BIOSC 1540 - Computational Biology

Genomics Quiz - Key

Feb 15, 2024

50 points

Please read the following instructions carefully before beginning your assessment.

- **Time limit:** You have 75 minutes to complete and turn in this assessment.
- **Open note:** You may use notes for this assessment with the following restrictions.
  - You must hand-write notes on paper or digitally using a tablet and stylus and print them.
  - You may not have more than two 8.5 × 11 in. sheets of paper.
  - All notes must be your work; sharing notes with peers is strictly prohibited.
  - Your name must be clearly written on the first page.
- **No digital devices:** No digital devices are allowed during the assessment.
- **Submission requirements:** You must submit all your notes with the completed assessment.
- **No calculators:** Calculators are not permitted.

I agree to the above instructions. I affirm that I will not give or receive any unauthorized help on this assessment and that all work will be my own. Your assessment will only be graded if you write you name, sign, and put your student ID below.

Name ______________________ Signature ______________________

Student ID ______________

Problem 1

What is the main limitation of Sanger sequencing? (1 point)

A. It can only sequence DNA molecules.
B. It has a high cost and low throughput.
C. It cannot be used to detect genetic mutations.
D. It cannot accurately sequence repetitive regions of DNA.

Problem 2

How can Sanger sequencing be used to help with next-generation sequencing (NGS) technologies? (1 point)

A. To confirm sequencing results and fill gaps.
B. Preprocess samples before NGS platforms sequence them.
C. Calibrate NGS machines and ensure their accuracy.
D. Provide continuous sequences across repetitive regions.

Problem 3

Which principle does Illumina sequencing rely on? (1 point)

A. Electrophoretic separation of DNA fragments.
B. Fluorescence resonance energy transfer for nucleotide detection.
C. Sequencing by synthesis using reversible terminator nucleotides.
D. Microarray hybridization to detect specific DNA sequences.

Problem 4

How does the read length of Illumina sequencing generally compare to that of Sanger sequencing? (1 point)

This question was poorly worded and I accepted any reasonable answer.

A. Illumina has longer reads than Sanger sequencing.
B. Sanger produces longer reads than Illumina sequencing.
C. Both technologies produce reads of similar lengths.
D. Read length does not apply to Illumina sequencing.

Problem 5

What is a significant challenge particular to de novo genome assembly? (1 point)

A. Finding a reference genome to align to.
B. Managing computational resources.
C. Identifying single nucleotide polymorphisms.
D. Handling repetitive DNA sequences.

Problem 6

What does a directed edge represent in de Bruijn graphs? (1 point)

A. A mutation in the genomic sequence.
B. A direct repetition of a sequence.
C. The overlap between k-mer sequences.
D. The physical distance between two genes.

Problem 7

What computational complexity is characteristic of de novo assembly? (1 point)

A. The total number of reads.
B. The length of the reads.
C. The similarity of k-mers within the reads.
D. The size of the reference genome.

Problem 8

What is the primary benefit of using both short and long reads in de novo genome assembly? (1 point)

A. It simplifies the computational algorithms required for assembly.
B. Long reads can span repetitive regions, while short reads improve coverage.
C. It reduces the overall cost of genome sequencing.
D. Hybrid approaches are faster because they require less data.

Problem 9

Which algorithm is commonly used for local pairwise sequence alignment? (1 point)

There were (accidently) two correct answers.

A. Smith-Waterman algorithm
B. Needleman-Wunsch algorithm
C. BLAST algorithm
D. Greedy algorithm

Problem 10

What is the main difference between global and local sequence alignment? (1 point)

There is a subtle difference between the two, but I will accept A or C.

A. Global alignment matches sequences in full; local alignment focuses on best-matching parts.
B. Global alignment uses penalties for gaps across entire sequences; local alignment finds high-scoring segments without gap penalties.
C. Local alignment identifies conserved regions; global for full sequence comparison.
D. Global alignment for evolutionary relationships; local for functional domains or motifs.

Problem 11

What is the significance of the numerical value in the bottom-right of the Needleman-Wunsch alignment matrix? (3 points)

The numerical value in the bottom-right corner of the Needleman-Wunsch alignment matrix represents the optimal score of the global alignment between two sequences.

Problem 12

What are the inherent limitations of greedy algorithms for genome assembly? In particular, how might these limitations affect the assembly outcome when dealing with complex eukaryotic genomes? (5 points)

Greedy algorithms for sequence assembly can be efficient for simple genomes but struggle with complex ones due to issues with repetitive sequences, limited global perspective, and handling of genetic variation. These limitations can lead to misassemblies, especially in genomes with high complexity, such as those with repetitive regions or structural variations.

Problem 13

Using the greedy algorithm, assemble the following error-free reads into a contig. (8 points)

- `5'- ATCGA -3'`,
- `5'- CGATT -3'`,
- `5'- GATTC -3'`,
- `5'- ATTCA -3'`.

Possible first merges:

```
ATCGA--    ATCGA---    ATCGA----    ----ATCGA    ----CGATT    GATTC-
--CGATT    ---GATTC    ----ATTCA    ATTCA----    GATTC----    -ATTCA
```

Check for possible second merges with best first merge:

```
GATTCA----    ---GATTCA    -GATTCA
-----ATCGA    ATCGA----    CGATT--
```

Check if we can merge our last read after best second merge:

```
CGATTCA----    --CGATTCA
------ATCGA    ATCGA
```

Our resulting contig: ATCGATTCA.

Problem 14

Fill out the matrix below using the Needleman-Wunsch algorithm with the following scoring system: (10 points)

- match: +1,
- mismatch: −1,
- gap: −1.

	-	A	T	G	C	A
-	0	-1	-2	-3	-4	-5
A	-1	↘ 1	→ 0	-1	-2	-3
G	-2	↓ 0	0	↘ 1	0	-1
T	-3	-1	↘ 1	0	↘ 0	-1
A	-4	-2	0	↘ 0	-1	↘ 1
C	-5	-3	-1	-1	↘ 1	↓ → 0

Perform a traceback on your filled-out matrix; what is the resulting alignment? (4 points)

Alignment 1:

```
ATGCA-
A-GTAC
```

or Allignment 2:

```
A-TGCA
AGTAC-
```

Problem 15

Given the sequence assembly of

```
AATGGCGTA--------
-----CGTAAAT-----
-------TAAATGCGAA
```

construct a complete de Bruijn graph of the contig for $k = 4$. (7 points)

```
AATGGCGTA———————
———— CGTAAAT————
————— TAAATGCGAA
```

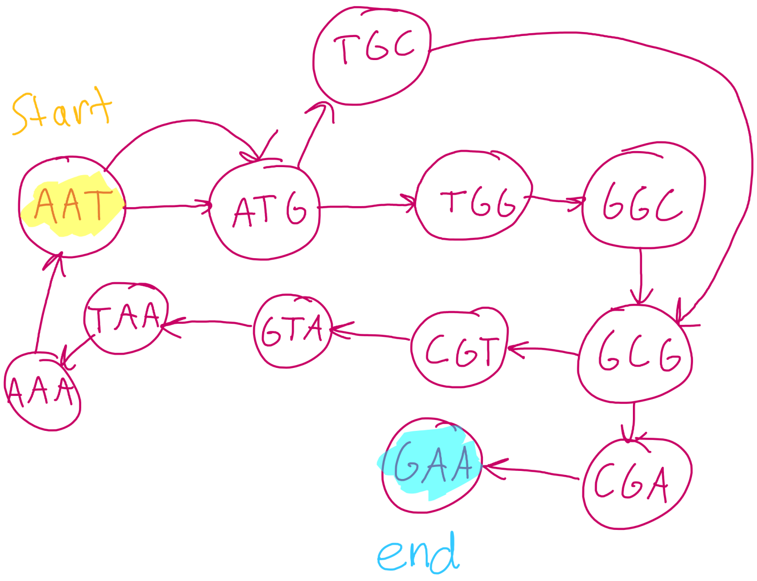
This assembly produces the following contig.

AATGGCGTAAATGCGAA

we split these into 4-mers

- AATG
- ATGG
- TGGC
- GGCG
- GCGT
- CGTA
- GTAA
- TAAA
- AAAT
- AATG
- ATGC
- TGCG
- GCGA
- CGAA

Now we build our graph where our nodes are K-1



Problem 16

Suppose you are given an ASCII representation of Phred scores for an Illumina sequencing read as: `gGA$` Explain the steps you would take to compute the percent error for each base call and write out, but do not solve, the equations (i.e., plug in the numbers). (3 points)

You subtract 33 from the ASCII value of each character to get its Phred score.

- For `g`: 103 - 33 = 70
- For `G`: 71 - 33 = 38
- For `A`: 65 - 33 = 32
- For `$`: 36 - 33 = 3

This formula gives the probability of an incorrect base call.

$$P = 10^{-\frac{Q}{10}}$$

We then need to multiply by 100 to get a percent error.

$$P = 100 \cdot 10^{-\frac{70}{10}}$$

$$P = 100 \cdot 10^{-\frac{38}{10}}$$

$$P = 100 \cdot 10^{-\frac{32}{10}}$$

$$P = 100 \cdot 10^{-\frac{3}{10}}$$

Appendices

A.1 Phred quality score

$$P = 10^{-\frac{Q}{10}}$$

A.2 ASCII table

```
Dec  Char                           Dec  Char     Dec  Char     Dec  Char
---------                           ---------     ---------     ----------
  0  NUL (null)                      32  SPACE     64  @         96  `
  1  SOH (start of heading)          33  !         65  A         97  a
  2  STX (start of text)             34  "         66  B         98  b
  3  ETX (end of text)               35  #         67  C         99  c
  4  EOT (end of transmission)       36  $         68  D        100  d
  5  ENQ (enquiry)                   37  %         69  E        101  e
  6  ACK (acknowledge)               38  &         70  F        102  f
  7  BEL (bell)                      39  '         71  G        103  g
  8  BS  (backspace)                 40  (         72  H        104  h
  9  TAB (horizontal tab)            41  )         73  I        105  i
 10  LF  (NL line feed, new line)    42  *         74  J        106  j
 11  VT  (vertical tab)              43  +         75  K        107  k
 12  FF  (NP form feed, new page)    44  ,         76  L        108  l
 13  CR  (carriage return)           45  -         77  M        109  m
 14  SO  (shift out)                 46  .         78  N        110  n
 15  SI  (shift in)                  47  /         79  O        111  o
 16  DLE (data link escape)          48  0         80  P        112  p
 17  DC1 (device control 1)          49  1         81  Q        113  q
 18  DC2 (device control 2)          50  2         82  R        114  r
 19  DC3 (device control 3)          51  3         83  S        115  s
 20  DC4 (device control 4)          52  4         84  T        116  t
 21  NAK (negative acknowledge)      53  5         85  U        117  u
 22  SYN (synchronous idle)          54  6         86  V        118  v
 23  ETB (end of trans. block)       55  7         87  W        119  w
 24  CAN (cancel)                    56  8         88  X        120  x
 25  EM  (end of medium)             57  9         89  Y        121  y
 26  SUB (substitute)                58  :         90  Z        122  z
 27  ESC (escape)                    59  ;         91  [        123  {
 28  FS  (file separator)            60  <         92  \        124  |
 29  GS  (group separator)           61  =         93  ]        125  }
 30  RS  (record separator)          62  >         94  ^        126  ~
 31  US  (unit separator)            63  ?         95  _        127  DEL
```