2110495 Advance Topics in Computer Engineering (Bioinformatics I) (140 คะแนน)
ชื่อ-นามสกุล ...นาย ดนุภัทร คำนวนสินธุ์ 5730196321 .....

1. (10 คะแนน) เซลล์ นิวเคลียส โครโมโซม จีโนม ดีเอ็นเอ ยีน มีความเกี่ยวข้องกันอย่างไร (6 คะแนน) เกี่ยวข้องกับการถอดรหัสจีโนมอย่างไร (2 คะแนน) และเกี่ยวข้องกับเซ็นทรัลดอกม่าอย่างไร (2 คะแนน)

   ตอบ: เซลล์คือหน่วยย่อยๆ พื้นฐานของร่างกายสิ่งมีชีวิต (สิ่งมีชีวิตประกอบขึ้นมาจาก cell) นิวเคลียสเป็นส่วนประกอบสำคัญภายในเซลล์ทำหน้าที่เก็บข้อมูลสำคัญที่ใช้ในการระบุลักษณะต่างๆ หรือ การทำงานต่างๆ ภายในเซลล์ ซึ่งภายในนิวเคลียส จะมีโครมาโซม ซึ่งจะบรรจุไปด้วย ดีเอ็นเอ โดยใน ดีเอ็นเอ ก็จะมี ยีน และ จีโนม อยู่ข้างใน โดยจีโนมจะทำหน้าที่เก็บข้อมูลวิธีการสร้างสิ่งมีชีวิต เปรียบเสมือนพิมพ์เขียวของสิ่งมีชีวิต การถอดรหัสจีโนม คือการพยายามอ่าน จีโนมของสิ่งมีชีวิตเพื่อหาคำตอบการเกิด หรือ การทำงานต่างๆ ของสิ่งมีชีวิต และ ยังสามารถช่วยในการทำนาย หรือ ระบุแนวโน้มการเกิดโรคบางอย่าง รวมถึง ลักษณะต่างๆ ของสิ่งมีชีวิตได้อีกด้วย สุดท้าย เซ็นทรัลดอกม่า คือการถ่ายทอดลักษณะทางพันธุกรรม โดยถ่ายทอดมาในรูปของโปรตีน จากกระบวนการแบ่งตัวของดีเอ็นเอ

2. (10 คะแนน) จงอธิบายว่า การประกอบร่างจีโนม (genome assembly) แบบ *de novo* และแบบ reference based แตกต่างกันอย่างไรในมิติของลักษณะการทำงาน (4 คะแนน) และอภิปรายความถูกต้องของผลการทำงาน (3 คะแนน) เวลาที่ใช้ในการประกอบร่างจีโนม (3 คะแนน) ของทั้งสองแนวทาง

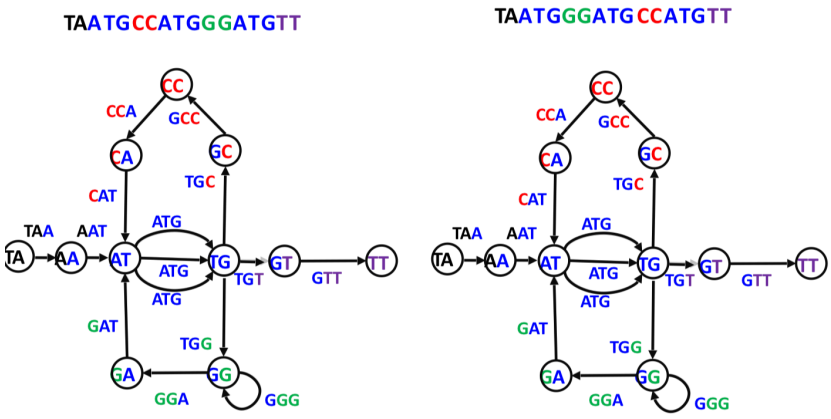
   ตอบ: de novo sequencing คือการพยายามประกอบร่างจีโนมทั้งเส้น จากข้อมูลจีโนมทีละส่วนย่อยๆ ที่อ่านมาได้ โดยไม่มีพื้นฐานความรู้เกี่ยวกับลักษณะของเส้นจีโนมที่มีผู้อื่น เคยศึกษาเอาไว้ มาใช้ในการประกอบ จะอาศัยเพียง อัลกอริทึม ในการประกอบเท่านั้น ซึ่งในแบบ reference-based sequencing จะมีการนำข้อมูลลักษณะของเส้นจีโนมที่ได้มีการศึกษาไว้ มาใช้ในการประกอบร่วมด้วย ความถูกต้องของผลการทำงานนั้น แบบ reference-based จะทำได้ดีกว่าเนื่องจากมีข้อมูลในการประกอบที่มากกว่าแบบ de novo ต่อเมื่อจีโนมดังกล่าว มีความคล้ายคลึงไม่แตกต่างไปกับจีโนมที่นำมาใช้เป็น reference และในทางกลับกัน de novo จะทำได้ดีกว่า และเวลาที่ใช้ในการประกอบร่างนั้น แบบ refernce-based มีแนวโน้มที่จะทำได้เร็วกว่าแบบ de novo ตามขนาดของข้อมูลที่เพิ่มขึ้น

3. (10 คะแนน) ถ้ากำหนดว่าดีเอ็นเอเส้นหนึ่งๆใน **reads.fastq** ที่อ่านได้จากเครื่องถอดรหัสจีโนมจะถือว่าผ่านเกณฑ์ถ้ามีจำนวนคลีโอไทด์อย่างน้อย 90% ที่มี Phred quality score >= 30 จงเขียนโคดที่ใช้ในการคัดกรองรีดที่ผ่านเกณฑ์และไม่ผ่านเกณฑ์ออกจากกัน

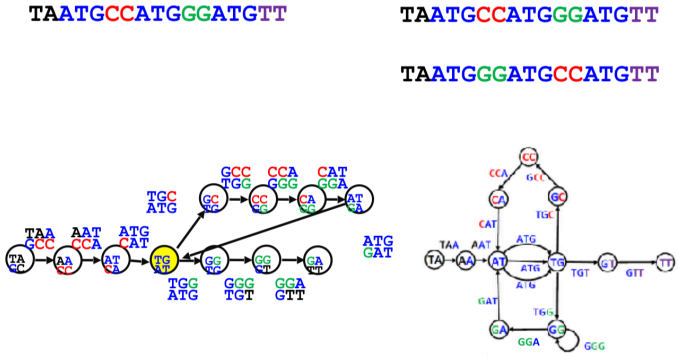
```python
# Open file
inf = open("example.fastq","r")
for line in inf:
    if line[0] not in ["+","#","@"]:
        # Read DNA seq
        DNA = line[:-1]
        count_n = 0
        # Counting not 'N' (N mean can't read)
        for each_position in DNA:
            if each_position != "N":
                count_n += 1
        if count_n/len(DNA) >= 0.90:
            # Pass
            print("PASS")
        else:
            # Not Pass
            print("NOTPASS")
```

4. (10 คะแนน) จงอธิบายพร้อมยกตัวอย่างว่า pair-end reads ช่วยลดความคลุมเครือในการประกอบร่างจีโนม (genome assembly) ได้อย่างไร

ตอบ: ช่วยลดความคลุมเครือในการประกอบร่างจีโนมจากอัลกอริทึมที่ใช้สร้าง De Bruijn Graph เนื่องจากจีโนมย่อยๆ ที่อ่านได้ อาจจะสามารถสร้าง De Bruijn Graph ได้หลากหลายรูปแบบ



เช่นตัวอย่างข้างต้น Genome เล็กๆ TAA AAT CAT CCA .... ที่อ่านมาได้อาจถูกประกอบกันได้หลายแบบหากอ่านมาเพียงชุดเดียวด้านเดียว แต่หากอ่านสองด้านมาพร้อมกัน จะทำให้สามารถสร้างกราฟได้แบบเดียวเท่านั้นเนื่องจากเมื่อเจอสิ่งที่ต้องตัดสินใจจะสามารถใช้ข้อมูลชุดที่สองมาชี้ขาดได้ ทำให้ลดความคลุมเครือในการประกอบได้



5. (10 คะแนน) จงออกแบบและอธิบายอัลกอริทึมที่ใช้ในการเปรียบเทียบความแตกต่างระหว่างจีโนมของ trio (พ่อ แม่ ลูก) โดยใช้แนวทางของการทำ *de novo* assembly หมายเหตุ อาจแสดงในรูปแบบของ pseudocode หรือแผนภาพ

ตอบ: เมื่อได้รับจีโนมของพ่อ แม่ ลูก ให้ทำการทำ de novo assembly จากนั้นทำการเปรียบเทียบความแปรผัน ซึ่ง ความแปรผันมีได้หลากหลายรูปแบบ เช่น การแปรผันในลำดับเบสเดียวที่บริเวณ หนึ่งๆ ที่เรียกว่าสนิปส์ (SNP) หรือการแปรผันเชิงโครงสร้าง (structural variation) เช่นมีชุดของรีพีทที่แตกต่างกันระหว่าง จีโนม เป็นต้น

6. (10 คะแนน) วิธีการ map หรือ align ข้อมูลดีเอ็นเอสายสั้นจำนวนมาก ไปยังจีโนมอ้างอิง วิธีการใดที่มีการใช้งานอย่างแพร่หลาย (2 คะแนน) จงยกตัวอย่างโปรแกรมที่ใช้วิธีการนี้มา 2 โปรแกรม (2 คะแนน) ข้อจำกัดของวิธีการนี้ (2 คะแนน) รูปแบบไฟล์ข้อมูลเข้า (2 คะแนน) รูปแบบไฟล์ที่เก็บผลลัพธ์มีอะไรบ้าง (2 คะแนน)

   ตอบ: Burrows-Wheeler-Tranform ใช้ในการลดขนาดของข้อมูลทีต้องเก็บให้มีขนาดเล็กลง และ Suffix Array/Tree ใช้ในการค้นหาตำแหน่งที่เหมือนกัน
   ตัวอย่างของโปรแกรม : 1. Bowtie, Bowtie2 2. CUSHAW
   ข้อจำกัดของวิธีการนี้: หาก reference genom ไม่ใกล้เคียงกับตัวอย่าง genome ที่ต้องการ map จะทำให้ได้ผลลัพธ์ที่ไม่ดี
   รูปแบบไฟล์ข้อมูลเข้า: เป็นสายสตริง DNA ที่ต้องการ map และ ฐานข้อมูล reference genome เพื่อใช้ในการค้นหา อาจอยู่ในรูปแบบไฟล์ใดก็ได้ (FASTA, FASTQ, etc.)
   รูปแบบไฟล์ข้อมูลออก: ไฟล์ที่ระบุตำแหน่งที่ตรงกับ reference genome บนสายตริง DNA

7. (20 คะแนน) จากไฟล์ Position Frequency Matrix **whatis.txt** ที่ post ใน FB จงเขียนโคดเพื่อสแกนหา regulatory motifs ในไฟล์ **test.fasta** โดยให้เปลี่ยนตัวอักษรในบริเวณที่เป็นไบน์ดิงไซต์ให้เป็นตัวอักษรใหญ่ (15 คะแนน) นำไบน์ดิงไซต์เหล่านี้ไปสร้างเป็น sequence logo (5 คะแนน)

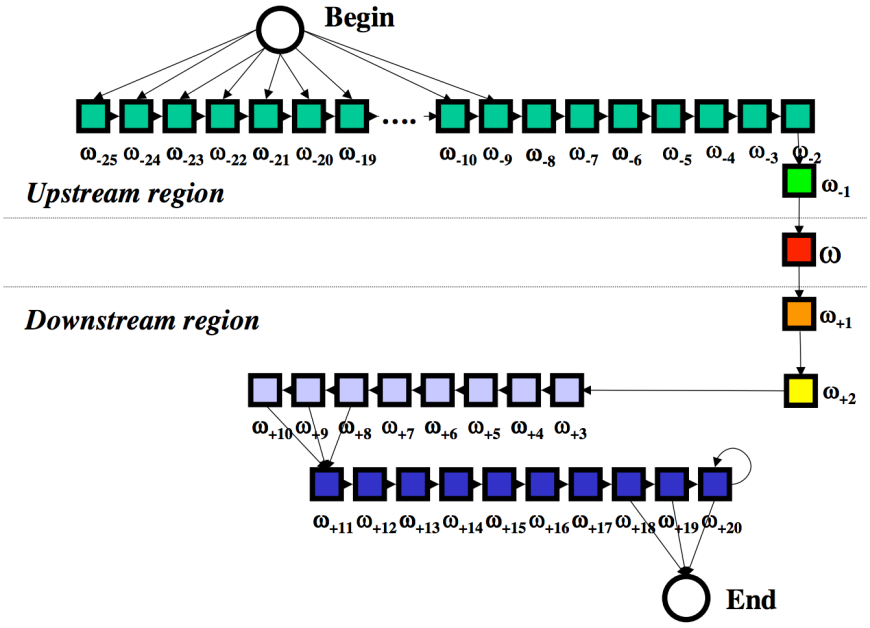
8. (10 คะแนน) จากไฟล์ **proteins.fasta** จงอธิบายว่าโปรตีนทั้ง 5 เส้นนี้มีความเกี่ยวข้องกันอย่างไร Hint: ให้ลองใช้เครื่องมือในฐานข้อมูลสาธารณะเช่น Pfam, BLAST, เพื่อหาคำตอบ

   ตอบ: โปรตีนทั้ง 5 เส้นล้วนมีความเกี่ยวข้องกับ แบคทีเรีย Staphylococcus โดยโปรตีนแรกจะเป็น Putative IgG-binding โปรตีนที่สองและห้าจะเป็น hypothetical protien และโปรตีนที่สามและสี่ จะเป็น immunoglobulin G-binding protein A

9. (20 คะแนน) ไฟล์ **terminals.fasta** เป็นสายของโปรตีนจำนวน 26 เส้นในบริเวณที่เป็น C-terminal (ปลายสายโปรตีนทางขวา) โดยมีการระบุว่ากรดอะมิโนที่เท่าไหร่เป็น ω site โดยนับจากกรดอะมิโนทางขวาสุด ตัวอย่างเช่น กรดอะมิโน S เป็น ω site

```
>5NTD_HUMAN | -25
MKVIYPAVEGRIKFSTGSHCHGSFSLIFLSLWAVIFVLYQ
```

โดยในส่วนของ C-terminal นี้มีบริเวณต่างๆที่เป็นไปได้ตามภาพ



คำถาม

a. เราจะประยุกต์ใช้ HMM ในการตอบคำถามว่าสายข้อมูลเข้าหนึ่งๆที่มีความยาว 40 กรดอะมิโนนั้น ตำแหน่งไหนน่าจะเป็น ω site มากที่สุด ได้อย่างไร (10 คะแนน)

- กำหนด hidden-state คือ เป็น ω site และ ไม่เป็น ω site ทั้งหมด 2 state
- กำหนด observing-state คือตัวอักษรต่างๆของกรดอะมิโน
- ใช้ Viterbi Algorithm ในการหา ตำแหน่งของ ω site ที่ fit กับ probabilities ต่างๆ เมื่อ given input string โดยกำหนดให้สนใจเฉพาะที่ ω site มีตำแหน่งเดียว และมีค่า probabilities ในการเกิด input string ดังกล่าวมีค่ามากที่สุด

b. ค่าเริ่มต้นของ transition probabilities และ emission probabilities สามารถคำนวณได้อย่างไร (10 คะแนน)

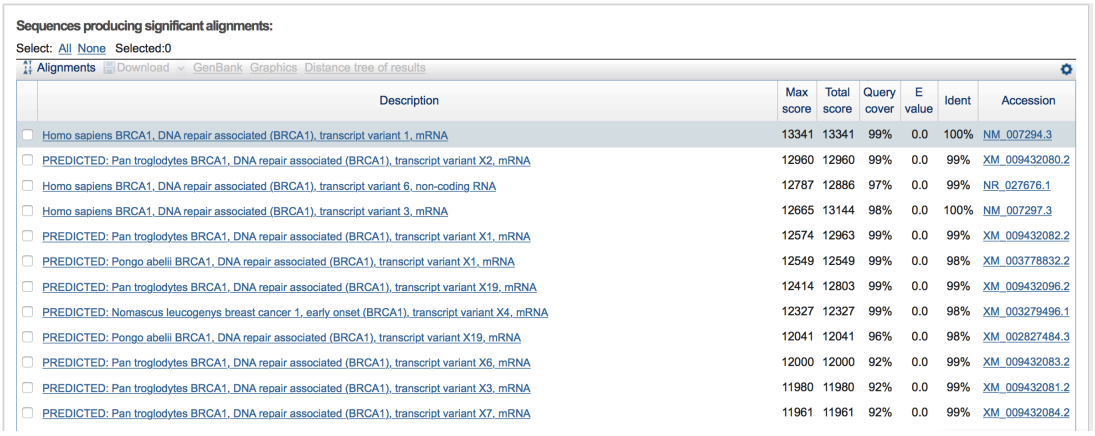
- สามารถคำนวณได้จากการใช้ Baum-Welch Learning ในการเรียนรู้แต่ละ input ที่เข้ามาเพื่อปรับค่า probabilities ในทั้งสองตารางให้มีความถูกต้องแม่นยำมากขึ้นได้ โดยจะมีความแม่นยำ (ใกล้เคียงค่าจริง) มากขึ้นเมื่อได้รับ input เข้ามาเป็นจำนวนมากขึ้น

10. (10 คะแนน) อัลกอริทึมที่ใช้ในการหา regulatory motifs, การทำ sequence alignment, และการสร้าง profile HMM มีความเหมือนและหรือแตกต่างกันและหรือเกี่ยวข้องกันอย่างไร (5 คะแนน) ถ้ามีชุดของ long-noncoding RNAs อย่างตัวอย่างในไฟล์ **lncRNAs.fasta** ที่ต้องการหาบริเวณที่มีความอนุรักษ์ระหว่างกันควรประยุกต์ใช้อัลกอริทึมในกลุ่มไหนบ้างในการช่วยหาคำตอบ (ถ้ามีข้อจำกัดใดๆ ให้บอกด้วย) (5 คะแนน)

11. (10 คะแนน) ถ้านักวิทยาศาสตร์พบว่านอกจาก 4 นิวคลีโอไทด์ A, T, C และ G แล้ว ยังมีนิวคลีโอไทด์ที่ค้นพบใหม่อีก 2 ตัวคือ X และ Y คำถามคือ codon usage table จะมีการเปลี่ยนไปอย่างไร และมีอาจมีผลกระทบกับเซ็นทรัลดอกมา (Central dogma) อย่างไรบ้าง

    ตอบ: จากเดิมที่มี 4 ชนิดคือ A, T, C, G จะได้ ทำให้เกิดตารางขนาด 64 ช่อง เมื่อมี X, Y เพิ่มเข้ามาจะทำให้ตารางมีขนาดใหญ่ขึ้น เนื่องจาก โอกาสเกิดแบบต่างๆ จะมีเพิ่มขึ้น เช่น XXY AXX เป็นต้น ต้องมีการคำนวณใหม่ทั้งหมด และอาจมีผลกระทบกับเซ็นทรัลดอกมา คือ จะทำให้เกิดความหลากหลายในการถ่ายทอดข้อมูลของ DNA เพิ่มมากขึ้นจากเดิมเนื่องจากมีตัวใหม่เพิ่มเข้ามาคือ X, Y

12. (10 คะแนน) สายข้อมูลดีเอ็นเอ sequence.fasta มีความคล้ายคลึงกับโปรตีนใดในฐานข้อมูลโปรตีนที่ NCBI ให้ capture หน้าจอผลมาแสดงเป็นคำตอบได้ และยกตัวอย่างโปรตีนโดเมนที่อยู่ในสายโปรตีนนี้มา 2 โดเมน



Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Homo sapiens BRCA1, DNA repair associated (BRCA1), transcript variant 1, mRNA	13341	13341	99%	0.0	100%	NM_007294.3
PREDICTED: Pan troglodytes BRCA1, DNA repair associated (BRCA1), transcript variant X2, mRNA	12960	12960	99%	0.0	99%	XM_009432080.2
Homo sapiens BRCA1, DNA repair associated (BRCA1), transcript variant 6, non-coding RNA	12787	12886	97%	0.0	99%	NR_027676.1
Homo sapiens BRCA1, DNA repair associated (BRCA1), transcript variant 3, mRNA	12665	13144	98%	0.0	100%	NM_007297.3
PREDICTED: Pan troglodytes BRCA1, DNA repair associated (BRCA1), transcript variant X1, mRNA	12574	12963	99%	0.0	99%	XM_009432082.2
PREDICTED: Pongo abelii BRCA1, DNA repair associated (BRCA1), transcript variant X1, mRNA	12549	12549	99%	0.0	98%	XM_003778832.2
PREDICTED: Pan troglodytes BRCA1, DNA repair associated (BRCA1), transcript variant X19, mRNA	12414	12803	99%	0.0	99%	XM_009432096.2
PREDICTED: Nomascus leucogenys breast cancer 1, early onset (BRCA1), transcript variant X4, mRNA	12327	12327	99%	0.0	98%	XM_003279496.1
PREDICTED: Pongo abelii BRCA1, DNA repair associated (BRCA1), transcript variant X19, mRNA	12041	12041	96%	0.0	98%	XM_002827484.3
PREDICTED: Pan troglodytes BRCA1, DNA repair associated (BRCA1), transcript variant X6, mRNA	12000	12000	92%	0.0	99%	XM_009432083.2
PREDICTED: Pan troglodytes BRCA1, DNA repair associated (BRCA1), transcript variant X3, mRNA	11980	11980	92%	0.0	99%	XM_009432081.2
PREDICTED: Pan troglodytes BRCA1, DNA repair associated (BRCA1), transcript variant X7, mRNA	11961	11961	92%	0.0	99%	XM_009432084.2

ตอบ: เป็นโปรตีนเกี่ยวกับมะเร็งเต้านมในมนุษย์ ซึ่งมีโปรตีนโดเมนในสายโปรตีนนี้ คือ

1. Zinc ring finger domain
2. Serine cluster domain
3. BRCT domains (แถม)