MATH564 L22: Genomics

- DNA sequencing of a genome
- Mathematics of gene sequencing
- BLAST algorithm

DNA sequencing of a genome

- Nucleotides: monomeric units of the nucleic acid polymers deoxyribonucleic acid (DNA) and ribonucleic acid (RNA)

  $$\text{nucleotide} = \text{nucleobase} + \text{sugar} + \text{phosphate}$$

  Nucleobases: adenine (A) - thymine (T), cytosine(C) - guanine (G) in DNA, A- uracil (T), C-G in RNA

  Sugars: ribose (RNA) or deoxyribose (DNA)
- Genome: complete sequence of nucleotides forming genetic material (e.g., chromosomes)
- Chromosome: DNA polymer encoding genetic information, divisible into genes that encode a specific protein. Humans have 23 chromosomes with 200-2000 genes, and 50 to 250 million base pairs (BP) each, 3 billion BPs overall
- Sequencing: DNA extraction, fragmentation (~ 0.1 MBPs), cloning, sequencing

DNA combinatorics

- Reference: Combinatorics of Genome Rearrangements, Fertin et al.
- Gene: a base-4 number with $\mathcal{O}(10^5)$ digits, $4^{100000} \simeq 10^{60000}$
- Approach: break gene into subsegments of $\sim 10^3$ BPs, with overlap.
- BLAST: basic local alignment search tool

→ Smith-Waterman algorithm for local sequence alignment:

  – compare segments of all possible lengths

  – define a similarity measure to be optimized

→ Needleman-Wunsch algorithm