Genome 540 Discussion

Conor Camplisson

January 31st, 2023

Genome Sciences
UNIVERSITY OF WASHINGTON

Outline

- Homework 3 wrap-up

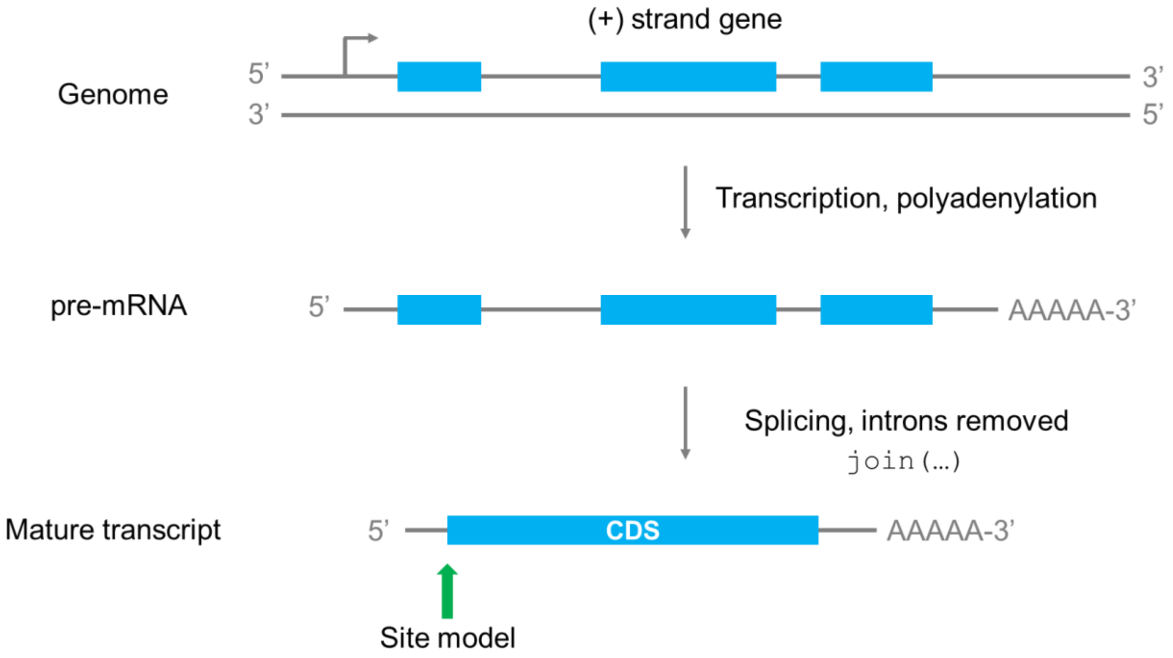
- Homework 4 overview & questions

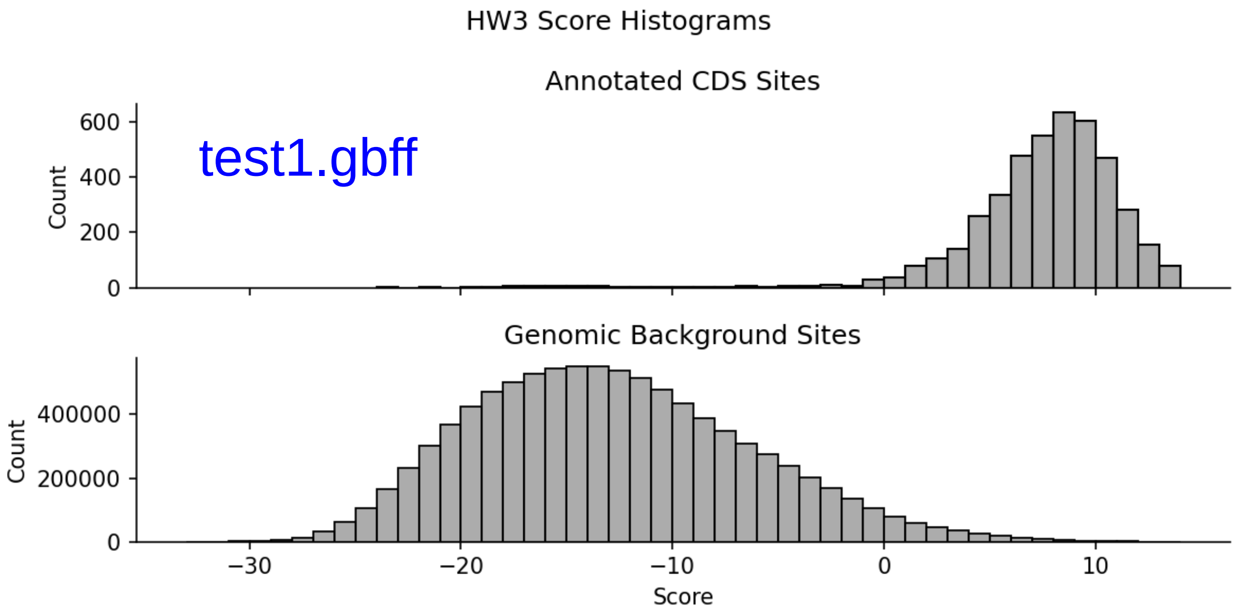
Outline

- Homework 3 wrap-up

- Homework 4 overview & questions

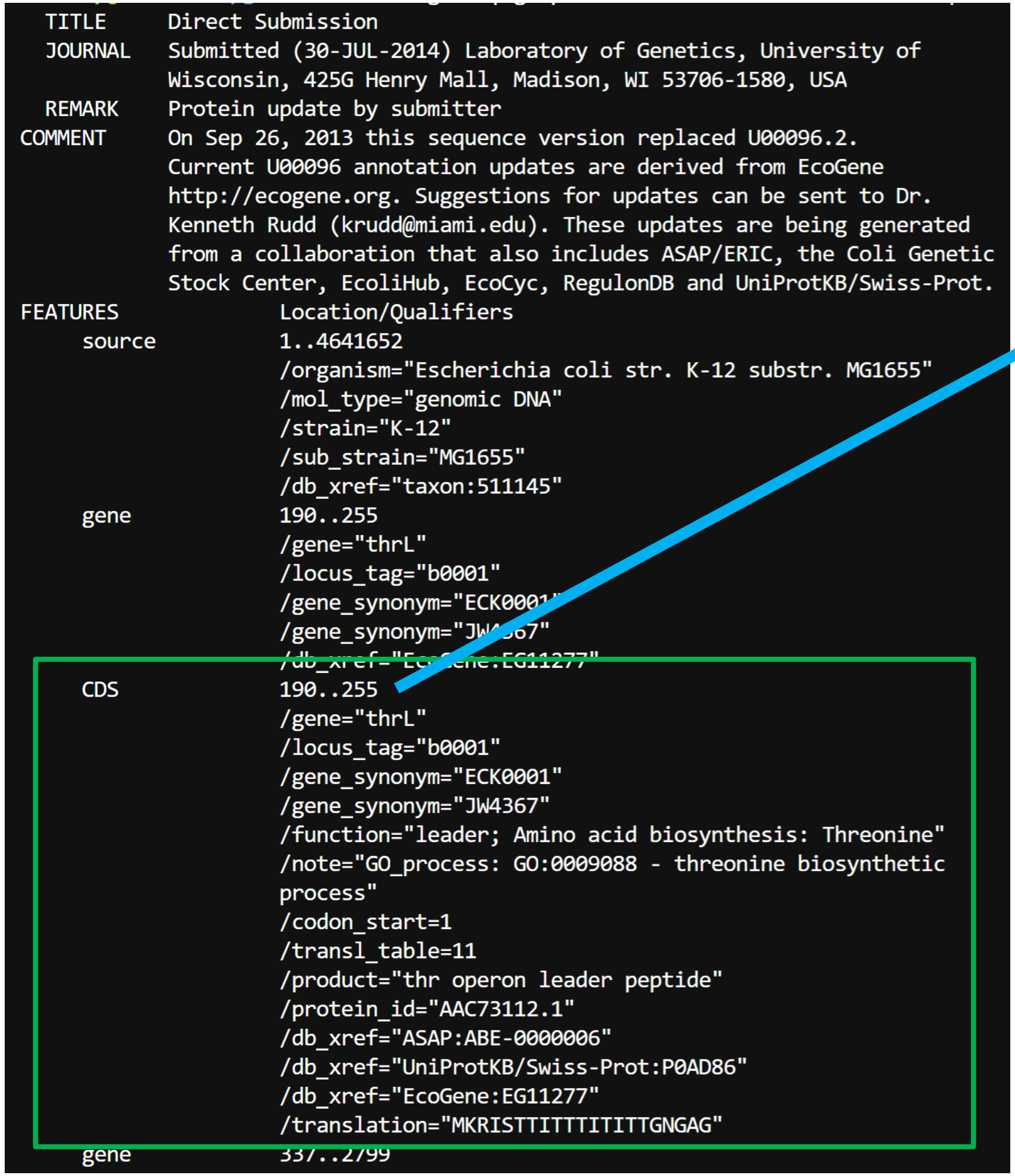
Homework 3 Wrap-up





test1.gbff

Homework 3 Wrap-up



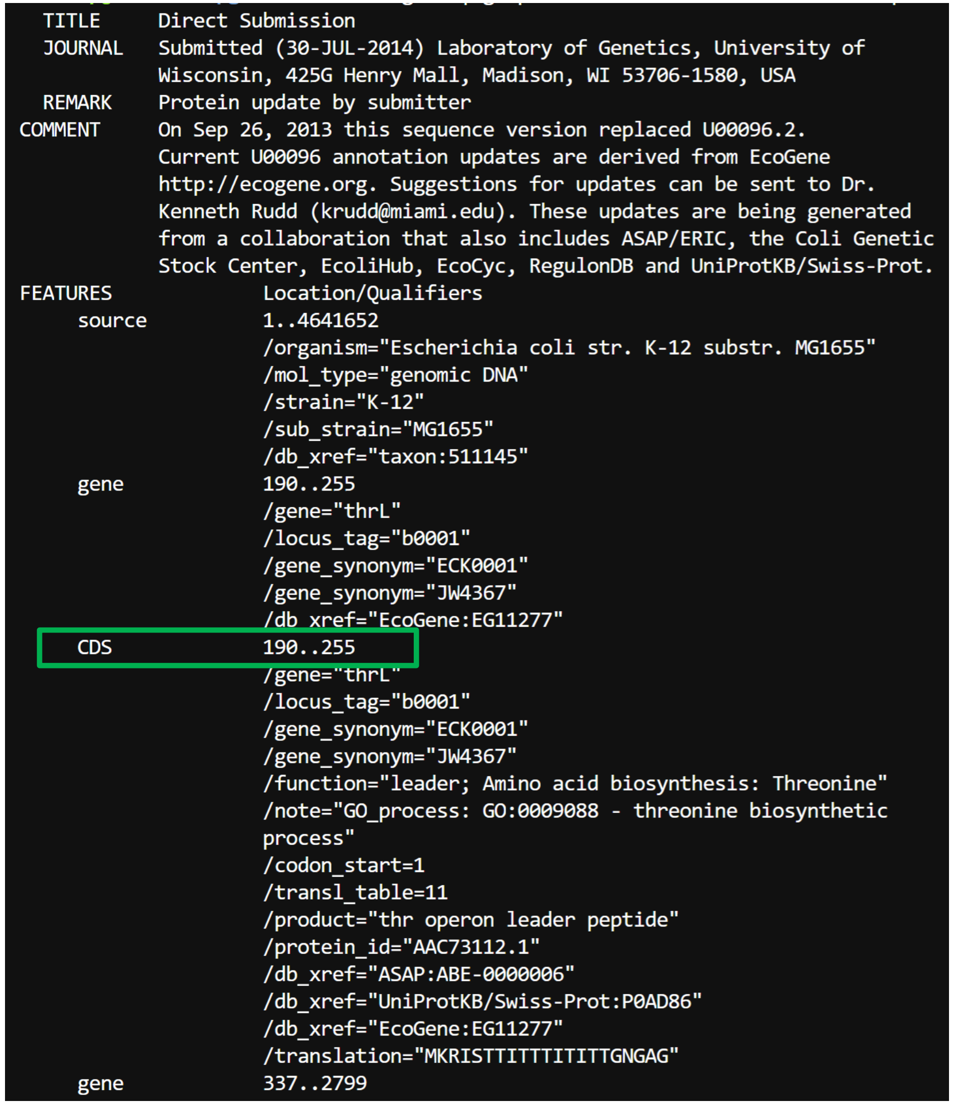
```
  TITLE     Direct Submission
  JOURNAL   Submitted (30-JUL-2014) Laboratory of Genetics, University of
            Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
  REMARK    Protein update by submitter
COMMENT     On Sep 26, 2013 this sequence version replaced U00096.2.
            Current 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))
```

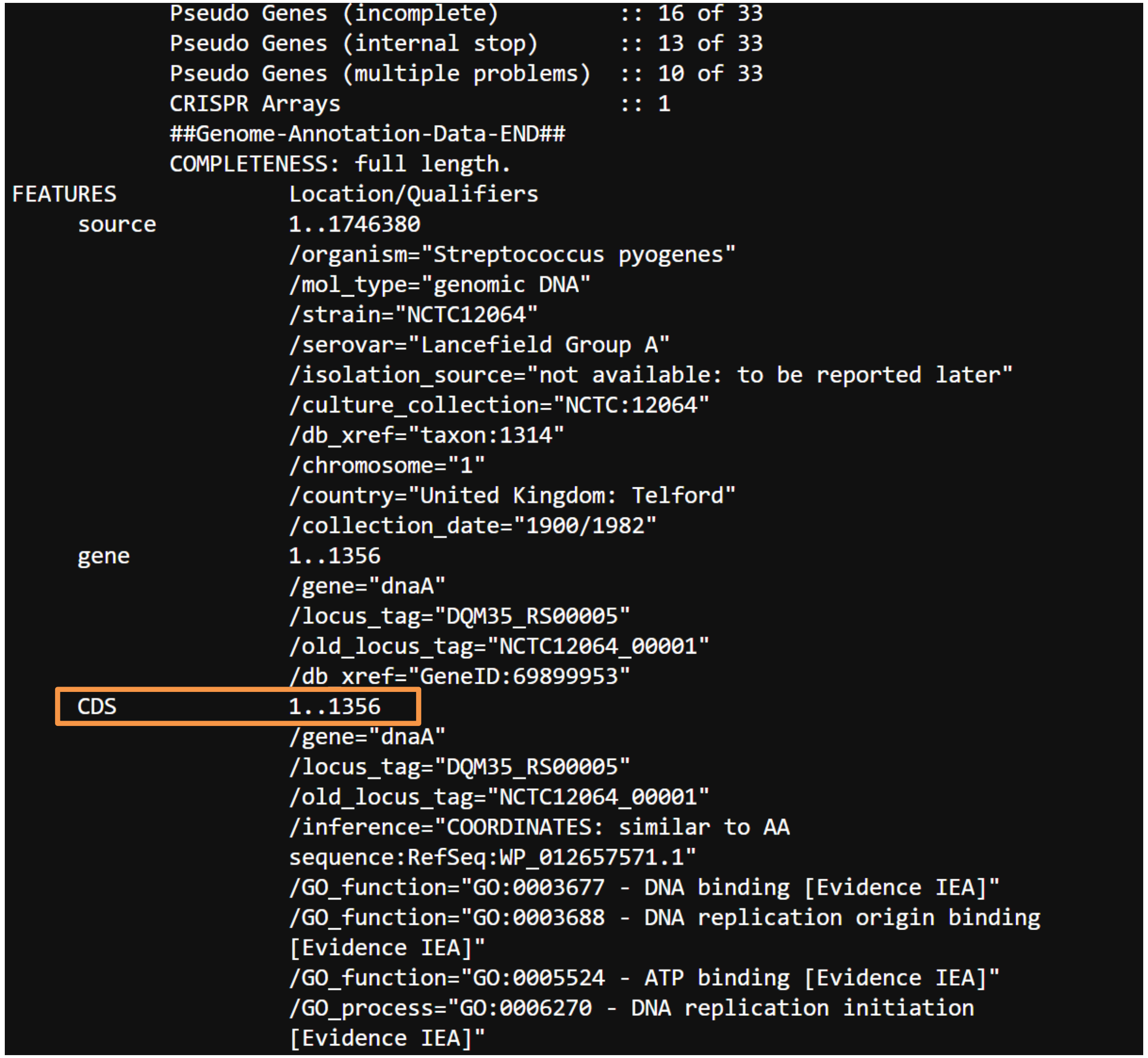
Homework 3 Wrap-up

test1.gbff



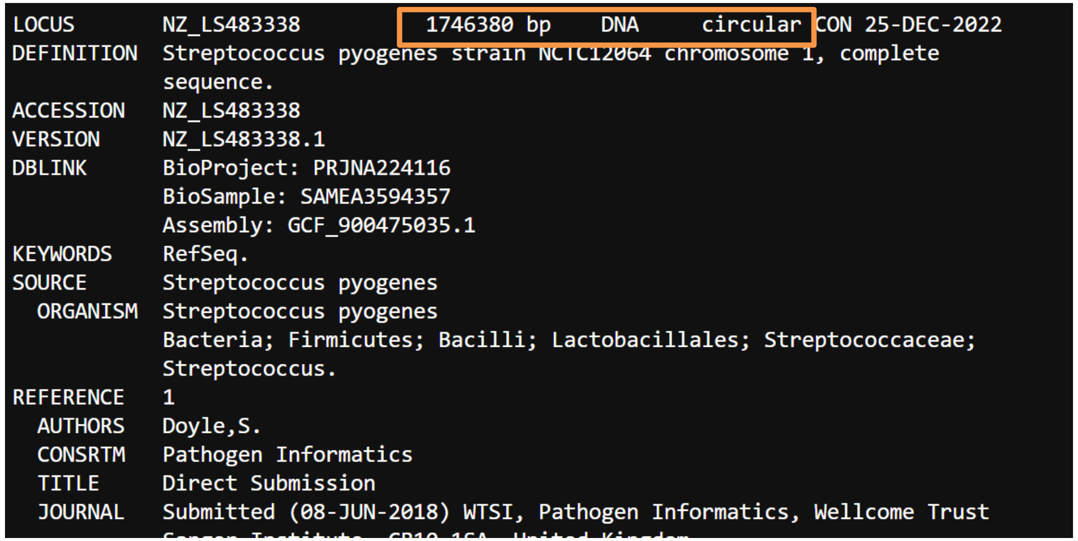
```
  TITLE     Direct Submission
  JOURNAL   Submitted (30-JUL-2014) Laboratory of Genetics, University of
            Wisconsin, 425G Henry Mall, Madison, WI 53706-1580, USA
  REMARK    Protein update by submitter
COMMENT     On Sep 26, 2013 this sequence version replaced U00096.2.
            Current 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
```

s_pyogenes.gbff



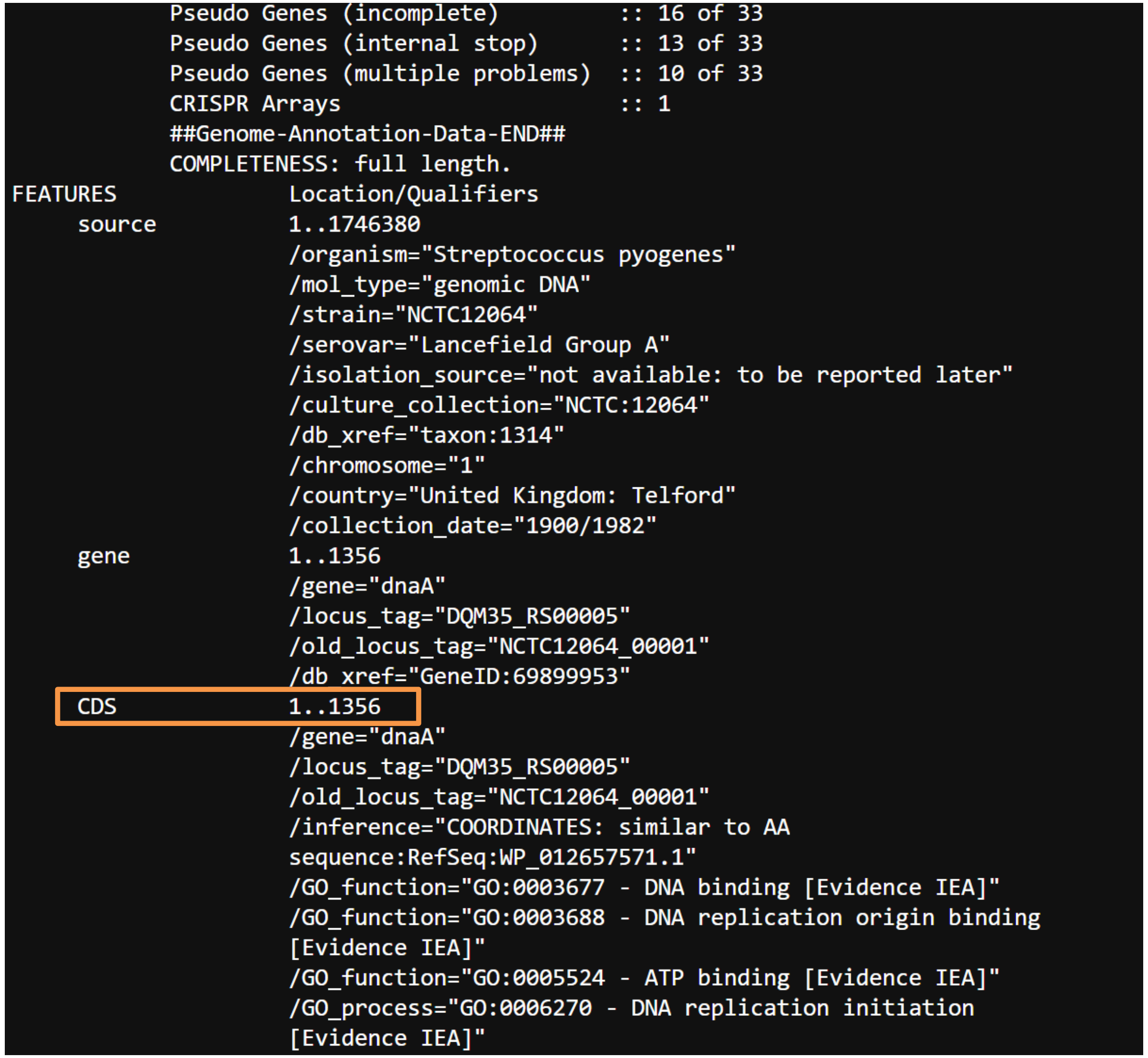
```
            Pseudo Genes (incomplete)          :: 16 of 33
            Pseudo Genes (internal stop)       :: 13 of 33
            Pseudo Genes (multiple problems)   :: 10 of 33
            CRISPR Arrays                      :: 1
            ##Genome-Annotation-Data-END##
            COMPLETENESS: full length.
FEATURES             Location/Qualifiers
     source          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]"
```

Homework 3 Wrap-up



```
LOCUS       NZ_LS483338          1746380 bp    DNA     circular CON 25-DEC-2022
DEFINITION  Streptococcus pyogenes strain NCTC12064 chromosome 1, complete
            sequence.
ACCESSION   NZ_LS483338
VERSION     NZ_LS483338.1
DBLINK      BioProject: PRJNA224116
            BioSample: SAMEA3594357
            Assembly: GCF_900475035.1
KEYWORDS    RefSeq.
SOURCE      Streptococcus pyogenes
  ORGANISM  Streptococcus pyogenes
            Bacteria; Firmicutes; Bacilli; Lactobacillales; Streptococcaceae;
            Streptococcus.
REFERENCE   1
  AUTHORS   Doyle,S.
  CONSRTM   Pathogen Informatics
  TITLE     Direct Submission
  JOURNAL   Submitted (08-JUN-2018) WTSI, Pathogen Informatics, Wellcome Trust
            Sanger Institute, CB10 1SA, United Kingdom
```

s_pyogenes.gbff



```
             Pseudo Genes (incomplete)          :: 16 of 33
             Pseudo Genes (internal stop)       :: 13 of 33
             Pseudo Genes (multiple problems)   :: 10 of 33
             CRISPR Arrays                      :: 1
             ##Genome-Annotation-Data-END##
             COMPLETENESS: full length.
FEATURES             Location/Qualifiers
     source          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]"
```

For a (complete) circular genome, technically one should append the sequence at the end of the genome to the front of this sequence. But it is fine just to ignore such cases (i.e. exclude the CDS).

Phil

Outline

- Homework 3 wrap-up

- Homework 4 overview & questions

Homework 4 Overview

Part one: max weight path through a WDAG

- Write a program to find the max weight path
- Convert a WDAG diagram to a text representation (by hand)
- Determine the max weight path using your program
  - Both: unconstrained, constrained start/stop vertices

Part two: GC-rich genomic sub-sequence

- Write a program to represent a genome as a WDAG, export .txt
  - GC vs. AT scoring scheme
- Determine the max weight path using your program
  - GC-rich sub-sequence, lookup feature in .gbff file

Homework 4 Overview

Part one: max weight path through a WDAG

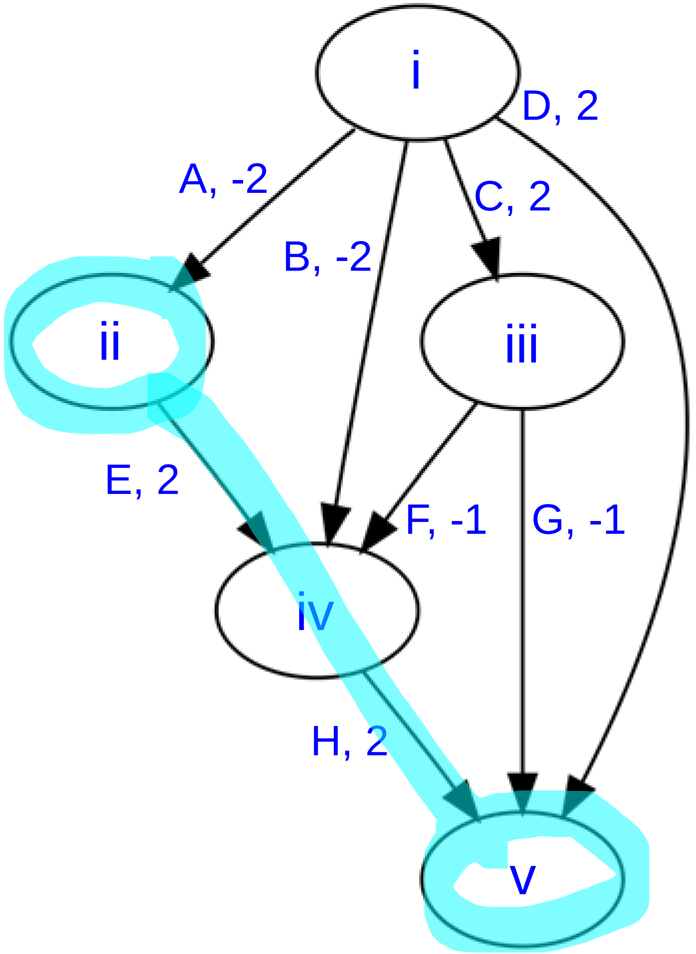
- Write a program to find the max weight path
- Convert a WDAG diagram to a text representation (by hand)
- Determine the max weight path using your program
    - Both: unconstrained, constrained start/stop vertices

Part two: GC-rich genomic sub-sequence

- Write a program to represent a genome as a WDAG, export .txt
    - GC vs. AT scoring scheme
- Determine the max weight path using your program
    - GC-rich sub-sequence, lookup feature in .gbff file

Homework 4 Overview

Example WDAG



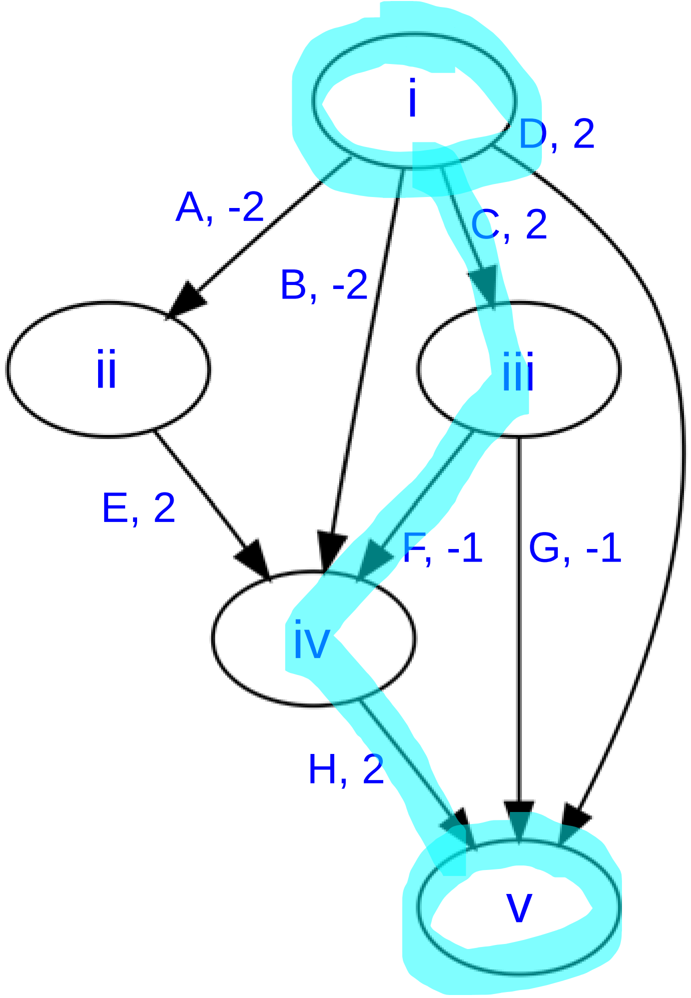
wdag_unconstrained.txt

```
V i
V ii
V iii
V iv
V v
E A i ii -2
E B i iv -2
E C i iii 2
E D i v 2
E E ii iv 2
E F iii iv -1
E G iii v -1
E H iv v 2
```

```
Score: 4.0
```

Homework 4 Overview

Example WDAG



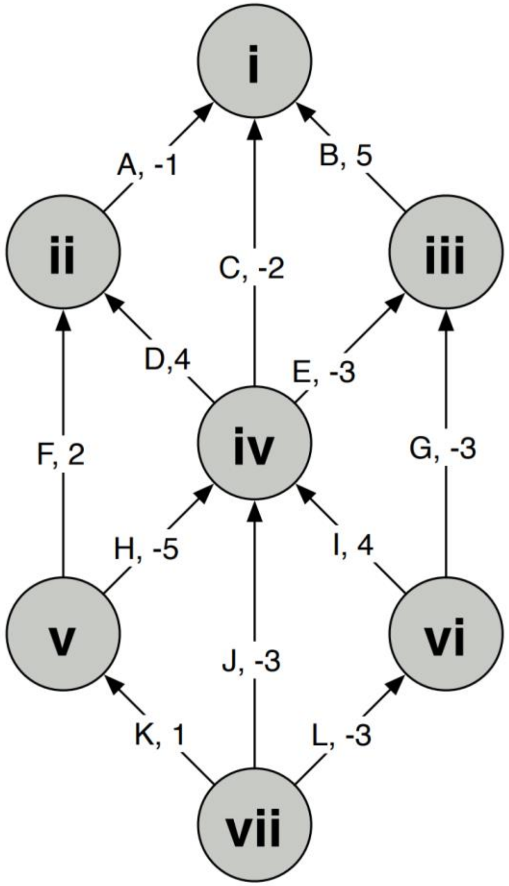
wdag_constrained.txt

```
V i START
V ii
V iii
V iv
V v END
E A i ii -2
E B i iv -2
E C i iii 2
E D i v 2
E E ii iv 2
E F iii iv -1
E G iii v -1
E H iv v 2
```

```
Score: 3.0
```

Homework 4 Overview

Test WDAG

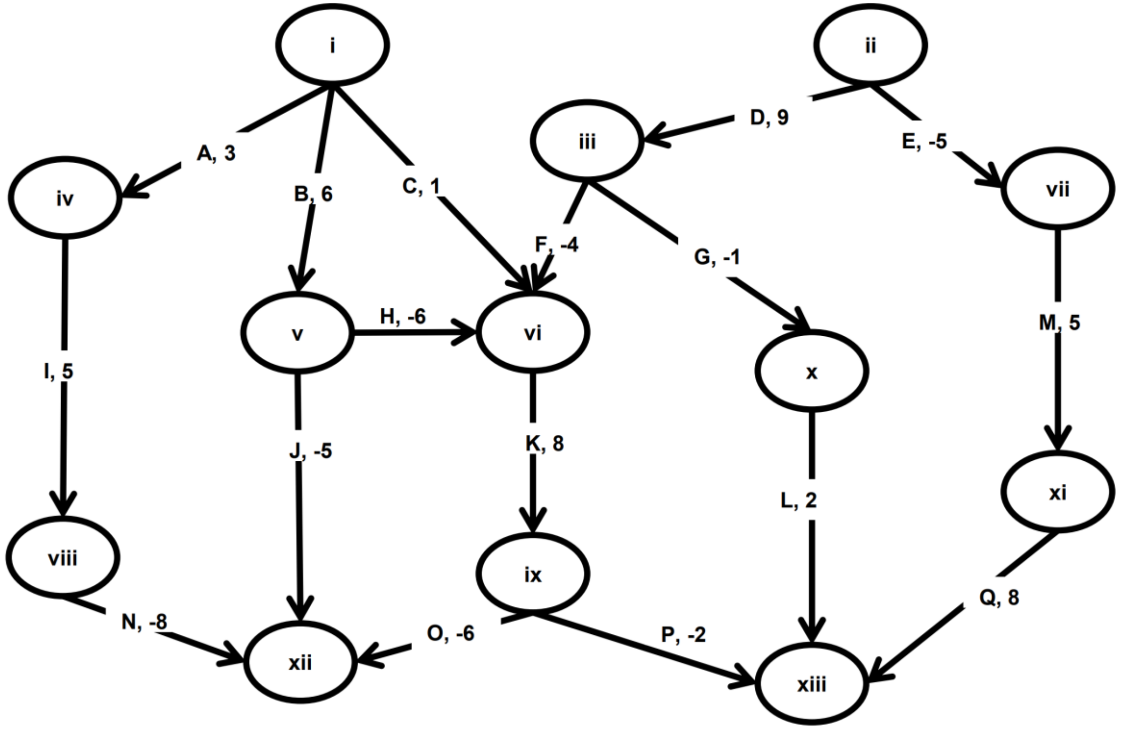


```
Assignment: GS 540 HW4
Name: Conor Camplisson
Email: concamp@uw.edu
Language: C++/Python
Runtime: 0m17.545s

Part 1
Score: 8
Begin: vi
End: ii
Path: ID

Part 2
Score: 4
Begin: vii
End: i
Path: LIDA
```

Homework WDAG



Homework 4 Overview

Part one: max weight path through a WDAG

- Write a program to find the max weight path
- Convert a WDAG diagram to a text representation (by hand)
- Determine the max weight path using your program
  - Both: unconstrained, constrained start/stop vertices

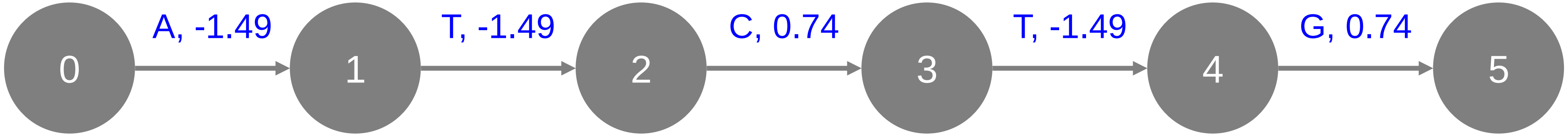
Part two: GC-rich genomic sub-sequence

- Write a program to represent a genome as a WDAG, export .txt
  - GC vs. AT scoring scheme
- Determine the max weight path using your program
  - GC-rich sub-sequence, lookup feature in .gbff file

Homework 4 Overview

Example sequence: 5’-ATCTG-3’

WDAG representation:



Homework 4 Overview

genome.fa

5’-ATCTG-3’

soring_scheme.txt

```
A -1.49
C 0.74
G 0.74
T -1.49
N 0
```

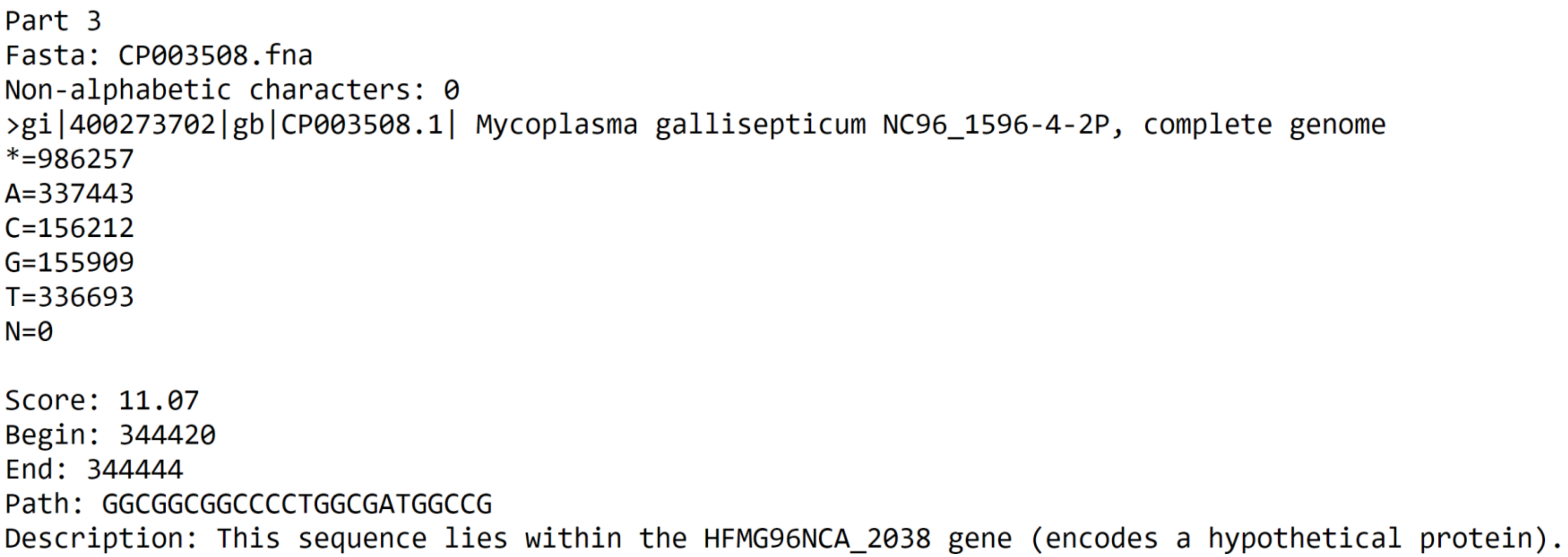
Program 2 →

```
V 0
V 1
V 2
V 3
V 4
V 5
E A 0 1 -1.49
E T 1 2 -1.49
E C 2 3 0.74
E T 3 4 -1.49
E G 4 5 0.74
```

Program 1 → ✓

Homework 4 Overview

Test Genome



```
Part 3
Fasta: CP003508.fna
Non-alphabetic characters: 0
>gi|400273702|gb|CP003508.1| Mycoplasma gallisepticum NC96_1596-4-2P, complete genome
*=986257
A=337443
C=156212
G=155909
T=336693
N=0

Score: 11.07
Begin: 344420
End: 344444
Path: GGCGGCGGCCCCTGGCGATGGCCG
Description: This sequence lies within the HFMG96NCA_2038 gene (encodes a hypothetical protein).
```

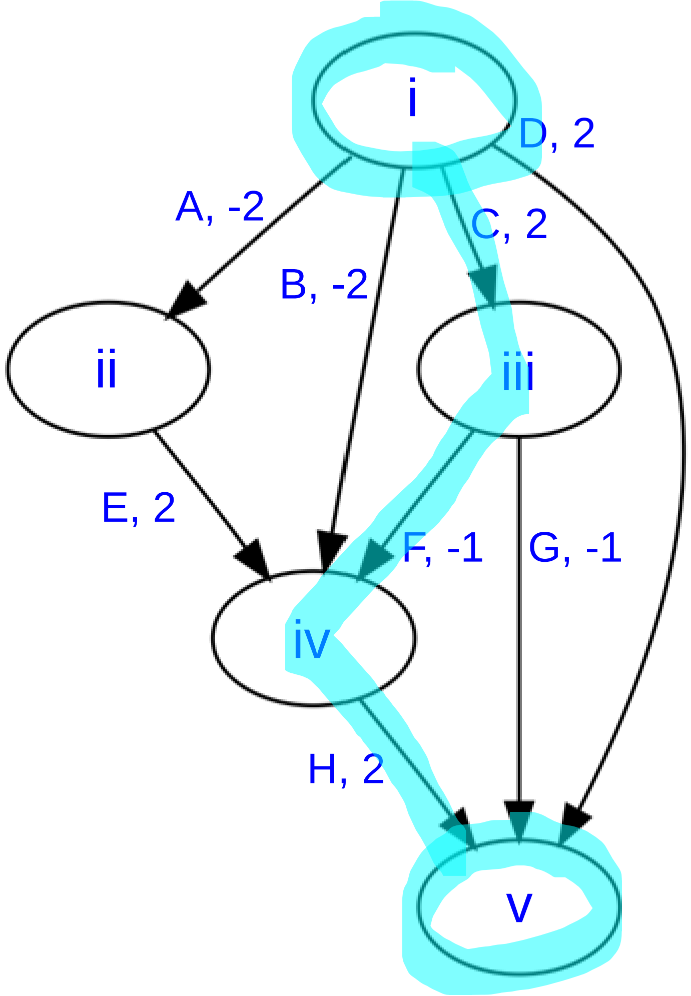
Homework Genome

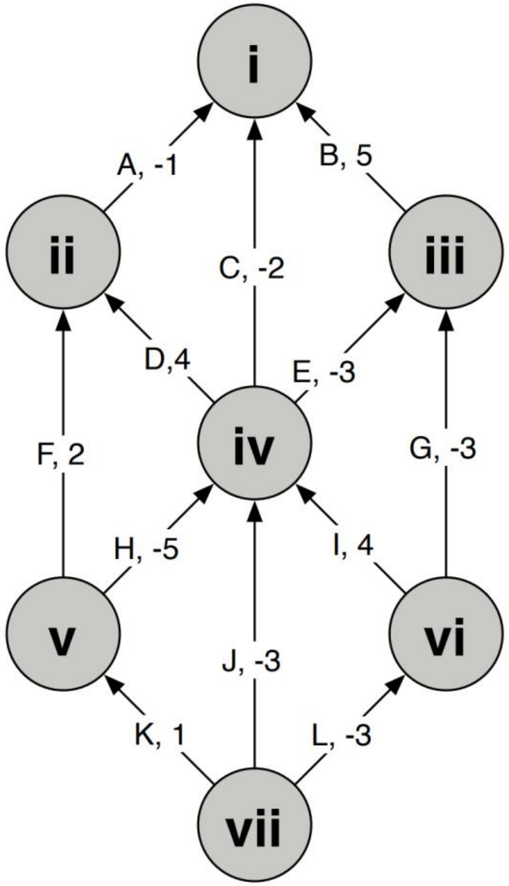
S. pyogenes

Home > Genome Editing > Products > Cas9 Nuclease, *S. pyogenes*

Cas9 Nuclease, *S. pyogenes*

Homework 4 Questions ?





```
Assignment: GS 540 HW4
Name: Conor Camplisson
Email: concamp@uw.edu
Language: C++/Python
Runtime: 0m17.545s

Part 1
Score: 8
Begin: vi
End: ii
Path: ID

Part 2
Score: 4
Begin: vii
End: i
Path: LIDA
```

Reminders

- Homework 4 due this Sunday Feb. 5, 11:59 pm
  - name in the file: `camplisson_hw3.txt.gz`

- Homework 5 will be posted tomorrow