bit
 - 1 bit for the sign, 8 bits for the exponent, and 23 for the value
 - 7 decimal digits of precision
- Double
 - 1 bit for the sign, 11 bits for the exponent, and 52 bits for the value.
 - 15 decimal digits of precision
- For this homework **use doubles** over floats

Reminders

- HW2 due this Sunday, 11:59pm
- Please have your name in the filename of your homework assignment and match the template