LIU Yuze BMEG3105

1155141485 Scribing: Lec-4 September 16, 2022

(Tip: Content with blue title is just my additional search, not included in the class)

Outline of lecture

- More about Dynamic Programming
- Gene expression matrix
- Sequence assembly and sequence mapping

0 Recap from last lecture

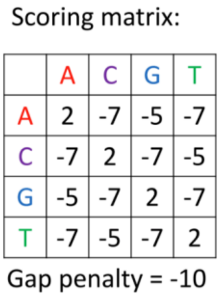
- Dynamic programming
  - ✓ Purpose: An efficient way to find optimal sequence alignment, which could be applied in identification of sequence similarity
  - ✓ Definition: Simplifying a complicated problem by breaking it down into simpler sub-problems in a recursive manner
  - ✓ Key Idea: “Divide & Conquer”
  - ✓ Dynamic table: Storing answers of sub-problems and the construction path, instead of listing all numeric results

The idea of “Divide & Conquer” behind dynamic programming is more important than the algorithm itself!!

1 More about Dynamic programming

- Procedure

1. Define a scoring matrix

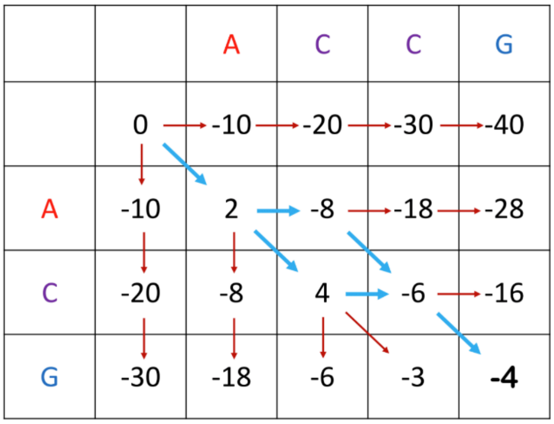


Scoring matrix:

	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

2. Fill the table with arrows preserved (for final alignment tracking)



		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**

3. The best alignment score locates in the last cell

4. Track back the arrows to get the alignment

Global and local alignment

- For local alignment, we only consider the value in the last cell
- For two sequence a and b, local alignment focuses on best alignment sub-sequences of them, which could find similar components, motifs and domains in dissimilar sequences.

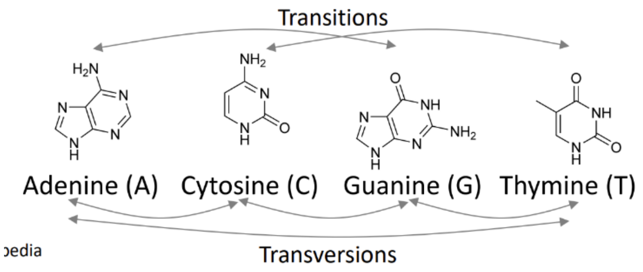
Scoring matrix

- Biological significance
  1. Mismatch: mutations
  2. Gap: Insertion/deletion, gene duplication
- Different scoring matrix for different purpose
  1. An example in on-class uReply

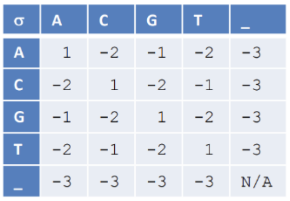
Question: What's the scoring matrix if we define the sequence similarity as "the total number of matched bases"?

Possible answer: Matches: 1; Mismatches: 0; Gap: 0; (To maximize the effect of matches)

2. Character similarity (From BMEG3102)



A possible scoring matrix:



σ	A	C	G	T	_
A	1	-2	-1	-2	-3
C	-2	1	-2	-1	-3
G	-1	-2	1	-2	-3
T	-2	-1	-2	1	-3
_	-3	-3	-3	-3	N/A

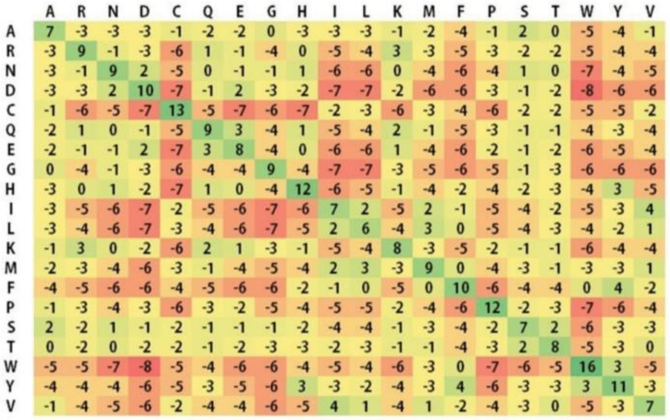
A & G are purines (two rings) while C & T are pyrimidines (one ring)

Milder consequences for changes to the same type (purine to purine or pyrimidine to pyrimidine): "transition" (-2 in the matrix)

More drastic consequences for changes to the other type (purine to pyrimidine or pyrimidine to purine): "transversion" (-1 in the matrix)

3. BLOSUM

BLOcks SUbstitution Matrix (BLOSUM)

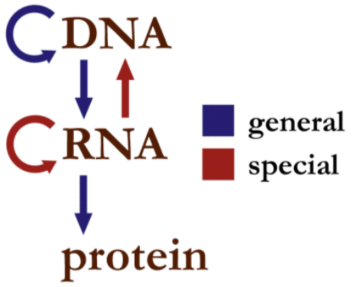


	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	7	-3	-3	-3	-1	-2	-2	0	-3	-3	-3	-1	-2	-4	-1	2	0	-5	-4	-1
R	-3	9	-1	-3	-6	1	-1	-4	0	-5	-4	3	-3	-5	-3	-2	-2	-5	-4	-4
N	-3	-1	9	2	-5	0	-1	-1	1	-6	-6	0	-4	-6	-4	1	0	-7	-4	-5
D	-3	-3	2	10	-7	-1	2	-3	-2	-7	-7	-2	-6	-6	-3	-1	-2	-8	-6	-6
C	-1	-6	-5	-7	13	-5	-7	-6	-7	-2	-3	-6	-3	-4	-6	-2	-2	-5	-5	-2
Q	-2	1	0	-1	-5	9	3	-4	1	-5	-4	2	-1	-5	-3	-1	-1	-4	-3	-4
E	-2	-1	-1	2	-7	3	8	-4	0	-6	-6	1	-4	-6	-2	-1	-2	-6	-5	-4
G	0	-4	-1	-3	-6	-4	-4	9	-4	-7	-7	-3	-5	-6	-5	-1	-3	-6	-6	-6
H	-3	0	1	-2	-7	1	0	-4	12	-6	-5	-1	-4	-2	-4	-2	-3	-4	3	-5
I	-3	-5	-6	-7	-2	-5	-6	-7	-6	7	2	-5	2	-1	-5	-4	-2	-5	-3	4
L	-3	-4	-6	-7	-3	-4	-6	-7	-5	2	6	-4	3	0	-5	-4	-3	-4	-2	1
K	-1	3	0	-2	-6	2	1	-3	-1	-5	-4	8	-3	-5	-2	-1	-1	-6	-4	-4
M	-2	-3	-4	-6	-3	-1	-4	-5	-4	2	3	-3	9	0	-4	-3	-1	-3	-3	1
F	-4	-5	-6	-6	-4	-5	-6	-6	-2	-1	0	-5	0	10	-6	-4	-4	0	4	-2
P	-1	-3	-4	-3	-6	-3	-2	-5	-4	-5	-5	-2	-4	-6	12	-2	-3	-7	-6	-4
S	2	-2	1	-1	-2	-1	-1	-1	-2	-4	-4	-1	-3	-4	-2	7	2	-6	-3	-3
T	0	-2	0	-2	-2	-1	-2	-3	-3	-2	-3	-1	-1	-4	-3	2	8	-5	-3	0
W	-5	-5	-7	-8	-5	-4	-6	-6	-4	-5	-4	-6	-3	0	-7	-6	-5	16	3	-5
Y	-4	-4	-4	-6	-5	-3	-5	-6	3	-3	-2	-4	-3	4	-6	-3	-3	3	11	-3
V	-1	-4	-5	-6	-2	-4	-4	-6	-5	4	1	-4	1	-2	-4	-3	0	-5	-3	7

BLOSUM matrices are used to score alignments between evolutionarily divergent protein sequences. They are based on local alignments.

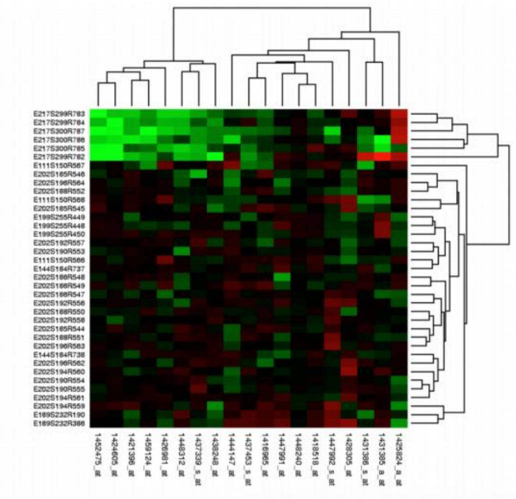
2 Gene expression matrix

➢ Central dogma



By transcription and translation, DNA sequence encodes protein expression, which further controls the phenotype of human being.

➢ Gene expression matrix



- Column: Sample (Cell)

  Row: gene

  Color: gene expression level (Green: positive correlated; Red: negative correlated)

  Clustering: The similarity/distance between the samples/genes

- How to generate?

1. mRNA isolation

mRNA sequences are isolated from samples and then prepared to be sequenced.

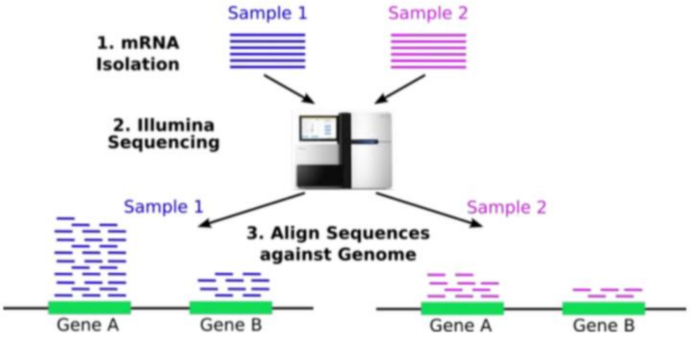
2. Illumina sequencing

   This sequencing method is based on reversible dye-terminators that enable the identification of single nucleotides as they are washed over DNA strands.

3. Align sequences against genome

   After mRNA are cut into short reads and sequenced, they are mapped to the genome. Then we count the number of these reads to get the specific expression level of different genes in these samples.

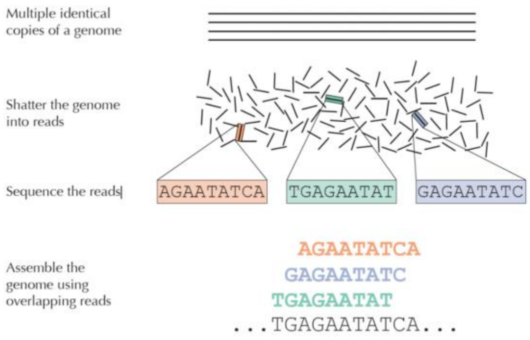
   Note: Longer genes will have more reads, so we should normalize by gene length to determine expression.

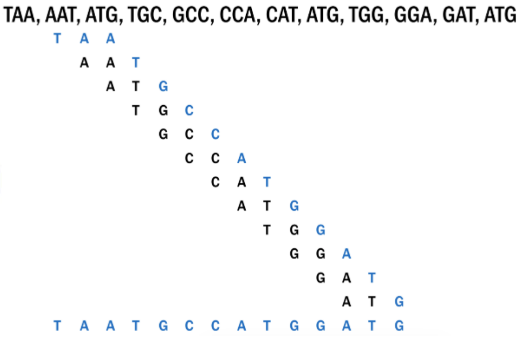


3 Sequence assembly and sequence mapping

- Genome assembly

  Human genome length: 3 * 10^9 bp
  Illumina sequencing length: 200 bp
  Key idea: Use overlap region to connect the short reads to construct the whole genome


Multiple identical copies of a genome
Shatter the genome into reads
Sequence the reads
AGAATATCA
TGAGAATAT
GAGAATATC
Assemble the genome using overlapping reads
AGAATATCA
GAGAATATC
TGAGAATAT
...TGAGAATATCA...

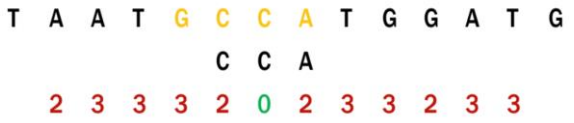

TAA, AAT, ATG, TGC, GCC, CCA, CAT, ATG, TGG, GGA, GAT, ATG
T A A
A A T
A T G
T G C
G C C
C C A
C A T
A T G
T G G
G G A
G A T
A T G
T A A T G C C A T G G A T G

Possible problems:

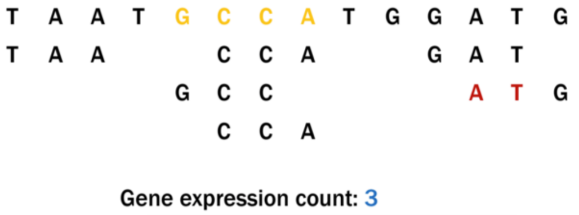
1. Mutation: When mutation occurs, we may find there are no or multiple overlaps to align
2. Conflict: For example, we have AAC, AAG, AAA to align with TAA, which one to be chosen is conflict
3. Repeat sequences: If the genome has such a following part "AAAAAAAAAA", we could not determine the real length of A by short reads such as "AAA"
4. Copy number: If the genome has such a following part "ATTATTATT", we could not determine the real copy number of "ATT"

➢ Sequence mapping

Method: Slide each read along the genome, calculate the difference



```
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
```



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

      Gene expression count: 3
```

Just slide the "CCA" read along the genome "TAATGCCATGGATG" and find the minimum difference place, where "CCA" matches

➢ Useful alignment tool (From BMEG3102)

BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool)

- Proposed by Altschul et al. in 1990
  (Altschul et al., J. Mol. Biol. 215(3):403-410, 1990)
- Probably the most frequently use algorithm in bioinformatics
- Main idea: finding local matches, then extending and combining them

  For example: we have a query sequence r: ACGTTGCT
  First, BLAST considers the sub-sequence of length-3 "ACG", then looks for the locations of ACG and other similar length-3 sequences like "ACG CCG GCG…" in the database. After that, BLAST extends it by including the adjacent characters in the two ends until the match score drops below a threshold

Potential project-1

A pipeline to get the gene expression matrix from reads

- Find the genome
- Find the reads
- Map reads to reference genome
- Count reads for each gene
- Use Google to find the software and the data
- Explain each step in the report to let us know you understand what are you doing

Useful links

Biopython: https://biopython.org

BLAST search: https://blast.ncbi.nlm.nih.gov/Blast.cgi

YouTube video for gene expression: https://www.youtube.com/c/niemasd