1. Suppose we have a population 5 made up of type A and type B individuals. Let i be the number of type A individuals. Suppose in each time step the probability of the number of A's increasing by 1 is $\frac{1}{3}$, and the probability of the number decreasing by 1 is $\frac{2}{3}$. In addition, assume if we ever reach a state with 0 A's or 5 A's then no further changes happen. If we start with 1 A and 4 B's what is the probability we end up with 5 A's? If we start with 4 A's and 1 B, what is the probability we end up with 0 A's?

$$\gamma_i = \frac{\beta_i}{\alpha_i} = \frac{\frac{2}{3}}{\frac{1}{3}} = 2$$

so

$$\rho_A = \frac{1}{1 + \sum_{k=1}^{N-1} \prod_{j=1}^{k} \gamma_j} = \frac{1}{1 + 2 + 2^2 + 2^3 + 2^4} = \frac{1}{31}$$

Similarly for B

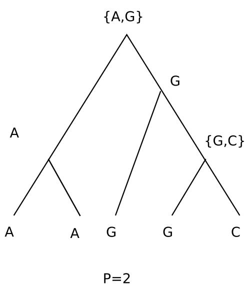
$$\gamma_i = \frac{\beta_i}{\alpha_i} = \frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}$$

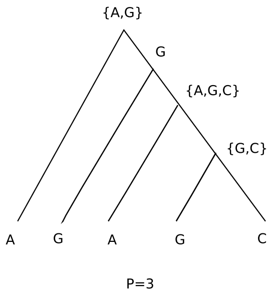
so

$$\rho_A = \frac{1}{1 + \sum_{k=1}^{N-1} \prod_{j=1}^{k} \gamma_j} = \frac{1}{1 + \frac{1}{2} + \frac{1}{2}^2 + \frac{1}{2}^3 + \frac{1}{2}^4} = \frac{16}{31}$$

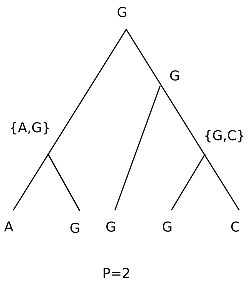
2. In the maximum parsimony method we build a phylogenetic tree for a number of different taxa by comparing bases at each site of the aligned sequences. If a base at a particular site is the same for all sequences then this site will not help us determine a tree since we would assume there has bee no mutation there. In addition, if all the sequences have the same base at a particular site (say A) except for up to three sequences which each have a different one of the other three bases (G, C and T), then this site does no provide information to distinguish trees. This is because by placing the most common base (A in this example) at each internal vertex we can assure that all trees have the minimum number of mutations. For example, suppose we have 5 sequences, S1,S2,S3,S4, S5, which have the following bases at a particular sites A, A, G, C, and T, respectively. Then by placing A at each internal vertex of any tree we guarantee that all trees have same parsimony score of 3, which is the minimum number required to get the observed bases. This means that in the maximum parsimony method we only need to consider sites where at least two different bases occur at least twice. These sites are called informative sites. For example, with 5 sequences, the following site with bases A,A,G,G,C is informative, but the site with bases A,G,G,G,C is not. Draw a 2 or 3 different trees with 5 taxa and convince yourself that the first site can have trees with different parsimony numbers, but that the parsimony number for the second site will always be 2.

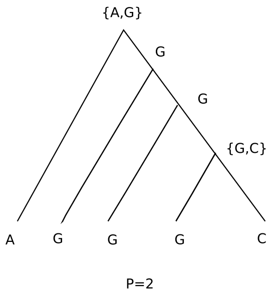
Here are some examples of trees based on the first site.





Here are some examples of trees based on the second site.





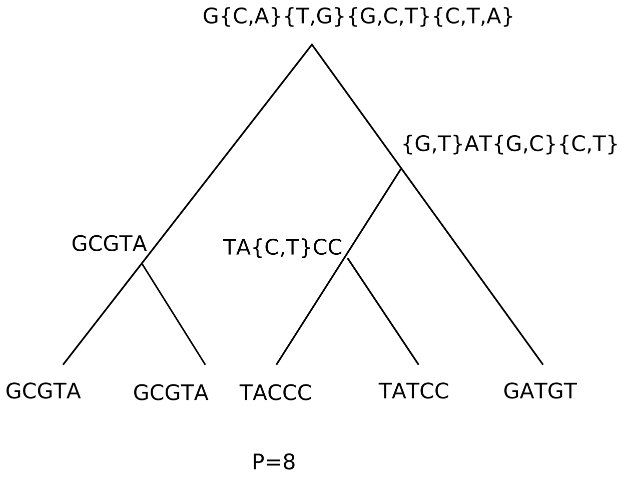
3. Given the five aligned sequences corresponding to five different taxa

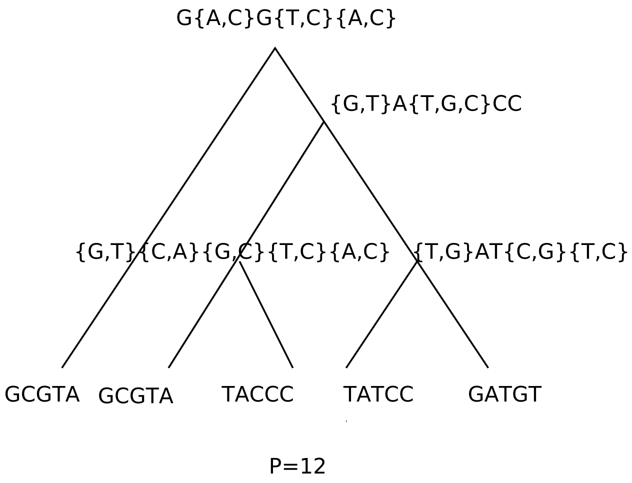
$$\begin{aligned} S1: &\quad AGCTGACGTTAACCG \\ S2: &\quad GGCTGAACTTAACTG \\ S3: &\quad ATATCAGCCGCACCG \\ S4: &\quad ATATTAGCCTCACCG \\ S5: &\quad AGATTAGTGTTACCG \end{aligned}$$

(a) Circle all the informative sites
The informative sites are blue.

(b) Of the informative sites, how many give distinct information about the parsimony of a tree? (ie how many distinct patterns are there?)
The last two informative sites give the same information – that is they will contribute the same amount to the parsimony number because they have the same pattern, These are the site with TTCCG and AACCT. To save time we could find the score for the one of these and multiply by two.

(c) Using only informative sites determine which of the trees below is most parsimonious.





(e) Explain why no tree could have a parsimony number less than 8.

At the first informative site there are two different bases (G,T), so at least one mutation is required. The same is true at the second informative site. The last three informative sites have three bases each, so at least 2 mutations are required for each. The minimum number of mutations is $1 + 1 + 2 + 2 + 2 = 8$.

4. Use the average distance method (UPGMA) to determine the phylogenetic tree for the five taxa listed in the phylogenetic distance table below.

	A	B	C	D	E
A		0.4	0.2	0.7	0.3
B			0.6	0.4	0.4
C				0.5	0.5
D					0.3

The closest two are A and C separated by 0.2, so we paired them and recalculate the table as follows:

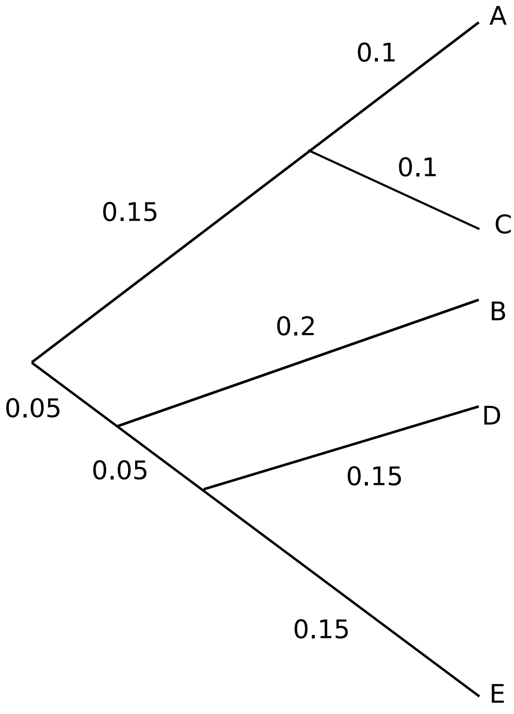
	B	D	E	AC
B		0.4	0.4	(0.4 + 0.6)/2=0.5
D			0.3	(0.5+0.7)/2 =0.6
E				(0.3 + 0.5)/2 =0.4

The next closest two are E and D, separated by 0.3, so we paired them and recalculate the table as follows:

	B	DE	AC
B		(0.4+0.4)/2 = 0.4	0.5
DE			(0.6+0.4)/2 =0.5

Finally B and DE are next closest, separated by 0.4 so we find $d(BDE, AC) = (0.5+0.5)/2 = 0.5$.

Putting this all together we get the following metric tree:


A
0.1
0.1
C
0.15
0.2
B
0.05
D
0.05
0.15
0.15
E