Jennifer Yu Bi/CS183 PSET #2

2a. According to the article, a one sentence definition of the gene that summarizes the definitions from the years:

1860s – 1900s: The gene is a discrete unit of heridity and generally stands for the ability to determine characteristics of an organism that may be inherited.

1910s: The gene was defined as a locus determined by its mutations that affected traits following the discovery and creation of the first genetic map.

1940s: After studying mutations, the gene was redefined as the information behind proteins and molecules in a biochemical pathway when it was discovered that mutations caused defects in metabolic pathways.

1950s: the gene is realized as a physical molecule, after it was discovered that x-rays could cause mutations and that bacteriophages transferred DNA to their progeny.

1960s: The gene is understood as transcribed code from the DNA molecule following the discovery of Watson and Crick on the double helix structure and leading to the explanation of base pairing and the RNA transcript of protein coding sequences.

1970s – 1980s: The gene is defined by its predicted sequence and identified based on similarity to other genes following the development of sequencing and cloning techniques along with more knowledge about genetic code.

1990s – 2000s: A gene is currently defined generally as a region of genomic sequence which consists of different regions such as transcribed regions, regulatory regions, etc, and alludes to the annotation of a gene / how its sequenced.

b. I think of a gene similarly to the most modern definition, as a region of DNA that contributes to phenotype expression or otherwise, can be sequenced and/or transcribed.

c. Looking at table 1, the epigenetic modifications is one possible contradiction, as the inherited information may not be based off the DNA sequence. Additionally, RNA editing is a potential issue, since the information on the DNA is not directly encoded into RNA.

f. An algorithm that would work to output a random global alignment of s_1 and s_2 with probability proportional to its score is to implement the Needleman-Wunsch algorithm with the scores interpreted as the logarithms of probabilities or if there are negative/positive values, as log/odds probabilities, as described in Chapter 2 of Biological sequence analysis (given to us on Moodle as reading). For the probability to be proportional to the score, we take the random probability that aligned pairs occur together assuming that a base occurs with some independent frequency, resulting in the following product sum: $P(a,b) = \prod_i q_{a_i} \prod_i q_{b_i}$, frequency of base. This then divides ~~We then need to divide this~~ ~~by~~ a joint probability that residues are derived together, which is simply $P(a,b) = \prod_i p_{a_i b_i}$. The odds ratio is:

L joint probability

$$\frac{\prod_i p_{a_i b_i}}{\prod_i q_{a_i} \prod_i q_{b_i}} = \prod_i \frac{p_{a_i b_i}}{q_{a_i} q_{b_i}}$$

* this works like the forward algorithm for a HMM.

As such, we can set the score equal to the log of this ratio. Needleman-Wuntsch is a recursive algorithm, and essentially replacing the max with addition and the addition of scores with multiplication, and log(probability) with probabilities means it can compute total probability of a sequence ~~an alignment with a score~~ alignment.* By using ~~proportional to the~~ the logarithms of the probability ratio, as the scoring, we can express the sum as a product, resulting in a score proportional to probability.