1 Grade

The course’s book - Biological Sequence Analysis by Durbin Eddy et al.

Final grade components –

- 50% Exercises (3-5 exercises)
- 20% Scribes
- 5% Participation (?)
- 25% Hackaton / project / home exam

2 Introduction

What is special about DNA? why do we want to analyze genetic sequences?

2.1 Central Dogma in short:

- **DNA**: We can look at the DNA as a language where its alphabet is $\Sigma = \{A, C, G, T\}$. Human DNA is a string of $3 \cdot 10^9$ letters from the alphabet. The DNA is double stranded, where each nucleotide
base ($\sigma \in \Sigma$) in one strand has a complementing base in the other ($A - T$, $G - C$). Each strand can be used as a DNA replication template, in order to make a new DNA molecule. In addition, The DNA molecule is directional, meaning that reading in opposite directions gives two different results. The length of the molecule in its unraveled conformation is 2 meters, which is compressed into a nucleus of 6m.

Example:

(\textgreater \textless) ATTACGGA

TAATGCCT (\textgreater \textless)

- **Transcription**: A central biological process called “Transcription” takes a sequences of DNA code from the double-strand, and generates a complementing RNA strand. The transcription process generates multiple copies of RNA per one DNA sequence transcribed.

- **RNA**: Unlike DNA, RNA is a single-strand. RNA degrades rapidly (as opposed to the much more stable DNA) since it has only one strand and there are enzymes in the cell that actively break it down.

- **Translation**: Each mRNA sequence is read in triplets - every triplet of nucleotides encodes a specific amino acid (one of 20 naturally available in the body). The translation process connects amino acids which are encoded according to the RNA sequence, to create a linear protein - which later may fold into secondary and tertiary structures (3D structures, as opposed to linear sequence).

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{gene_structure.png}
\caption{Diagram of basic structure of a gene}
\end{figure}

2.2 DNA - The Book of Life:

The rate of RNA transcription is not the same across all genes. Control systems may enhance or reduce “gene expression” - or RNA transcription.

In each cell in our body the DNA sequence is identical, but in different tissues (for example, the brain and the liver), different proteins are needed. How can we know which gene should be expressed in which cell? In the promoter\(^1\) there is a “switch” that can suppress the gene expression in a particular cell type and enhance it in another.

Our research may regard DNA or RNA, but what is actually important in biology are the proteins expressed in the cells themselves. So why is most of the research dedicated to the genome itself? This is mostly because it is much easier to sequence and analyze compared to proteins.

There are a few technologies for DNA sequencing, for example Sanger sequencing method.

\(^{1}\text{Promoter is a regulatory region upstream the gene.}\)
3 Dealing with a Scientific Question

3.1 Research approaches: How to approach a scientific problem as computational biologists?

- Problem 1 - Suppose a biologist provides us with a protein sequence and some biological question, such as “What does this protein do?”. How do we translate this question into a mathematical one? Logically, we can assume that the sequence has an effect on the structure, which determines the protein’s function. So if we can find a similar sequence in a known protein for which the function (and/or structure) is known - we can then deduce that this protein has a similar function. This brings us to a new question - what is similarity? How do we quantify similarity. We will explain later how we score similarity. Another question is how can we determine the statistical significance of the similarity (is this similarity just a coincidence?).

- Problem 2 - A recently new field of research called “Systems biology” investigates the scope of an entire system. For instance - say that protein A binds to the promoter sequence of a gene for protein B, which in turn binds to the promoter of gene C which finally goes back to regulate gene A. Say that another biologist friend asks us whether or not such cyclical regulation is coincidental, or does this happen a lot within the scope of a biological system?

![Cyclical regulation (problem 2)](image)

3.2 The seven stages of solving a problem:

When approaching a scientific problem, we describe the following stages:

1. Describing the scientific question (in biological terms) as general and simple as possible.
   - What is the function of a protein?
   - What is the structure of the protein?
   - Are there similar proteins in different organisms?
   - Do similar proteins share a common ancestor? That is - is the function shared across the phylogenetic subtree.

2. Gathering information and relevant biological background about the problem.
Evolution of genetic sequences.
Properties of amino acids.
Biology of transcription regulation.
What is the significance of letter changes? When does it happen?
Relation between proteins and the DNA encoding them.

3. Translate the problem into formal mathematical/computational terms and models.
   - Describe the problem formally in terms that can be encoded in a computer program.
   - For example: Define what is the distance (in terms of similarity) between sequences.

4. Find or devise an algorithm that solves the problem.
   - For instance, find an algorithm that generates a score for sequence similarity.

5. Estimate the parameters for solving the problem.
   - Finding pairs of proteins with common ancestor to find parallels to our problem and devise parameters for it.

6. Statistical significance
   - Is it surprising to get the given outcome?
   - Does the observed phenomenon, or result of the algorithm, better than random?

7. Visualization and Biological insights
   - How to analyze and present our results in a manner that generates insights into the biological question.

4 Biological Sequence Alignment

Recall problem number 1 while using the 7 stages-

1. The question: Is there a common biological ancestor to the two sequences?

2. The relevant background: Frequencies of changes (insertions, deletions and substitutions) in biological sequence with a common ancestor.

3. The model and score: let there be sequences $\hat{S} = AGT$ and $\hat{T} = AACT$, a possible alignment is
   \[ a : A - G T. \]
   \[ A A C T. \]

A possible scoring scheme is a cumulative score where a match gives 1 point, and a mismatch or gap gives no points:

$$ score (a) = \sum_i \sigma (a_{1i}, a_{2i}) $$

\[ \sigma \] is the scoring matrix:

\[
\begin{array}{cccc}
A & C & G & T \\
A & 1 & 0 & 0 & 0 \\
C & 0 & 1 & 0 & 0 \\
G & 0 & 0 & 1 & 0 \\
T & 0 & 0 & 0 & 1 \\
- & 0 & 0 & 0 & 0
\end{array}
\]
• Examples:

\[ \text{SCORE}( \begin{array}{c} A \\ A \end{array} \begin{array}{c} G \\ A \end{array} \begin{array}{c} T \\ C \end{array} \begin{array}{c} T \\ T \end{array} ) = 2 \]

\[ \text{SCORE}( \begin{array}{c} A \\ - \\ - \\ - \\ A \end{array} \begin{array}{c} G \\ T \\ - \\ - \\ - \end{array} \begin{array}{c} - \\ A \\ A \\ C \\ T \end{array} ) = 0 \]

• Another possible scoring function:

\[
\begin{array}{ccccccc}
A & C & G & T & - \\
1 & -1 & -1 & -1 & -2 \\
-1 & -2 & -2 & -2 & \emptyset \\
\end{array}
\]

\[
\begin{array}{ccccccc}
A & C & G & T & - \\
1 & -1 & -1 & -1 & -2 \\
-1 & -2 & -2 & -2 & \emptyset \\
\end{array}
\]

\[
\begin{array}{ccccccc}
C & -1 & 1 & -1 & -1 & -2 \\
G & -1 & -1 & 1 & -1 & -2 \\
T & -1 & -1 & -1 & 1 & -2 \\
- & -2 & -2 & -2 & -2 & \emptyset \\
\end{array}
\]

• Examples for \( \sigma_2 \):

\[ \text{SCORE}( \begin{array}{c} A \\ A \end{array} \begin{array}{c} G \\ A \end{array} \begin{array}{c} T \\ C \end{array} \begin{array}{c} T \end{array} ) = -1 \]

\[ \text{SCORE}( \begin{array}{c} A \\ - \\ - \\ A \end{array} \begin{array}{c} G \\ T \\ - \\ - \end{array} \begin{array}{c} - \\ A \\ A \\ C \end{array} \begin{array}{c} T \end{array} ) = -14 \]

• So we prefer mistakes rather than gaps, because the price for a gap is higher.

• In how many ways can we align a sequence of length \( n \)? It is exponential in \( n \).