Bioinformatics Notes

Artem Vasenin

Last updated December 26, 2016

Contents

Chapter 1

Biology background

1.1 DNA

Structure of DNA DNA is structured as a double helix. Each helix is a sequence of nucleotides, there are four types of nucleotides: A,C,T and G. The sequences are aligned with each other such that at each position the nucleotide on one helix is paired with nucleotide on the other helix in one of the four possibilities: A-T, T-A, C-G or G-C.

Structure of proteins DNA uses a triplet code, triplet of nucleotides is called a "codon". Each codon is a code for a specific amino acid. When multiple amino acids join together they form a protein. There are 20 types of amino acids which can be used to form a protein. Some codons are used for punctuation. A gene is a sequence of codons between start and stop codons. Each gene is responsible for one protein.

DNA - Protein translation To copy DNA, it is first untwisted and unzipped inside the nucleus. Then free RNA nucleotides form

complementary base pairs with one of the DNA strands. RNA uses a U nucleotide instead of DNA's T, so they can be paired as follows: A-U, T-A, C-G and G-C; DNA on the left, RNA on the right. When multiple RNA nucleotides bond next to each other, weak hydrogen bons form between those nucleotides. mRNA transcribes amino acid code using codons. mRNA strand is then synthesised from the sequence of RNA nucleotides. mRNA strand peels off the DNA and moves out of the nucleus into the cytoplasm.

Translation takes place on the ribosome in the cytoplasm. The ribosomes are sites of protein synthesis. mRNA strand attaches to a ribosome and tRNA molecules transport specific amino acids to the ribosome. The anti-codons and codons match up to form complementary base pairs. Peptide bonds form between the adjacent amino acids to form the polypeptide, which is a protein.

Chapter 2

Sequence Alignment

The goal of the alignment algorithms is to find the longest common subsequence of two sequences. Alignment of n sequences is an n-row matrix, where nth row represents the symbols of nth sequence (in order) interspersed by "-". Corresponding symbols in different sequences can have four different relations to each other:

Match The symbols are the same.

Mismatch The symbols are different.

Insertion "Primary" symbol is "-".

Deletion "Secondary" symbol is "-".

Matches in alignment of two sequences form "common subsequence".

Global Alignment We can represent the alignment search space as an edge-weighted directed acyclic graph. We can construct such DAC by creating a rectangular grid of nodes, with source in the top-left and destination in the bottom-right. Then we add edges, of which there are three types:

Match/Mismatch Connects neighbouring nodes diagonally from top-left to bottom-right.

Insertion Connects neighbouring nodes horizontally from left to right.

Deletion Connects neighbouring nodes vertically from top to bottom.

Global alignment can be found by finding the longest path from the source to the destination. To find the longest path we start at the source and calculate the length of path to all its children. To do that we consider all incoming edges and their source nodes. The longest path will go through the edge which has the greatest sum of its cost and path cost to its source node. Once we calculate which edge the longest path goes through that's the only information we need to store. Therefore we can iteratively calculate path length to all nodes. Finally to calculate the longest path from source to destination we start at the destination and iteratively move backwards along the stored edges until we arrive at the source.

Parametrisation We can parametrise the longest path by assigning different weights to different types of edges. Match/mismatch edges can be weighted by considering how likely it is for the codon in primary sequence to mutate into codon in the secondary sequence. These weights are usually calculated empirically. Insertions and deletions on the other hand depend only on their length. Combined insertions/deletion should usually have lower penalty that separate ones. There are two costs: cost to begin insertion/deletion and cost to extend it. Our DAC can be extended to represent that by adding vertical and horizontal edges that go across multiple nodes.

Local alignment Local alignment refers to a long sequence of uninterrupted matches. This corresponds to the idea that two sequences should not align perfectly, but rather only their ends should

align. We can compute local alignment by computing global alignment in sub-rectangle of the graph. To do this we add a zero-weight edge between top-left node in the sub-rectangle and source and another between bottom-right node and destination. Finding the longest path requires $O(n)$ memory and $O(n^k)$ time where n is the length of the longest sequence and k is the number of sequences to align.