Lecture 1 - An Introduction to Algorithms in Molecular Biology

Gidon Rosalki

2025-10-19

Notice: If you find any mistakes, please open an issue at `https://github.com/robomarvin1501/notes_algorithms_computat`

1 Course details

This course is based on the book *Biological Sequence Analysis*, by Richard Durbin.

There is no exam in this course. The final grade is comprised of 15% of writing a final paper (without AI). 35% of homework (allowed to be done in pairs, and AI is allowed, with explicit mentioning of what the AI did). 40% from a final project or hackathon (groups of 2 or 4), and 10% on the midterm exam.

2 Introduction

To begin with, we shall create some definitions:

1. Population (all cats in Tel Aviv)
2. Organisms (cat)
3. Organs / tissues (heart)
4. Cells ($10^{-5}m$)
5. Organelles ($10^{-6}m$)
6. Macro molecules (DNA, protein $10^{-7}m$)
7. Small molecules ($10^{-9}m$)
8. Atoms ($10^{-10}m$)

We will mainly focus on cells, organelles, and macro molecules. These are with what molecular biology concerns itself.

1. DNA is made up of nucleotides, A, C, G, and T.
2. RNA is also made of of nucleotides, but this time A, C, G, U.
3. Proteins are made up of 20 different amino acids.
4. Sugars. Surprising I know, but they are in fact made up of these sequences. They are not sugars like we eat on the day to day, but are similar. (will not be greatly discussed in this course).
5. Lipids.

DNA is comprised of the nucleotides A, C, G, and T, as stated above. It is in a double helix shape, where each strand is a sequence of these nucleotides, and the opposing strand has the opposite nucleotide. A is always opposite T, and C is opposite G. RNA is very similar, but we replace T with U. A central dogma of molecular biology is that DNA is used for **replication**, and RNA is used for **transcription**. This RNA is then **translated** into proteins. In translation overall, we take a sequence of DNA, which is translated into RNA, and then into proteins. We may consider the translation of letters into binary / hexadecimal for representation on computer hard drives. The same data is stored, but it is kept in a different form. Another example of this is binary to hexadecimal. An interesting thing to note is that in 4 numbers of binary, there are 16 options, and naturally there are 16 letters in hexadecimal. Therefore, each sequence of 4 binary numbers can be stored as a single hexadecimal character. Similarly, in molecular biology, each sequence of 3 nucleotides from AAA to TTT (of which there are 64) can be translated into 1 of 21 named codons, one of which is the STOP codon. Since $64 > 21$ note that many sequences of nucleotides translate into the same codon.

Given this, if you have a sequence of DNA: CAATGTAACTGGTTTA... Should you see a start codon within it (ATG) then it is probable that this is the start of the encoding for a protein. Finding the next STOP codon will indicate the end of said protein. It is possible that there will be many start codons before finding a stop codon.

How DNA is translated is dependent on where is the start codon, and in which direction the DNA is read, i.e. which strand is the active encoding strand, and which start codon is then used.

When considering molecular biology algorithms, we do not just want to learn the algorithm, but rather also its context, why it was made, and why it works as it does. To understand this, we need the following:

1. Biological question: Given a protein sequence, what is this protein? What is its function?
2. What is the relevant information / principle? If a sequence is similar, this can imply that there is a common ancestor, and thus these sequences have the same function.
3. Data: A protein database, that maps sequences to functions.
4. Algorithm / maths question: Sequence comparison / queries. This is the algorithmic question. Given someone who does not understand biology, we can give them access to these data, and ask them to find similar sequences and the like. This does leave us with questioning the definition: What is similarity?
5. Algorithm: Find the closest distance (As in, distances between sequences of letters) in the database
6. Parameter choice. Once upon a time we would simply choose the weight of a mistake. However, since we now live in a world of data, we can **learn** a value, since different mistakes may have different impacts. Therefore we can estimate the "distance" of true relationships.
7. Statistics: We also need to ask the probability that two things are similar by chance, rather than innately.
8. Visualisation / Explanation: Why should this be the case. This is distinct from statistics, since for example the statistics will say there is an 80% chance that this photo is a photo of a cat, but when presented with said photo, one may see that this photo is of a cat.