PARK, Songwon
1155152367

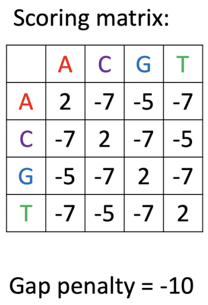
BMEG3105 Lecture 5
From Sequence to Gene Expression Matrix: Assembly and Mapping
Wednesday, 20 September 2023

Agenda

- More about DP
- Why and how do we get the gene expression matrix?
- Introduction to sequence assembly and sequence mapping
    - Overview of how we get the data
    - Not go into the algorithm details

Dynamic Sequencing

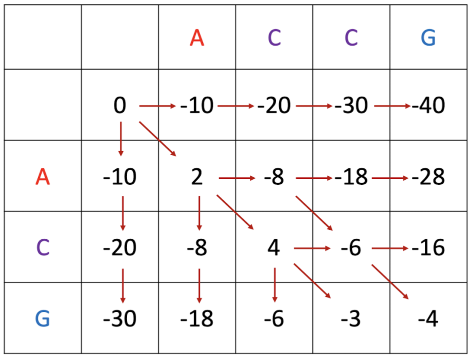
Example: finding the optimal alignment of sequences ACCG and ACG using a DP table



Scoring matrix:

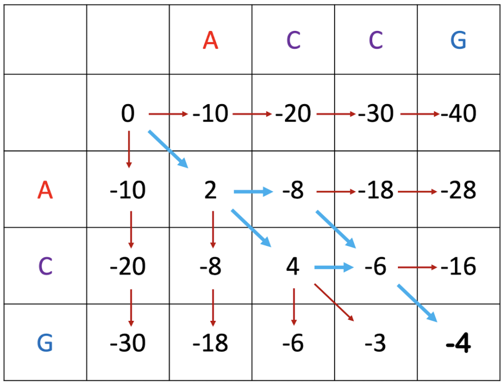
	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

Gap penalty = -10



		A	C	C	G
	0	-10	-20	-30	-40
A	-10	2	-8	-18	-28
C	-20	-8	4	-6	-16
G	-30	-18	-6	-3	-4

- Horizontal and vertical arrows = gap
- Diagonal arrows = match and mismatch
- From the last cell, trace back along the most positive numbers to find the optimal alignment



		A	C	C	G
	0	-10	-20	-30	-40
A	-10	2	-8	-18	-28
C	-20	-8	4	-6	-16
G	-30	-18	-6	-3	**-4**

- Two optimal alignments in this case
- ACCG and A_CG
- ACCG and AC_G

Summary:

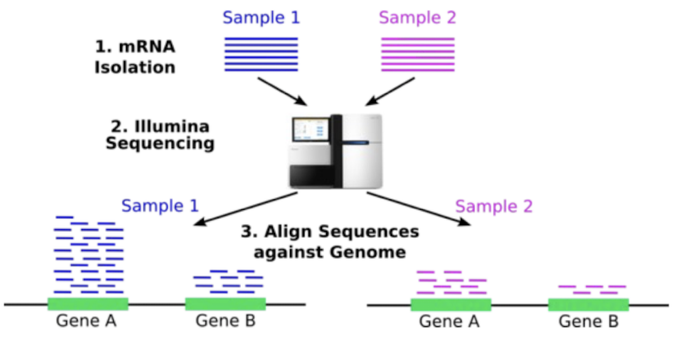
- Dynamic programming divides the original problem into smaller problems and solves the original problem recursively
- It is a much more efficient method than simple enumeration; when aligning two sequences of length 300, we only need to fill in 90000 cells of the DP table
- DP table stores answer of sub-problems and construction path
- Local alignment VS Global alignment
    - Local alignment tells us about similar components, motifs, and domains in dissimilar sequences
    - Global alignment considers all the bases in the sequences, eventually making the sequences the same length
- Available webserver for sequence alignment: https://www.ebi.ac.uk/Tools/psa/emboss_needle/
- Plugin in Python: https://biopython.org/

<u>Gene Expression Matrix</u>

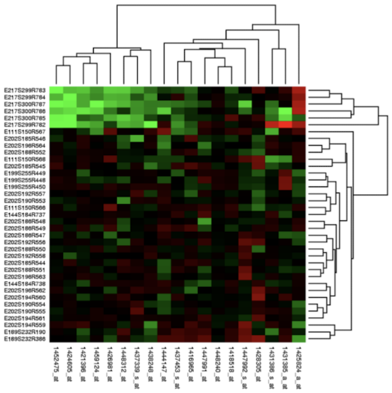
Why do we need the gene expression matrix?

- Phenotype = Genotype + Environment
- Biologists believe genotype is determined by the sequences
- However, our genomes are very similar
    - Genetic variation only accounts for 0.001% in the genome
    - Protein-encoding genes only account for 1% of the genome
- Therefore, we need to know the gene expression difference which may account for the phenotype difference

How do we get the gene expression matrix? → RNA Sequencing



→



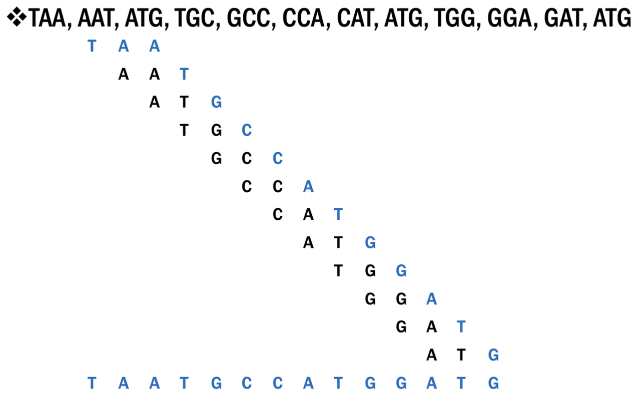
1. Isolate the mRNA in samples 1 and 2

2. Perform RNA sequencing (ex. Illumina sequencing), which will chop the mRNA into smaller pieces
3. Map the read we have sequenced back to the genome
4. Count how many reads are mapped for each gene → this number is visualized in the gene expression matrix
    a. A red cell indicates a large number of reads
    b. A green cell indicates a low number of reads

Sequence Assembly and Genome Mapping

Genome Assembly: assembled based on short reads and overlap regions

Example:



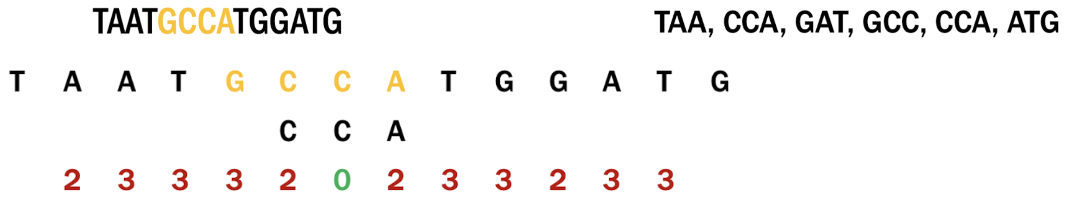
❖TAA, AAT, ATG, TGC, GCC, CCA, CAT, ATG, TGG, GGA, GAT, ATG

```
T A A
  A A T
    A T G
      T G C
        G C C
          C C A
            C A T
              A T G
                T G G
                  G G A
                    G A T
                      A T G
T A A T G C C A T G G A T G
```

Problems with this method:

- What if there are mutations?
- What if there are conflicts?
- What if there are repeat sequences?
- What if there are repeat genes and how do you determine the copy number?
- How do you make the algorithm faster?

Mapping example:



TAATGCCATGGATG   TAA, CCA, GAT, GCC, CCA, ATG

```
T A A T G C C A T G G A T G
          C C A
  2 3 3 3 2 0 2 3 3 2 3 3
```

- Slide each read along the genome and calculate how many bases are different from the gene
- “0” indicates a perfect match
- Count how many reads are perfectly matched = gene expression count
    - In this example, the gene expression count is 3