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 G.

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

In this matrix the length of the matrix (sequence) is represented as $\overline{G} = n$. 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:

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

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:

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

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.

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

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

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

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