Week 3
Discussion Section

Chengxiang Qiu

1/21/2021

Some comments on HW1

1. Please make an effort to match the template!

2. You only need to submit result on the real data. The test data and the template are used to help you debugging.

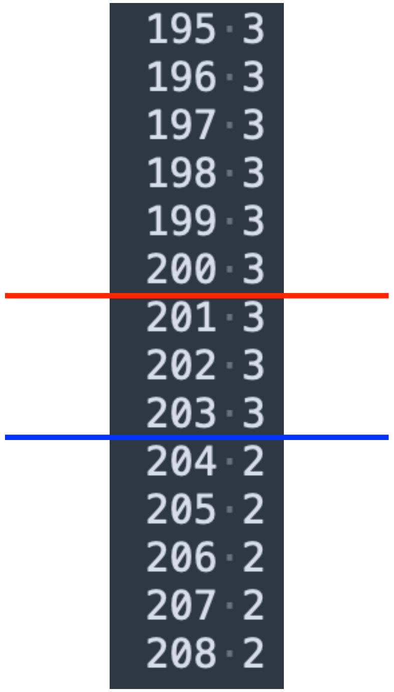
3. Please provide language and runtime.

Error 1: the position of the longest match substring is 1-bp shifting.

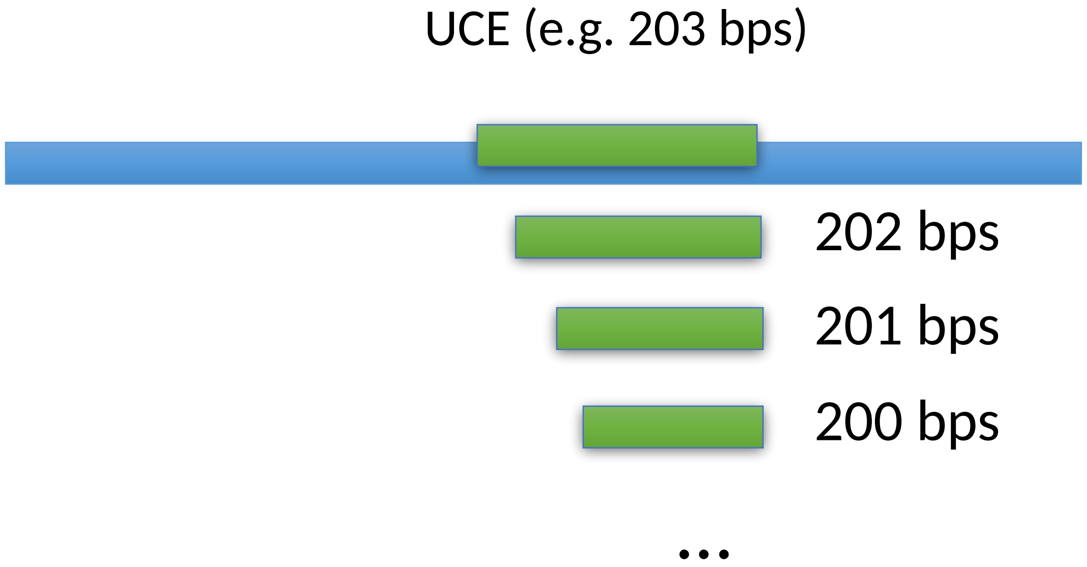
Error 2: the description of the longest match substring is not correct.

The extra credit question - how many UCEs?

Histogram (example)



```
195·3
196·3
197·3
198·3
199·3
200·3
201·3
202·3
203·3
204·2
205·2
206·2
207·2
208·2
```



UCE (e.g. 203 bps)

202 bps

201 bps

200 bps

…

HW2 Questions?

Program: Generate FASTA files of simulated genome using order-0 Markov and order-1 Markov models

- Run your HW1 program twice
  - Human sequence & Simulated mouse sequence using order-0 Markov model
  - Human sequence & Simulated mouse sequence using order-1 Markov model

  - what can you conclude about the statistical significance of matches between the orthologous mouse and human regions in homework 1?

- Due Jan 23, 11:59pm

HW3 Questions?

HW3: create a motif model for TSSs

- Due 11:59pm on Sunday, Jan. 30
- Assignment:
  - Parse a Genbank file (gbff format) with sequence info and annotated CDS locations
    - Write your own code to parse the file! Do not use a third-party Genbank file parser.
  - Using the CDS information, compute a site weight matrix for a 21bp motif centered at the translation start site
  - Using the weight matrix, compute scores for annotated CDS translation start sites and for non-annotated positions

Genbank flat file format (.gbff)

- Feature list
  - Each locus has entries for gene, mRNA, and CDS
  - CDS features are coding sequences (these are the entries we care about)
  - ‘complement’ indicates the reverse complement
- ORIGIN
  - Located after the feature list, at the end of the file
  - Contains the genome sequence

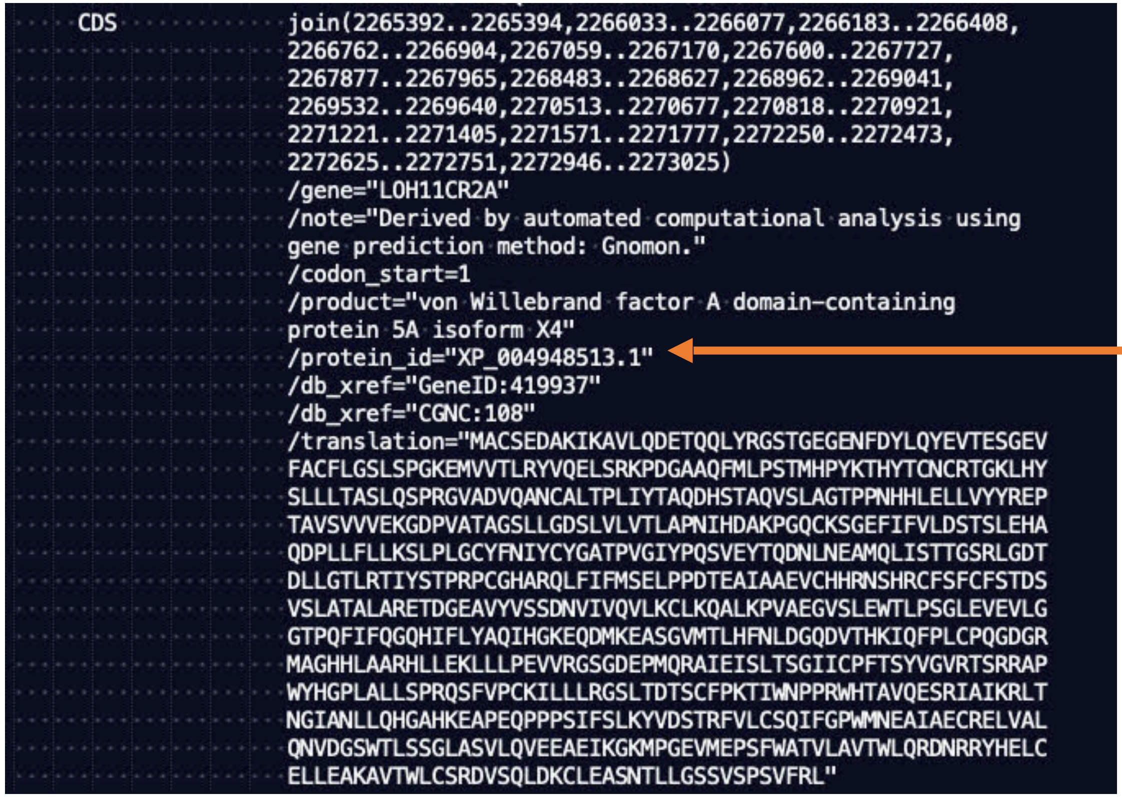
Genbank flat file format (.gbff)

```
FEATURES             Location/Qualifiers
     source          1..2895605
                     /organism="Plasmodium falciparum 3D7"
                     /mol_type="genomic DNA"
                     /isolate="3D7"
                     /db_xref="taxon:36329"
                     /chromosome="13"
     gene            21467..28890
                     /gene="VAR"
                     /locus_tag="MAL13P1.1"
                     /db_xref="GeneID:813647"
     mRNA            join(<21467..26641,27577..>28890)
                     /gene="VAR"
                     /locus_tag="MAL13P1.1"
                     /transcript_id="XM_001349702.1"
                     /db_xref="GI:124512763"
                     /db_xref="GeneID:813647"
     CDS             join(21467..26641,27577..28890)
                     /gene="VAR"
                     /locus_tag="MAL13P1.1"
                     /codon_start=1
                     /product="erythrocyte membrane protein 1, PfEMP1"
                     /protein_id="XP_001349738.1"
                     /db_xref="GI:124512764"
                     /db_xref="GOA:Q8IEV1"
                     /db_xref="InterPro:IPR008602"
                     /db_xref="UniProtKB/TrEMBL:Q8IEV1"
                     /db_xref="GeneID:813647"
                     /translation="MGPPGITGTQGETAKHMFDRIGKQVYETVKNEAENYISELEGKL
SQATLLGERVSSLKTCQLVEDYRSKANGDVKRYPCANRSPVRFSDESRSQCTYNRIKD…"
.
.
.

ORIGIN
        1 taaaccctga accctaaacc ctaaaccctg aaccctaaac cctaaaccct aaacctaaac
       61 ctaaaccctg aaccctaaac cctgaaccct gaaccctaaa ccctaaaccc tgaaccctaa…
```

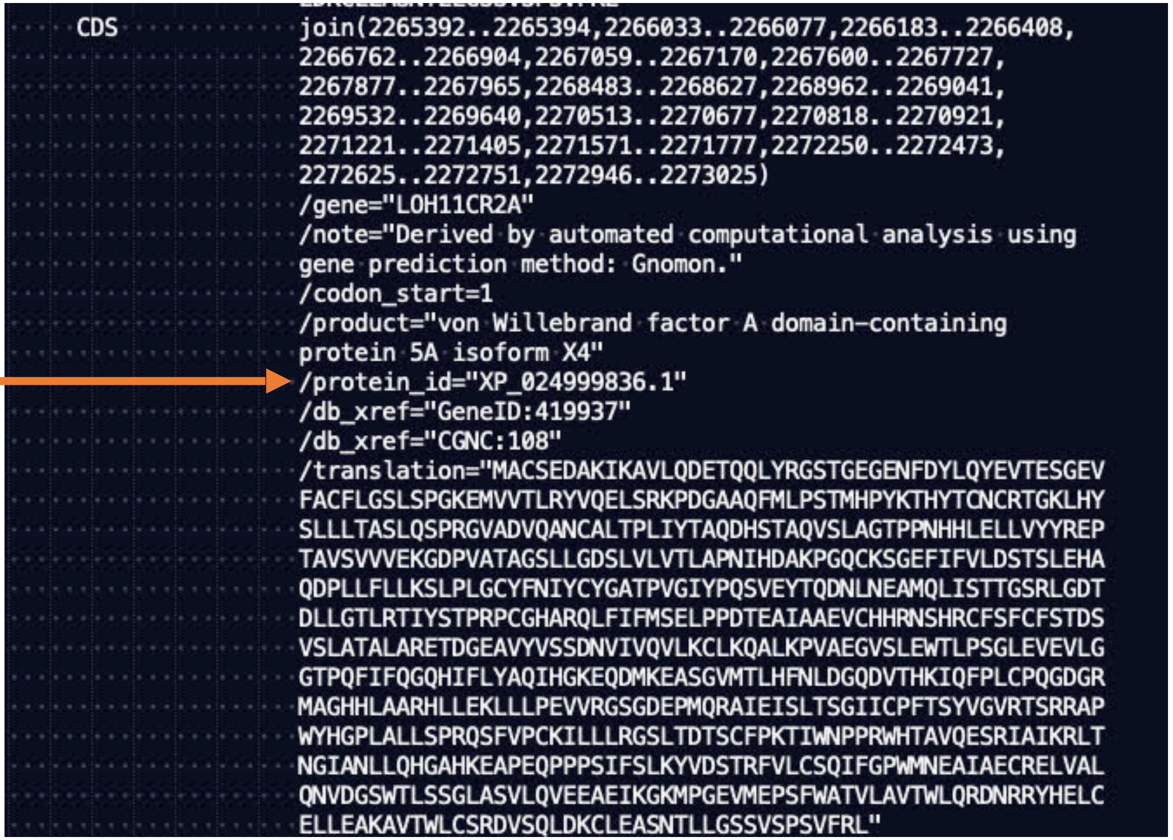
Some more CDS examples

```
CDS             96094..97215
                    /locus_tag="PTSG_00022"
                    /codon_start=1
                    /product="hypothetical protein"
                    /protein_id="EGD72006.1"
                    /db_xref="GI:326426436"
                    /translation="MVVAAGSGGASRPTNAPSCPLCPGGSVGGAVLMVVPLLVCIALL
                    AGCLSVSSLWRRNKRQRHAPQYASTCASGRAKPNKRAAPRVQPDLRLPHQQQQPQHPQ...“


CDS             join(10183..10943,11138..11246,11408..11525,11697..11815,
                    12006..12056,12284..12445,12661..12792,12989..13135,
                    13293..13400,13597..13661,13848..13957,14104..14208,
                    14364..14440,14606..14773,14909..15013)
                    /locus_tag="PTSG_00005"
                    /codon_start=1
                    /product="hypothetical protein"
                    /protein_id="EGD71989.1"
                    /db_xref="GI:326426419"
                    /translation="MMMMMMMMRPCCSLPSTWWLVVVVLAAACCAATPTAAAVPAAAP
                    AEAADPSVVNVGQFVVSLDEDGVLSAVRNPAQMPNPHLAWHSTGEILEVAASKMYLHG...“


CDS             complement(join(15291..15934,16108..16234,16358..16394,
                    16582..16790,17086..17196,17376..17456,17810..17877,
                    18020..18060,18199..18256,18556..18598,18767..19187,
                    19334..19410,19552..19631,19795..19917,20098..20183,
                    20449..20577,20789..20904,21261..21449,21667..21787,
                    21936..22108,22453..22549,22808..22934,23895..23970,
                    24140..24246,24389..27209))
                    /locus_tag="PTSG_11525"
                    /codon_start=1
                    /product="hypothetical protein"
                    /protein_id="EGD71990.1"
                    /db_xref="GI:326426420"
                    /translation="MWRSWRHGEVGSGVAGGENGKDAQQASSNSHGSHGSHGSNHPNG
                    NHGGSSDNVGSSHDERSSSDREQERGQVQRRKRRHARMHEKHASNHAASSVARPSRLT...“
```

Handling ‘Duplicate’ Entries

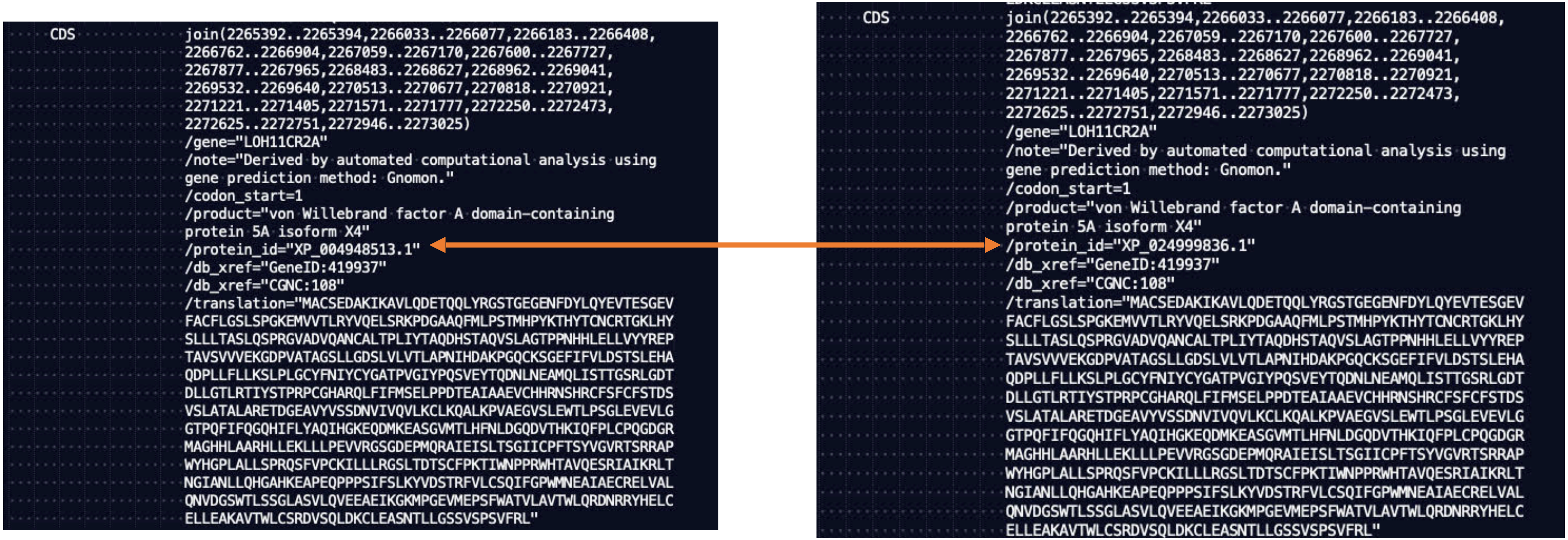


```
     CDS             join(2265392..2265394,2266033..2266077,2266183..2266408,
                     2266762..2266904,2267059..2267170,2267600..2267727,
                     2267877..2267965,2268483..2268627,2268962..2269041,
                     2269532..2269640,2270513..2270677,2270818..2270921,
                     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="MACSEDAKIKAVLQDETQQLYRGSTGEGENFDYLQYEVTESGEV
                     FACFLGSLSPGKEMVVTLRYVQELSRKPDGAAQFMLPSTMHPYKTHYTCNCRTGKLHY
                     SLLLTASLQSPRGVADVQANCALTPLIYTAQDHSTAQVSLAGTPPNHHLELLVYYREP
                     TAVSVVVEKGDPVATAGSLLGDSLVLVTLAPNIHDAKPGQCKSGEFIFVLDSTSLEHA
                     QDPLLFLLKSLPLGCYFNIYCYGATPVGIYPQSVEYTQDNLNEAMQLISTTGSRLGDT
                     DLLGTLRTIYSTPRPCGHARQLFIFMSELPPDTEAIAAEVCHHRNSHRCFSFCFSTDS
                     VSLATALARETDGEAVYVSSDNVIVQVLKCLKQALKPVAEGVSLEWTLPSGLEVEVLG
                     GTPQFIFQGQHIFLYAQIHGKEQDMKEASGVMTLHFNLDGQDVTHKIQFPLCPQGDGR
                     MAGHHLAARHLLEKLLLPEVVRGSGDEPMQRAIEISLTSGIICPFTSYVGVRTSRRAP
                     WYHGPLALLSPRQSFVPCKILLLRGSLTDTSCFPKTIWNPPRWHTAVQESRIAIKRLT
                     NGIANLLQHGAHKEAPEQPPPSIFSLKYVDSTRFVLCSQIFGPWMNEAIAECRELVAL
                     QNVDGSWTLSSGLASVLQVEEAEIKGKMPGEVMEPSFWATVLAVTWLQRDNRRYHELC
                     ELLEAKAVTWLCSRDVSQLDKCLEASNTLLGSSVSPSVFRL"
```



```
     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="MACSEDAKIKAVLQDETQQLYRGSTGEGENFDYLQYEVTESGEV
                     FACFLGSLSPGKEMVVTLRYVQELSRKPDGAAQFMLPSTMHPYKTHYTCNCRTGKLHY
                     SLLLTASLQSPRGVADVQANCALTPLIYTAQDHSTAQVSLAGTPPNHHLELLVYYREP
                     TAVSVVVEKGDPVATAGSLLGDSLVLVTLAPNIHDAKPGQCKSGEFIFVLDSTSLEHA
                     QDPLLFLLKSLPLGCYFNIYCYGATPVGIYPQSVEYTQDNLNEAMQLISTTGSRLGDT
                     DLLGTLRTIYSTPRPCGHARQLFIFMSELPPDTEAIAAEVCHHRNSHRCFSFCFSTDS
                     VSLATALARETDGEAVYVSSDNVIVQVLKCLKQALKPVAEGVSLEWTLPSGLEVEVLG
                     GTPQFIFQGQHIFLYAQIHGKEQDMKEASGVMTLHFNLDGQDVTHKIQFPLCPQGDGR
                     MAGHHLAARHLLEKLLLPEVVRGSGDEPMQRAIEISLTSGIICPFTSYVGVRTSRRAP
                     WYHGPLALLSPRQSFVPCKILLLRGSLTDTSCFPKTIWNPPRWHTAVQESRIAIKRLT
                     NGIANLLQHGAHKEAPEQPPPSIFSLKYVDSTRFVLCSQIFGPWMNEAIAECRELVAL
                     QNVDGSWTLSSGLASVLQVEEAEIKGKMPGEVMEPSFWATVLAVTWLQRDNRRYHELC
                     ELLEAKAVTWLCSRDVSQLDKCLEASNTLLGSSVSPSVFRL"
```

Handling ‘Duplicate’ Entries



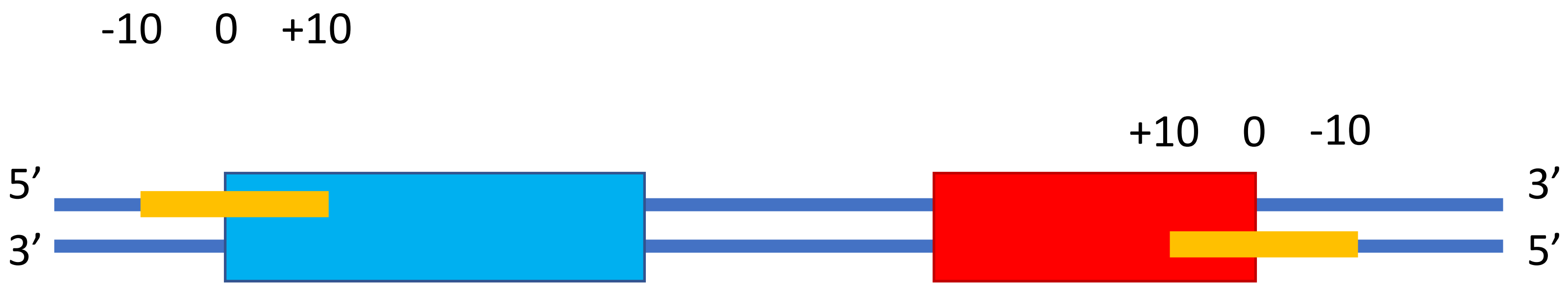
```
     CDS             join(2265392..2265394,2266033..2266077,2266183..2266408,
                     2266762..2266904,2267059..2267170,2267600..2267727,
                     2267877..2267965,2268483..2268627,2268962..2269041,
                     2269532..2269640,2270513..2270677,2270818..2270921,
                     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="MACSEDAKIKAVLQDETQQLYRGSTGEGENFDYLQYEVTESGEV
                     FACFLGSLSPGKEMVVTLRYVQELSRKPDGAAQFMLPSTMHPYKTHYTCNCRTGKLHY
                     SLLLTASLQSPRGVADVQANCALTPLIYTAQDHSTAQVSLAGTPPNHHLELLVYYREP
                     TAVSVVVEKGDPVATAGSLLGDSLVLVTLAPNIHDAKPGQCKSGEFIFVLDSTSLEHA
                     QDPLLFLLKSLPLGCYFNIYCYGATPVGIYPQSVEYTQDNLNEAMQLISTTGSRLGDT
                     DLLGTLRTIYSTPRPCGHARQLFIFMSELPPDTEAIAAEVCHHRNSHRCFSFCFSTDS
                     VSLATALARETDGEAVYVSSDNVIVQVLKCLKQALKPVAEGVSLEWTLPSGLEVEVLG
                     GTPQFIFQGQHIFLYAQIHGKEQDMKEASGVMTLHFNLDGQDVTHKIQFPLCPQGDGR
                     MAGHHLAARHLLEKLLLPEVVRGSGDEPMQRAIEISLTSGIICPFTSYVGVRTSRRAP
                     WYHGPLALLSPRQSFVPCKILLLRGSLTDTSCFPKTIWNPPRWHTAVQESRIAIKRLT
                     NGIANLLQHGAHKEAPEQPPPSIFSLKYVDSTRFVLCSQIFGPWMNEAIAECRELVAL
                     QNVDGSWTLSSGLASVLQVEEAEIKGKMPGEVMEPSFWATVLAVTWLQRDNRRYHELC
                     ELLEAKAVTWLCSRDVSQLDKCLEASNTLLGSSVSPSVFRL"
```

```
     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="MACSEDAKIKAVLQDETQQLYRGSTGEGENFDYLQYEVTESGEV
                     FACFLGSLSPGKEMVVTLRYVQELSRKPDGAAQFMLPSTMHPYKTHYTCNCRTGKLHY
                     SLLLTASLQSPRGVADVQANCALTPLIYTAQDHSTAQVSLAGTPPNHHLELLVYYREP
                     TAVSVVVEKGDPVATAGSLLGDSLVLVTLAPNIHDAKPGQCKSGEFIFVLDSTSLEHA
                     QDPLLFLLKSLPLGCYFNIYCYGATPVGIYPQSVEYTQDNLNEAMQLISTTGSRLGDT
                     DLLGTLRTIYSTPRPCGHARQLFIFMSELPPDTEAIAAEVCHHRNSHRCFSFCFSTDS
                     VSLATALARETDGEAVYVSSDNVIVQVLKCLKQALKPVAEGVSLEWTLPSGLEVEVLG
                     GTPQFIFQGQHIFLYAQIHGKEQDMKEASGVMTLHFNLDGQDVTHKIQFPLCPQGDGR
                     MAGHHLAARHLLEKLLLPEVVRGSGDEPMQRAIEISLTSGIICPFTSYVGVRTSRRAP
                     WYHGPLALLSPRQSFVPCKILLLRGSLTDTSCFPKTIWNPPRWHTAVQESRIAIKRLT
                     NGIANLLQHGAHKEAPEQPPPSIFSLKYVDSTRFVLCSQIFGPWMNEAIAECRELVAL
                     QNVDGSWTLSSGLASVLQVEEAEIKGKMPGEVMEPSFWATVLAVTWLQRDNRRYHELC
                     ELLEAKAVTWLCSRDVSQLDKCLEASNTLLGSSVSPSVFRL"
```

- The specific sequences were annotated by the RefSeq genome annotation pipeline (more info here), which is supposed to generate non-redundant annotations.

- **Consider each CDS entry listed in the file one time, regardless of whether there are other CDS entries that are similar/identical/overlapping.**

Computing a TSS site weight matrix



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

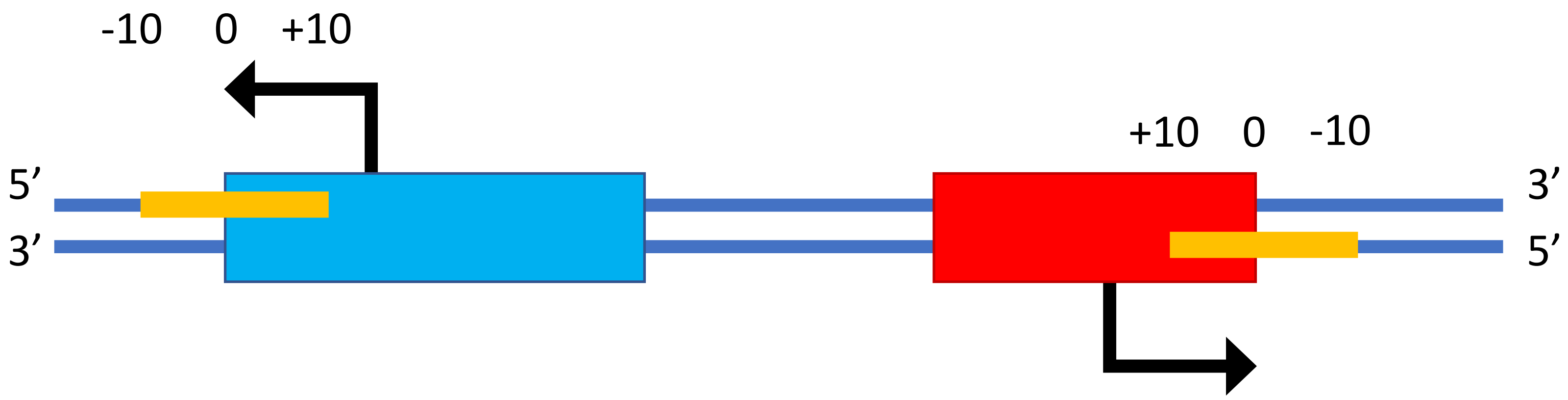
Step 1: Count matrix – record the number of times each nucleotide shows up at each motif position (-10 to +10).

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 ($2^{-99} = 1.6\text{x}10^{-30} \approx 0$)

Computing site scores



- 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)

Noncontiguous CDSs

- Positions downstream of the translation start site could be noncontiguous
  - join(1000…1008, 1200…1500)
- How would you construct the TSS motif?

Noncontiguous CDSs

- Positions downstream of the translation start site could be noncontiguous
  - join(1000…1008, 1200…1500)
- How would you construct the TSS motif?

```
-10 -9  -8  -7  -6  -5  -4  -3  -2  -1  0    1    2    3    4    5    6    7    8    9
10
990 991 992 993 994 995 996 997 998 999 1000 1001 1002 1003 1004 1005 1006 1007 1008
1200 1201
```

- Note in the gbff that ranges are **one indexed** and inclusive on both ends.

Reporting 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
```

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
...
```

HW3 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
- Score histogram for annotated CDS TSSs
- Score histogram for all positions
- List of non-CDS positions with score >=10

HW3 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.)