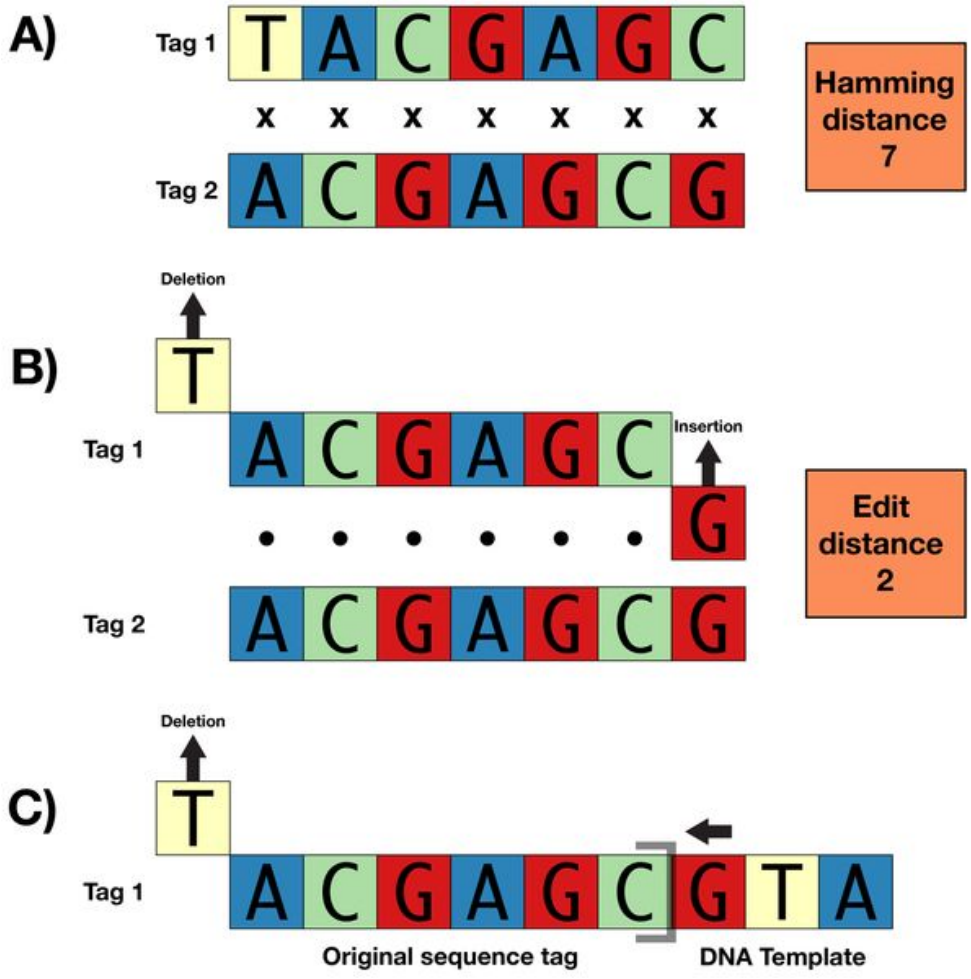
Measuring Similarity, continued

12/3/18

Last time:


A)
Tag 1
T A C G A G C
x x x x x x x
Tag 2
A C G A G C G
Hamming distance 7
B)
Deletion
T
Tag 1
A C G A G C
Insertion
G
Tag 2
A C G A G C G
Edit distance 2
C)
Deletion
T
Tag 1
A C G A G C G T A
Original sequence tag
DNA Template

Genetic distance?

Genes are located particular *loci* on the genome. Measuring variations in genes at the same loci -- called alleles -- can cause phenotypic variation in an organism.

The math for calculating these distances is pretty involved -- we won't delve into it here, but you're welcome to explore:

- https://www.journals.uchicago.edu/doi/10.1086/282771
- https://www.cell.com/ajhg/fulltext/S0002-9297(07)62616-0
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1762596/
-

Distance matrices

...are pretty self-explanatory! Using different metrics, these matrices hold information about how ‘far’ pairs of sequences are from one another. Some of these metrics:

- Euclidean distance
- Hamming distance
- Least squares
- etc!

BLOSUM##

...stands for **BLO**CK **SU**BSTITUTION **M**ATRIX. Great for comparing evolutionarily divergent sequences!

In an *ij* substitution matrix, the value at coordinate *(i, j)* will be equal to the probability that the base at *i* is replaced by the base at *j* in a dataset.

‘Each amino acid is more or less likely to mutate into various other amino acids.’

##: gene segments with similarity above a certain threshold (##%) are *clustered* together, reducing their weight during the analysis. (BLOSUM62 sets this threshold at 62%) Would you want a higher or lower BLOSUM number for very similar sequences? Very different sequences? Why?

From matrices to trees

Tree space is the collection of all possible trees satisfying the scores in your distance matrix.

Once you have your tree space, there are multiple ways to pare it down so that you've got the optimal tree.

See here for more info:
https://homes.cs.washington.edu/~ruzzo/courses/gs559/09wi/lectures/7A_distance.pdf