1.

For n OTUs, the number of pairwise distances D_n between OTUs is $\binom{n}{2}$.

For 2 OTUs, there is a single branch connecting them. For each additional OTU, a new bifurcation is created, introducing 2 additional branches. Therefore the number of branches L_n for a tree with n OTUs is $2n - 3$.

For $n = 2, 3$, there is a unique topological arrangement of the n OTUs. For $n > 3$, however, an additional OTU can branch off from any existing branch in the topology. We use this fact to prove the number of distinct topologies for $n \geq 3$ OTUs by induction.

Let T_n be the number of distinct topologies for n OTUs. First, we consider $n = 3$. There is a unique topological arrangement of 3 OTUs, so $T_3 = 1$. Now let us assume that we have a tree with k OTUs (and by our induction hypothesis, $T_k = \frac{(2k-5)!}{2^{k-3}(k-3)!}$ distinct topologies and $L_k = 2k - 3$ branches). If we want to create a tree with $k + 1$ OTUs, we simply add an OTU to one of the L_k branches on one of the T_k trees. Therefore, the number of distinct topologies for a tree with $k + 1$ OTUs is

$$T_k \cdot L_k = \frac{(2k-5)!}{2^{k-3}(k-3)!} \cdot 2k - 3 = ... = \frac{(2(k+1)-5)!}{2^{(k+1)-3}((k+1)-3)!}$$

The table below shows the growth of D_n, L_n, and T_n as n grows.

Description	**Symbol**	**Pattern of growth**	**Solution**
# species	n	$2, 3, 4, 5, ..., n$	$n + 1$
# distances	D_n	$1, 3, 6, 10, ..., \binom{n}{2}$	$D_n + n$
# branches	L_n	$1, 3, 5, 7, ..., 2n - 3$	$L_n + 2$
# topologies	T_n	$1, 1, 3, ..., 1 \cdot 3 \cdot 5 \cdot ... \cdot (2n - 5)$	$T_n \cdot L_n$

2.

Let q represent the fraction of identical sites in a pairwise sequence alignment, and let d represent the number of substitutions per site. These values are related by the following simple equation.

$$q = 1 - d$$

However, we want to correct for unobserved substitutions. We would like to find a function $f(q) : [0, 1] \rightarrow \mathbb{R}$ such that $f(q)$ is close to 0 as q approaches 1, but is greater than $1 - q$ as q approaches 0.

Let us proceed by defining λ as the substitution (or mutation) rate for the sequences, and Δt as some small time interval. The probability that there are no substitutions at a given site in time Δt is

$$P_0(\Delta t) = 1 - \lambda \Delta t$$

If we assume independence of disjoint time intervals, we can write the following probability for any arbitrary time interval.

$$P_0(t + \Delta t) = P_0(t)P_0(\Delta t) = P_0(t)[1 - \lambda \Delta t] = P_0(t) - P_0(t)\lambda \Delta t$$

$$P_0(t + \Delta t) - P_0(t) = -P_0(t)\lambda \Delta t$$

$$\frac{P_0(t + \Delta t) - P_0(t)}{\Delta t} = -\lambda P_0(t)$$

As Δt becomes arbitrarily small, this gives us the differential equation

$$P'(t) = -\lambda P_0(t)$$

with solution

$$P_0(t) = e^{-\lambda t}$$

Therefore, if $d = \lambda \Delta t$ and $q = e^{\lambda \Delta t}$, then the approximating function we desire can be written as follows.

$$d \approx f(q) = -ln(q)$$

3.

First we show that calculating $G_{i,j}$ is a constant time operation. There are two possible ways to have a gap ending with an arbitrary $G_{i,j}$: it is either the extension of a previously opened gap, or the opening of a new gap. For these two cases, the gap can be introduced/extended in either of the two sequences. Therefore we can rewrite the recurrence for $G_{i,j}$ as follows.

$$G_{ij} = \max \begin{cases} G_{i-1,j} - \beta \\ G_{i,j-1} - \beta \\ S_{i-1,j} + w(1) \\ S_{i,j-1} + w(1) \end{cases}$$

So calculating $G_{i,j}$ requires 4 operations. Calculating $D_{i,j}$ is also constant time, requiring only a single operation.

Recall the given recurrence for $S_{i,j} = \max\{D_{i,j}, G_{i,j}\}$. Therefore, for two sequences of lengths m and n, calculating S_{mn} will require $O(5mn) \in O(mn)$ operations.