Genome 540 Discussion

Conor Camplisson

January 31st, 2023
Outline

• Homework 3 wrap-up

• Homework 4 overview & questions
Outline

• Homework 3 wrap-up

• Homework 4 overview & questions
Homework 3 Wrap-up

**Genome**

5' → (+) strand gene → 3'

3' → 5'

**Transcription, polyadenylation**

5' → pre-mRNA → AAAAA-3'

**Splicing, introns removed**

5' → Mature transcript → AAAAA-3'

**Site model**

**HW3 Score Histograms**

- **Annotated CDS Sites**
  - test1.gbff

- **Genomic Background Sites**
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**

<table>
<thead>
<tr>
<th>TITLE</th>
<th>Direct Submission</th>
</tr>
</thead>
<tbody>
<tr>
<td>JOURNAL</td>
<td>Submitted (30-JUL-2014) Laboratory of Genetics, University of Wisconsin, 425G Henry Mall, Madison, WI 53796-1588, USA</td>
</tr>
<tr>
<td>REMARK</td>
<td>Protein update by submitter</td>
</tr>
<tr>
<td>COMMENT</td>
<td>On Sep 26, 2013 this sequence version replaced U00096.2. Current U00096 annotation updates are derived from EcoGene <a href="http://ecogene.org">http://ecogene.org</a>. Suggestions for updates can be sent to Dr. Kenneth Rudd (<a href="mailto:krudd@miami.edu">krudd@miami.edu</a>). These updates are being generated from a collaboration that also includes ASAP/ERIC, the Coli Genetic Stock Center, EcoHub, EcoCyc, RegulonDB and UniProtKB/Swiss-Prot.</td>
</tr>
</tbody>
</table>

**s_pyogenes.gbff**

<table>
<thead>
<tr>
<th>FEATURES</th>
<th>source</th>
</tr>
</thead>
<tbody>
<tr>
<td>location/qualifiers</td>
<td>1..1746380</td>
</tr>
<tr>
<td>/organism=&quot;Streptococcus pyogenes&quot;</td>
<td></td>
</tr>
<tr>
<td>/mol_type=&quot;genomic DNA&quot;</td>
<td></td>
</tr>
<tr>
<td>/strain=&quot;NCTC12864&quot;</td>
<td></td>
</tr>
<tr>
<td>/sequence=&quot;Lancefield Group A&quot;</td>
<td></td>
</tr>
<tr>
<td>/isolation_source=&quot;not available: to be reported later&quot;</td>
<td></td>
</tr>
<tr>
<td>/culture_collection=&quot;NCTC:12864&quot;</td>
<td></td>
</tr>
<tr>
<td>/db_xref=&quot;taxon:1314&quot;</td>
<td></td>
</tr>
<tr>
<td>/chromosome=&quot;1&quot;</td>
<td></td>
</tr>
<tr>
<td>/country=&quot;United Kingdom: Telford&quot;</td>
<td></td>
</tr>
<tr>
<td>/collection_date=&quot;1900/1982&quot;</td>
<td></td>
</tr>
</tbody>
</table>

**features**

**gene**

| 1..1356 |
| /gene="dnaA" |
| /locus_tag="DQ35_RS00005" |
| /old_locus_tag="NCTC12864_00001" |
| /db_xref="GeneID:698999953" |

**CDS**

| 1..1356 |
| /gene="dnaA" |
| /locus_tag="DQ35_RS00005" |
| /old_locus_tag="NCTC12864_00001" |
| /db_xref="GeneID:698999953" |
| /inference="COORDINATES: similar to AA sequence:RefSeq:WP_012657571.1" |
| /GO_function="60:0003677 - DNA binding [Evidence IEA]" |
| /GO_function="60:0003688 - DNA replication origin binding [Evidence IEA]" |
| /GO_function="60:0005524 - ATP binding [Evidence IEA]" |
| /GO_protein="60:000270 - 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
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

```
Example WDAG

A, -2
B, -2
C, 2
D, 2
E, 2
F, -1
G, -1
H, 2

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
```
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 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

5’-ATCTG-3’

genome.fa

soring_scheme.txt

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

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 2

Program 1
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: GCGCGCGCCCCTGCGATGGCG
Description: This sequence lies within the HFMG96NCA_2038 gene (encodes a hypothetical protein).
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