Lecture 3: The Foundation of Modern Biology and Genomics: Sequence Data

Course: BMEG3105, Data analytics for personalized genomics and precision medicine

Professor: Yu Li

Scriber: Asset Yermukhanbet

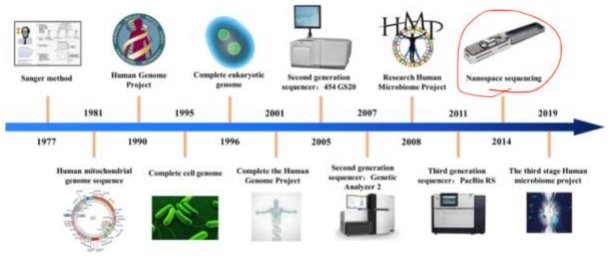
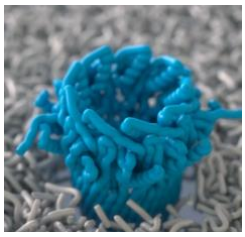
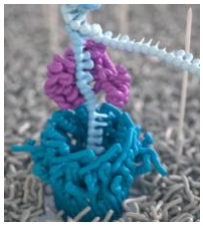
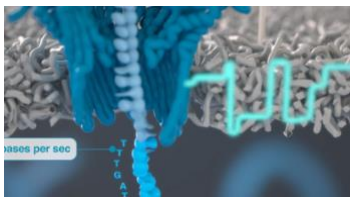
Brief Summary

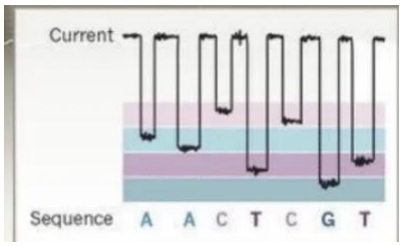
Module	**Key Points**
Sequence Data	- Why Sequence Data? - How do we sequence the data? - Nanopore sequencing - Protein sequencing - What to do with the raw data? - Sequence-to-structure-to-function paradigm is what?
Dynamic Programming	- Sequence alignment - Sequence alignment and sequence similarity - How to find the best alignment? - Example of sequence alignment using Dynamic Programming - Summary - Additional resources

Module I

Sequence Data

Why sequence data?	Sequential data plays an important role in genetics as DNA* and RNA* are stored in a sequential order. *DNA (or deoxyribonucleic acid) contains a combination of 4 nucleotides (Adenine, Guanine, Cytosine, Thymine) *RNA (or ribonucleic acid) contains a combination of 4 nucleotides (Adine, Cytosine, Guanine, Uracil)  But why we need to analyze sequential data? In biology, there is a terminology called phenotype, which essentially represents any **visible features** that we, living beings, have. Whether it is an eye color or nose size, phenotypes differentiate us from each other. It is clear that phenotype is determined and shaped by **genotype**, a being's genetic material.

	Genotype is powerful to make phenotype occur/exist. However, it is also important to understand that in addition to the genotype, the environment in which the being is exposed to shapes/impacts the phenotype (such as, climate and air quality) Since now we know that genotype is important to determine the conditions/features/mutations/diseases of a patient, we may use it in our data analysis and inference processes
How do we sequence the data?	DNA/RNA sequencing technologies have taken their roots back in 1977 and have been subject to innovation ever since then. The below chronological diagram represents transformations that sequencing technologies have experienced for the past 50 years.  The focus of this course would be on **nanopore sequencing** technology and its methodology
Nanopore Sequencing	Nanopore sequencing is a third generation approach used in the sequencing of biopolymers (specifically, DNA or RNA) Methodology: 1. There is a continuous current going through the nanopore  2. Once the DNA/RNA strand goes through the nanopore, the flow of the current gets disrupted  

	3. Each base can be identified by how much **disruption level** (electric signal) it causes to the flow of the current in real-time.  Advantages of nanopore sequencing? ➔ Capable of long reads of data ➔ Real-time read Any drawbacks? ➔ Error rate historically has been remaining high **Video Illustration:** https://www.youtube.com/watch?v=RcP85JHLmnI
Protein sequencing	Protein takes the form of a chain of amino acids These days, protein sequencing is mostly based on the mass spectrometry (MS) technology Methodology: 1. The sequence is partitioned/broken down into small pieces 2. Each piece can be identified by the Mass Spectrometry due to its weight value 3. The short pieces are then assembled into the raw sequence
What do we do with the raw data?	Several actions are taken when assembling the raw pieces into the chain: ➔ **Quality Control:** the process of removing the noise (random, inaccurate, meaningless, context-free information that disrupts the analysis and inference of the meaningful data) from the data ➔ **Alignment/Mapping:** Identify if the sequence from the data aligns with any reference genome ➔ **Variant Calling:** Identify the differences and variants from the sequence we have obtained to that of the reference sequence ➔ **Phenotype Associated Variants:** Candidate Diseases or Phenotypes associated with the observation

Sequence-to-Structure-to-Function Paradigm is what?	The paradigm essentially entails that the sequence of the protein signals the structure of the protein which signals the function of the protein

Module II

Sequence Alignment and Dynamic Programming

Sequence alignment	Sequence alignment helps us to determine the similarity between sequences and identify regions of similarity which might tell us about their relationship in terms of functionality and the possibility of sharing a **common ancestry (**homology) Why finding the sequence similarity is crucial? ➔ Allows to predict the biomolecular function and the property of the data ➔ Helps to identify the conservative regions (regions that have not been altered evolutionary) and the non-conservative regions (the areas that incur mismatch) ➔ Identify any mutations/changes in the newly sequenced data 
Sequence Alignment and Sequence Similarity	There are three type of instances when performing sequence alignment and calculating the score: ➔ Match: A <-> A ➔ Mismatch: A <-> C ➔ Gap: C <-> _ There is scoring matrix that we can use of as a reference to calculate the alignment score* and see which alignment version yields the highest value *The alignment score is the sum of the score for each pair in the alignment

<table>
<tr>
<td></td>
<td>
<table>
<tr><td></td><td>A</td><td>C</td><td>G</td><td>T</td></tr>
<tr><td>A</td><td>2</td><td>-7</td><td>-5</td><td>-7</td></tr>
<tr><td>C</td><td>-7</td><td>2</td><td>-7</td><td>-5</td></tr>
<tr><td>G</td><td>-5</td><td>-7</td><td>2</td><td>-7</td></tr>
<tr><td>T</td><td>-7</td><td>-5</td><td>-7</td><td>2</td></tr>
</table>
Gap penalty = -10
*All the values in the matrix have been calculated statistically
</td>
</tr>
<tr>
<td>How to find the best alignment?</td>
<td>Solution 1: Enumerate all the possible outcomes which is very naïve since the sequences might be very long and, hence, result in thousands or millions of combinations possible. Say, if we have a sequence of length m, where m = 300, we will be having 7 * 10^(88)

Thus, we may need to use another solution: dynamic programming

Dynamic Programming is one of the most popular approaches in computer science:
1. Divide the problem into smaller sub-problems
2. Solve these overlapping sub-problems optimally and recursively/iteratively*
3. Use the solutions obtained from the smaller problems to construct a solution for the bigger original problem

*for each sub-problem solution, we store the value into the memory</td>
</tr>
<tr>
<td>Example of sequence alignment using Dynamic Programming</td>
<td>Input sequences: ACCG and ACG

Let F be the function that determines the most optimal alignment score for two input sequences. Thus, the question persists, F(ACCG, ACG) = ?
For the simpler understanding, let’s use table representation with the author’s comments on each step

<table>
<tr><td></td><td></td><td>A</td><td>C</td><td>C</td><td>G</td></tr>
<tr><td></td><td>0</td><td>-10</td><td></td><td></td><td></td></tr>
<tr><td>A</td><td>-10</td><td>2</td><td></td><td></td><td></td></tr>
<tr><td>C</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>G</td><td></td><td></td><td></td><td></td><td></td></tr>
</table>

The comments: we start at the 0 point, on the top left corner</td>
</tr>
</table>

Rules we must follow:

- ➔ Moving diagonally implies either match or mismatch
- ➔ Moving vertically or horizontally implies gap
- ➔ We must consider all possible paths to each cell and calculate the best score possible
- ➔ The score of best alignment will be represented on the bottom right corner

Step 1: Since we are moving to the right and bottom, we will have -10 in each respective cell, since rule#2 obliges us to consider it as a gap. The diagonal movement leads us to A and A which is a match, and thus, rewards 2 points

Step 2:

		A	C	C	G
	0	-10	**-20**		
A	-10	2	**-8**		
C	**-20**	**-8**	**4**		
G					

Step 2: 4 has resulted from going from A-A cell to the C-C which is the best solution possible out of all yielding 2+2 = 4 points. -20 is obtained by going from -10 to the right, which clearly yields -10 – 10 = -20 points. -8 score is obtained from taking a step from 2 points A-A to the A-C (which implies the gap), yielding 2 – 10 = -8 points

Further steps are shown and self-explanatory

		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**

Optimal Score: -4
Optimal Alignment:

- ➔ ACCG
  A_CG

- ➔ ACCG

	AC_G
Sooo in summary…?	Dynamic Programming allows us to find **the best alignment solution** and the score by recursively solving the sub-problems Since the table takes the form of square, we will have a total of m * m combinations which is way better than using enumeration technique Score matrix is very important reference material as it helps to determine the path we will be taking when finding the optimal alignment score
Additional Resources	Webserver for sequence alignment: https://www.ebi.ac.uk/Tools/psa/emboss_needle/ Biopython: https://biopython.org/

Reference Materials:

Yu Li. (September 2025). The Foundation of Modern Biology and Genomics: Sequence Data. The Chinese University of Hong Kong.

[https://www.dropbox.com/scl/fi/1du4obeok262k7oj7rx9m/lec3-DP.pdf?rlkey=49t1ppaivd2roaqo6u0n1ej17&e=1&dl=0]