Numerical Representation of a Genome

Reza Mazloom

October 2022

What we refer to as a genome sequence is a set of short nucleotide sequences, called reads, identified by a sequencing machine and then assembled into longer and more contiguous sequences called scaffolds. In this section we will introduce a numerical notation to represent the assembled genome. We will be using equation (1) as our representation for a genome assembly A.

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & ... & a_{1n} \\ a_{21} & a_{22} & a_{23} & ... & a_{2n} \\ a_{31} & a_{22} & a_{33} & ... & a_{3n} \\ a_{41} & a_{42} & a_{43} & ... & a_{4n} \end{bmatrix} \tag{1}$$

In this matrix the length of the matrix (sequence) is represented as $\overline{A} = n$. Each column, also denoted by j below, represents the probability of the sequence having one of the four possible nucleotides at position j, also referred to as a base in the genome sequence. The subscript i represents each of the four nucleotide types, we formalize this such that:

$$\begin{aligned} a_{1j} &= P(A_j = \mathbf{A}denine) \\ a_{2j} &= P(A_j = \mathbf{C}ytosine) \\ a_{3j} &= P(A_j = \mathbf{G}uanine) \\ a_{4j} &= P(A_j = \mathbf{T}hymine) \end{aligned}$$

We expect, based on biological properties of a DNA, that each base j must have one of the four possible nucleotides Adenine, Cytosine, Guanine, or Thymine, referred to A, C, G, T in short. Therefore we can ascertain:

$$P(A_j) = \sum_{i=1}^{4} a_{ij} = 1 \tag{2}$$

Hence, in cases where the nucleotide at position j is an unknown nucleotide, denoted by"N" in the FASTA and FASTQ seqeunce formats, we will set all probabilities to an equal value, conveying uncertainty.

$$a_{1j} = a_{2j} = a_{3j} = a_{4j} = 0.25 \iff P(A_j = u\mathbf{N}known) = 1 \tag{3}$$

In the process of extracting the assembled genome A from the true biological genome sequence G each step of the process introduces uncertainties, possible errors, into the assembled sequence. Therefore, given A we can define $G = A - d(G, A)$ where $d(G, A)$ is the percentage of different nucleotides between the two sequences. if we define the matrix G similar to A, we can calculate the percentage difference as:

$$d(G, A) = 1 - \frac{\lfloor G \odot A \rceil}{max(\overline{G}, \overline{A})} = 1 - \frac{\sum_{j=0}^{max(\overline{G},\overline{A})} \sum_{i=0}^{4} \lfloor g_{ij} a_{ij} \rceil}{max(\overline{G}, \overline{A})} \tag{4}$$

Where $G \odot A$ is the Hadamard (element-wise) product of the two matrices and $\sum_{i=0}^{4} \lfloor g_{ij} a_{ij} \rceil$ is a binary sequence denoting correct nucleotide identification where 0.5 is rounded up. Note that the smaller matrix is padded with zeros to equalize the matrix sizes.