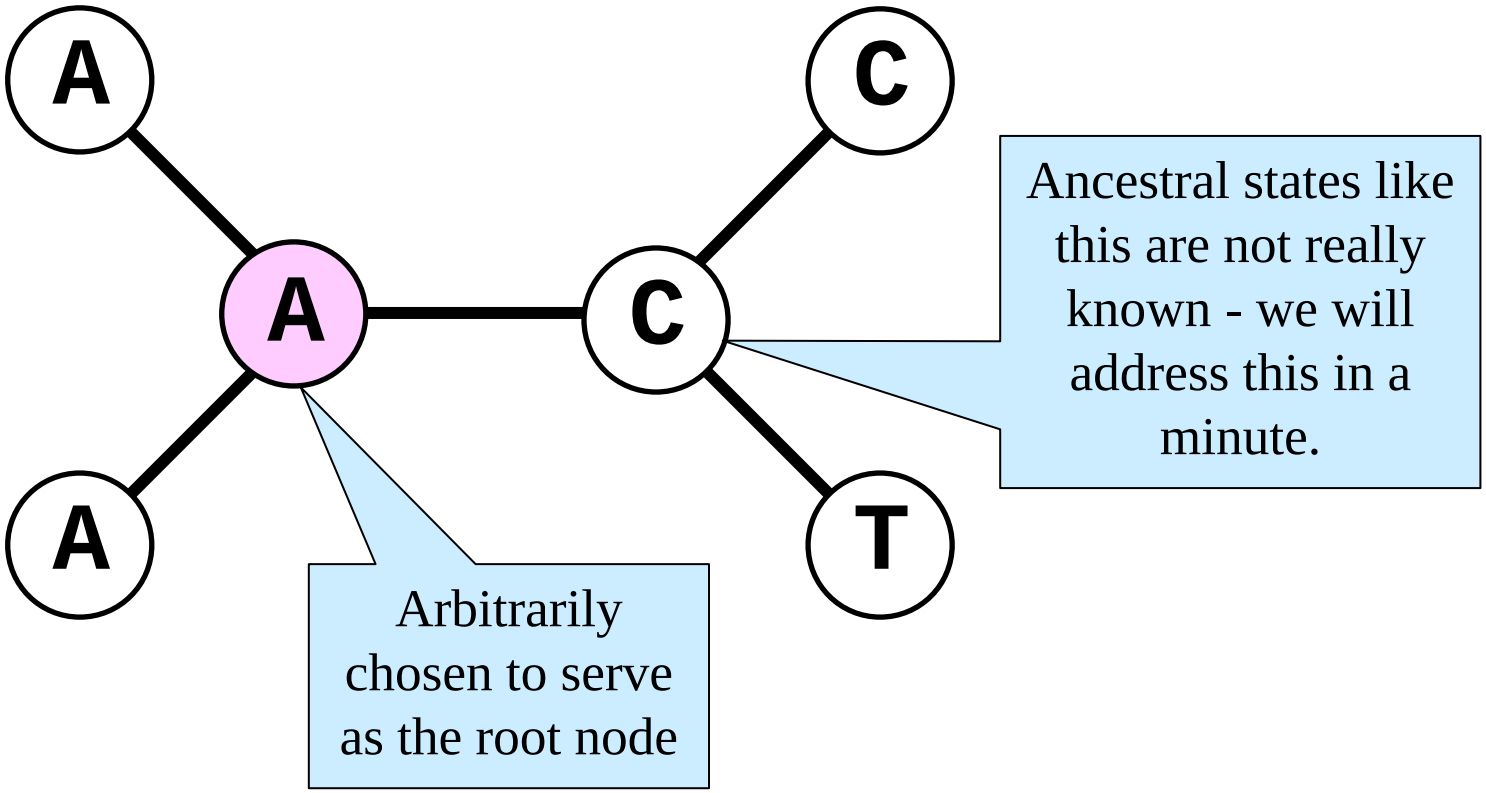
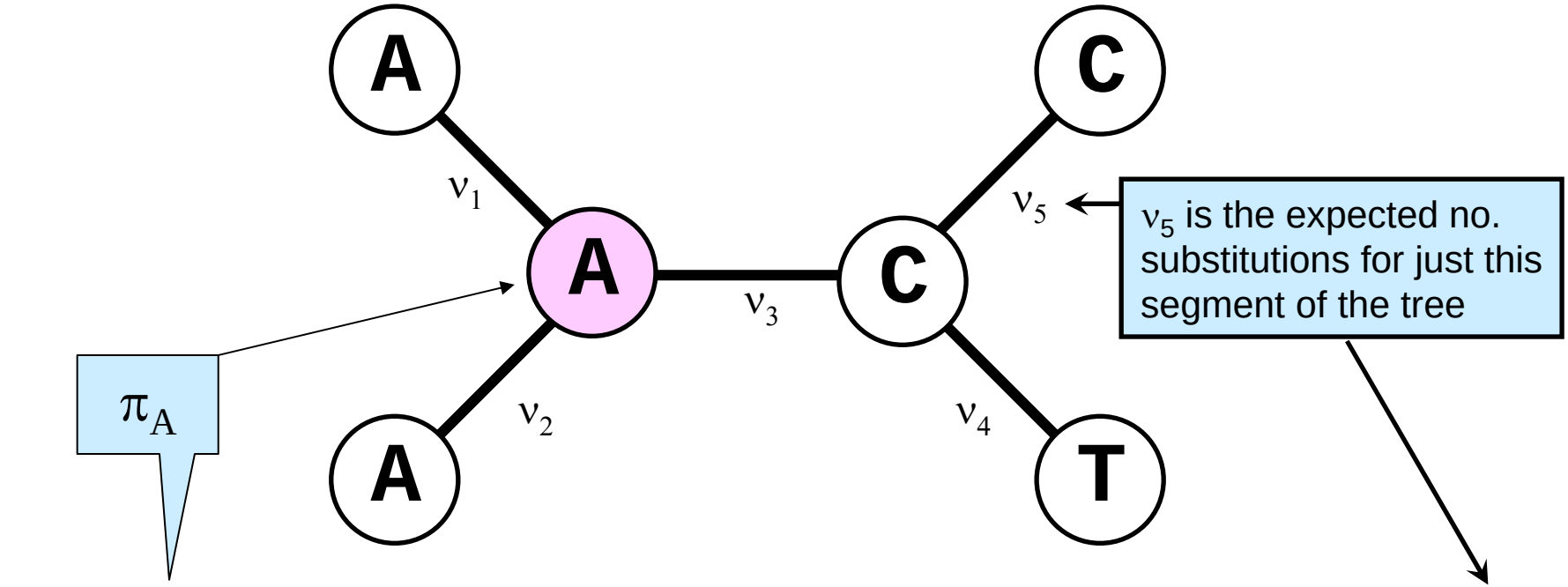
Likelihood of a tree

(data for only one site shown)



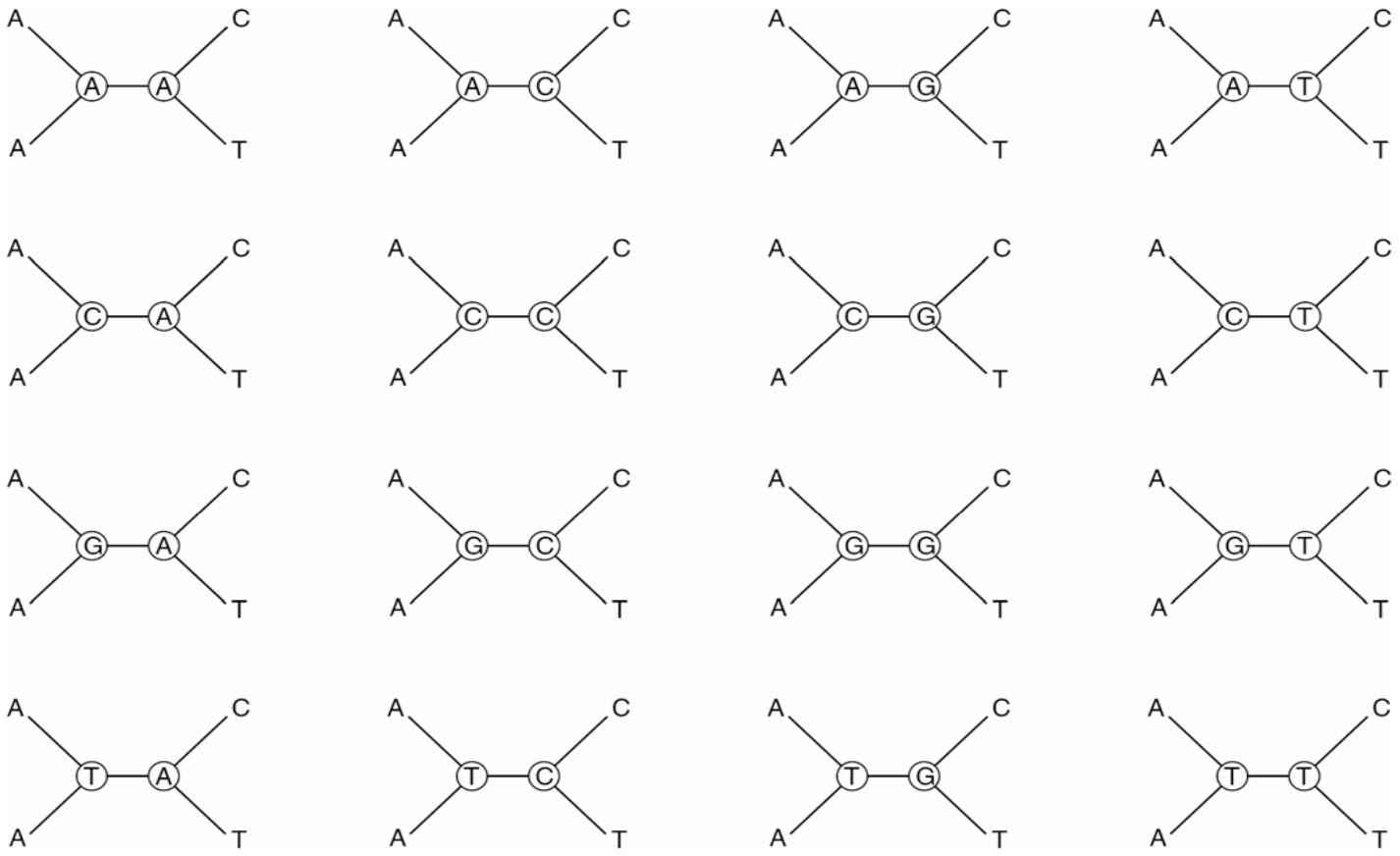
Likelihood for site k



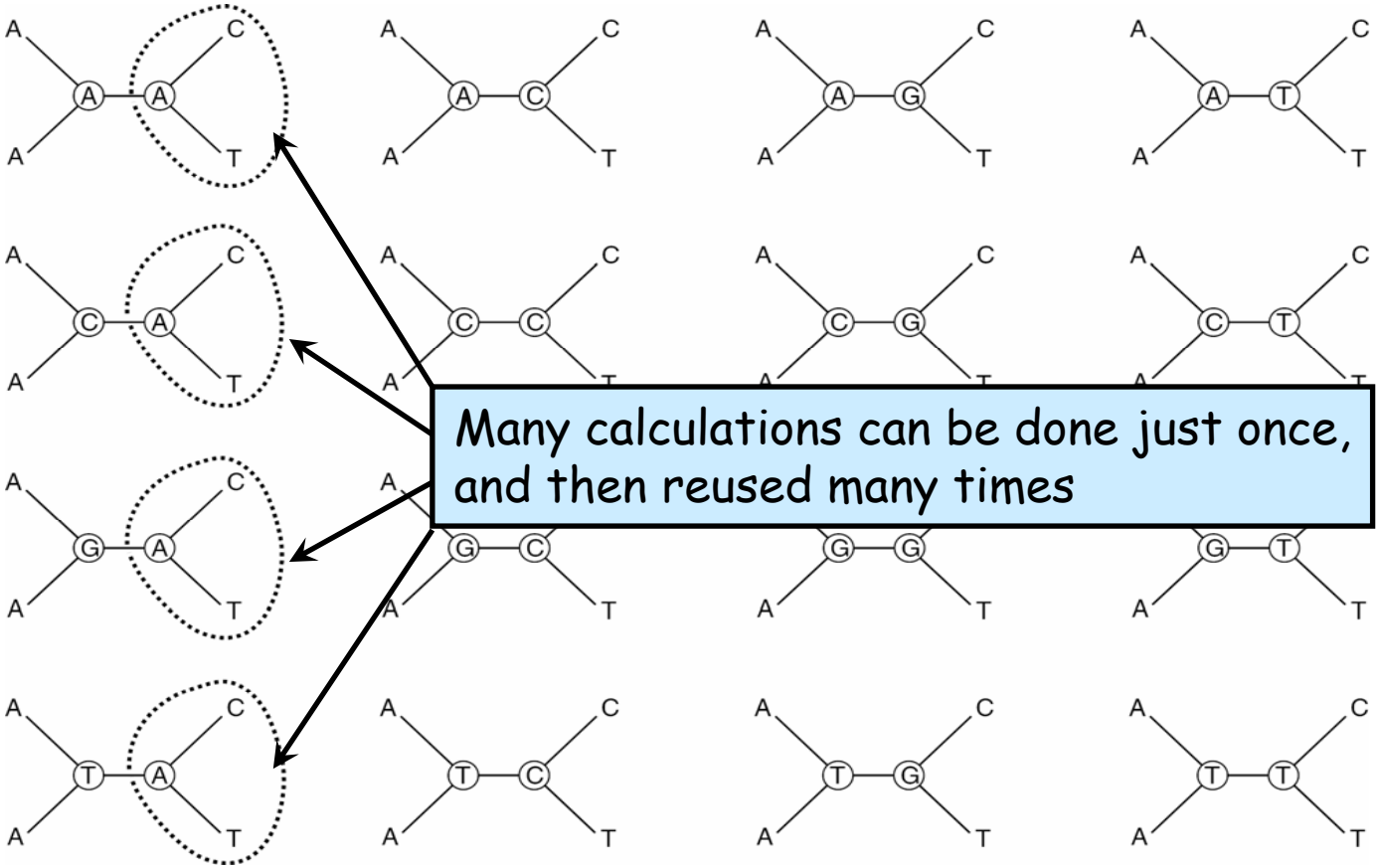
$$L_k = \tfrac{1}{4}\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4\nu_1/3}\right]\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4\nu_2/3}\right]\left[\tfrac{1}{4}-\tfrac{1}{4}e^{-4\nu_3/3}\right]\left[\tfrac{1}{4}-\tfrac{1}{4}e^{-4\nu_4/3}\right]\left[\tfrac{1}{4}+\tfrac{3}{4}e^{-4\nu_5/3}\right]$$

$P_{AA}(\nu_1)$ $P_{AA}(\nu_2)$ $P_{AC}(\nu_3)$ $P_{CT}(\nu_4)$ $P_{CC}(\nu_5)$

Brute force approach would be to calculate L_k for all 16 combinations of ancestral states and sum

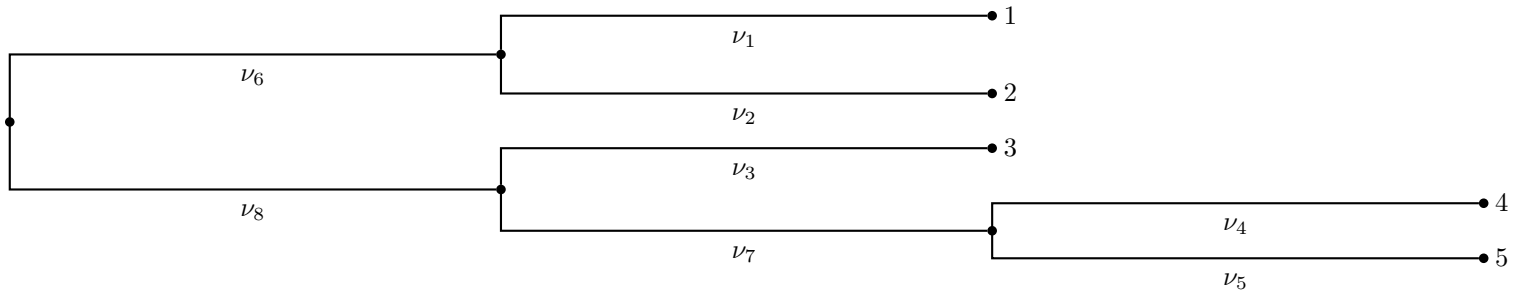


Pruning algorithm*
(same result, much less time)

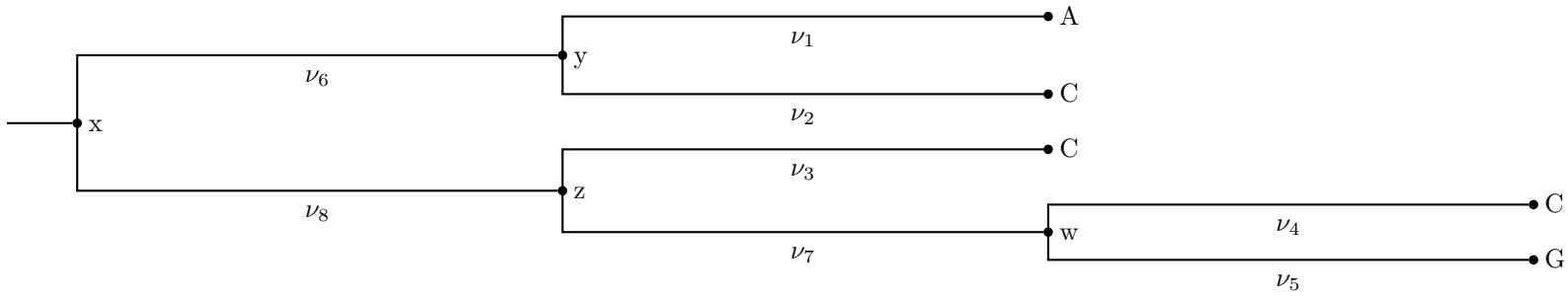


*The pruning algorithm was introduced by: Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**:368-376

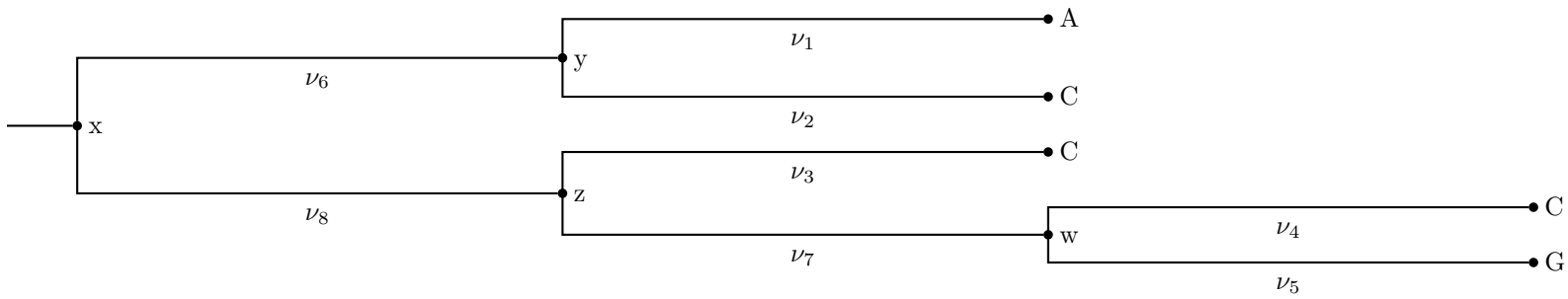
Taxon	Character
1	A
2	C
3	C
4	C
5	G



$$L = \sum_x \sum_y \sum_z \sum_w \Pr(x, y, z, w, A, C, C, C, G|\boldsymbol{\nu})$$

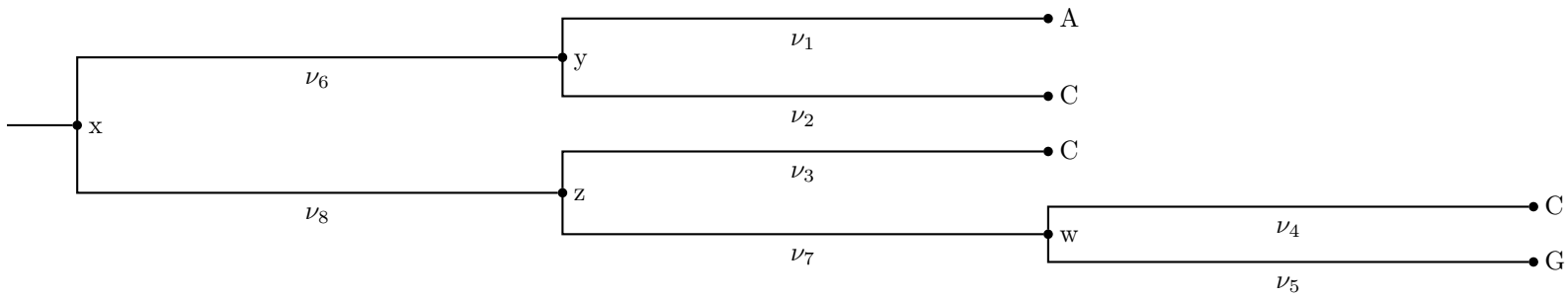


$$L = \sum_x \sum_y \sum_z \sum_w \Pr(x) \Pr(y|x,\nu_6) \Pr(A|y,\nu_1) \Pr(C|y,\nu_2) \cdots$$
$$\Pr(z|x,\nu_8) \Pr(C|z,\nu_3) \Pr(w|z,\nu_7) \Pr(C|w,\nu_4) \Pr(G|w,\nu_5)$$



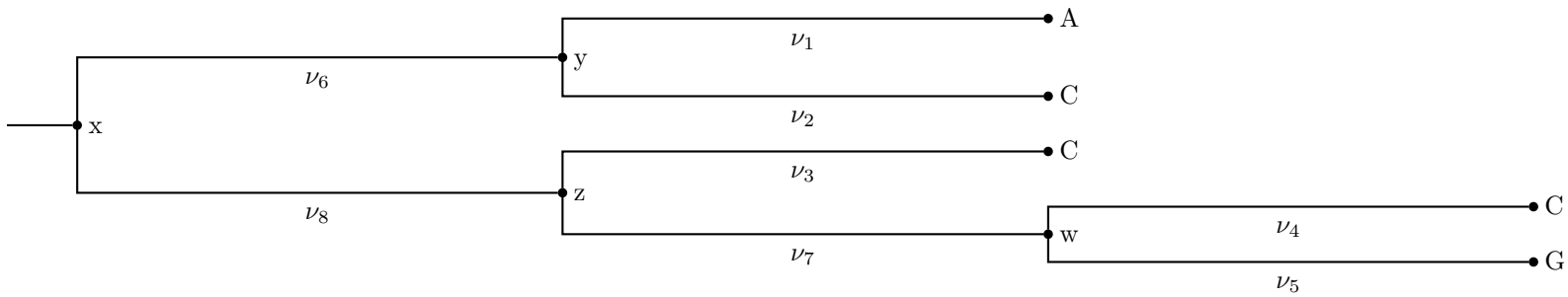
$$L = \sum_x \sum_y \sum_z \Pr(x)\Pr(y|x,\nu_6)\Pr(A|y,\nu_1)\Pr(C|y,\nu_2)\cdots$$

$$\Pr(z|x,\nu_8)\Pr(C|z,\nu_3)\left(\sum_w \Pr(w|z,\nu_7)\Pr(C|w,\nu_4)\Pr(G|w,\nu_5)\right)$$



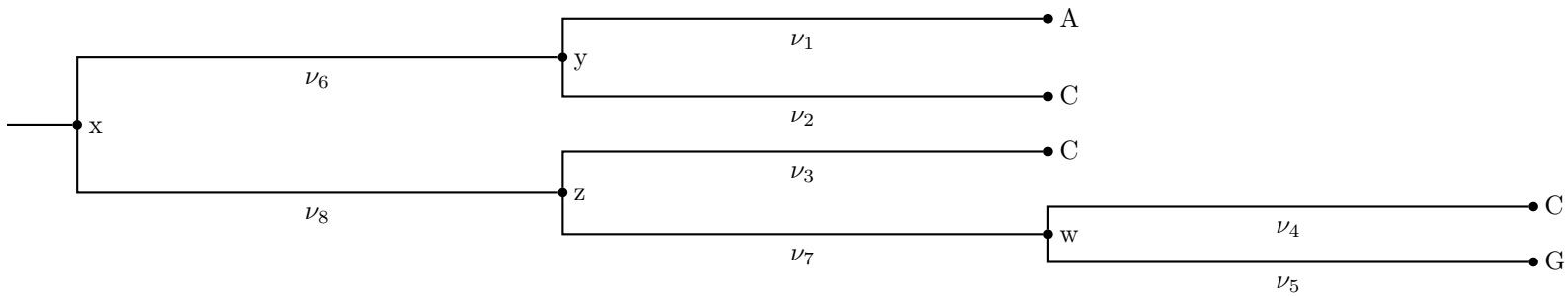
$$L = \sum_x \sum_y \Pr(x) \Pr(y|x, \nu_6) \Pr(A|y, \nu_1) \Pr(C|y, \nu_2) \cdots$$

$$\left(\sum_z \Pr(z|x, \nu_8) \Pr(C|z, \nu_3) \left(\sum_w \Pr(w|z, \nu_7) \Pr(C|w, \nu_4) \Pr(G|w, \nu_5) \right) \right)$$



$$L = \sum_x \Pr(x) \Big(\sum_y \Pr(y|x,\nu_6) \Pr(A|y,\nu_1) \Pr(C|y,\nu_2) \Big) \cdots$$

$$\Big(\sum_z \Pr(z|x,\nu_8) \Pr(C|z,\nu_3) \Big(\sum_w \Pr(w|z,\nu_7) \Pr(C|w,\nu_4) \Pr(G|w,\nu_5) \Big) \Big)$$



Maximum likelihood is a lot of work

- Site likelihoods involve products of transition probabilities, summed over ancestral states
- Overall log-likelihood for a tree is sum of site log-likelihoods
- Overall log-likelihood must be maximized!
  - must find MLEs for all edge lengths and all model parameters
  - this involves computing the overall log-likelihood **many, many times** (try turning on **logiter** in PAUP to get a feel for how much work this involves)
- Maximized lnL can now be compared to maximized lnL from other trees

Is it worth it?

- Uses all information
  - Parsimony ignores constant and autapomorphic sites
  - Distance methods ignore information not captured in pairwise comparisons
- Model generality
  - Some models possible with distance methods, but some quantities cannot be estimated reliably (e.g. variation in rates across sites)
  - Many parsimony variants exist, but parsimony does not allow estimation of the step matrix entries, for example
  - Many complex models are only possible under likelihood or Bayesian methods (which have a likelihood foundation)

References

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120.

Green Plant rbc*L*

First 88 amino acids, translation is for *Zea mays*

```
M--S--P--Q--T--E--T--K--A--S--V--G--F--K--A--G--V--K--D--Y--K--L--T--Y--Y--T--P--E--Y--E--T--K--D--T--D--I--L--A--A--F--R--V--T--P--
Chara           (green alga; land plant lineage)   AAAGATTACAGATTAACTTACTATACTCCTGAGTATAAAACTAAAGATACTGACATTTTAGCTGCATTTCGTGTAACTCCA
Chlorella       (green alga)                       .....C...C.T...................T..CC..C.A.....C.....T...C.T..A..G..C...A.G.....T
Volvox          (green alga)                       ........TC.T.....A.....C..A.....C...GT.GTA.....C........C.....A.........A.G......
Conocephalum    (liverwort)                        ........TC.........T........G..T...G............G..T........A......A.AA.G.....T
Bazzania        (moss)                             ........T........C..T.....G.....A...G.G..C.....G..A..T.....G..A.........A.G.....C
Anthoceros      (hornwort)                         ........T........CC.T.....C.....T..CG.G..C..G........T.....G..A..G.C.T.AA.G.....T
Osmunda         (fern)                             ........TC....G...C..........C..T...G.G..C..G........T.....G..A.....C..AA.G.....C
Lycopodium      (club "moss")                      .GG..............C.T..C........T.....G..C.....A..C..T...C.G..A........AA.G.....T
Ginkgo          (gymnosperm; Ginkgo biloba)        ..............G.....T..........A...C....C...........T..C..G..A.....C..A........T
Picea           (gymnosperm; spruce)               ...................T..........A...C.G..C........G..T.....G..A.....C..A........T
Iris            (flowering plant)                  ..............G.....T..........T..CG....C...........T..C..G..A.....C..A........T
Asplenium       (fern; spleenwort)                 ........TC..C.G.....T..C..C..C..A..C..G..C........C..T..C..G..A..T..C..GA.G..C...
Nicotiana       (flowering plant; tobacco)         .....G....A...G.....T..............CC....C..G........T..A..G..A.....C..A........T
```

```
Q--L--G--V--P--P--E--E--A--G--A--A--V--A--A--E--S--S--T--G--T--W--T--T--V--W--T--D--G--L--T--S--L--D--R--Y--K--G--R--C--Y--H--I--E--
CAACCTGGCGTTCCACCTGAAGAAGCAGGGGCTGCAGTAGCTGCAGAATCTTCTACTGGTACATGGACTACTGTTTGGACTGACGGATTAACTAGTTTGGACCGATACAAAGGAAGATGCTACGATATTGAA
.....A..T........A........G..T..G........A........A..A........T.....G.....A........T..T...........A.....T........TC.T..T..T..C..C..G
.....A..T..............TGT..T.....T..T.....T.....A..A..A.....T.....A.....A........T..T.....A...C.T.....T........TC.T..T..T..C..C..G
..G.....G..A...G.A...........A..A.....T.....T.........................A...........T..TC.T....ACC.T..T..T..T.....TC.......T.G......C
.....G..A..A.................A..G...........T........A..C.....G.....C..G........C..T..GC.T..A...C.C..T..T........TC.......T..C..C...
T....A..G..G.................A..C...........T........A.........................C..T...C.T..C..CC.T.....T........TC..........C......
.....C..A..A..GG....G.....T..A..............G...........A.....G.....C.....A.....G..T...C.T..C...C.T..T..T..T..G..TC................
....T...A..A.....C..G.....G..A..C...........T........C.........................C..T...C.T..C...C.C..T..C........TC.G.....T..A......
........A..G........G.....G..A..............C........C..............C...........C..T...C.T..C...C.T..T..T.....G...........T..C..C..G
.....A..G..G..C..G.....G..A..A...........T........C..C...........C...........C..T...C.T......C.T..T..T.....G..GC.......T..C..C..G
.....C..A.....TG..........G.....C..G........C.....................A..A..G........T...C.T..C...C.T..T..T.........C........C.C..C..G
.....C..A..A...G..........C..A.................G..C.....A...........C.....G.....A.....G..G..C..CC.T.....T.....G..CC.............C..G
........A.....................C..G........C.....................A.....A.....C..T...C.T..C..CC.T..T..T........GC........CGC..C..G
```

All four bases are observed at some sites...

...while at other sites, only one base is observed

Question: Why is rate heterogeneity ubiquituous?

Answer: Differences in mutational rates and (mainly) **selective constraint**

- Many sites are under purifying (stabilizing) selection:
  - Any mutation results in a different amino acid, AND
  - A amino acid replacement at the site results in dramatically worse functioning of the protein.
  - These sites will show *low* rates of evolution on a tree.
- Other sites are less constrained.
  - A mutation results in the same amino acid, OR
  - Many amino acids will work equally well at that position in the protein.
  - These sites will show *high* rates of evolution on a tree.

Rate heterogeneity in protein-coding genes: terms

- **Synonymous** mutations result in the same amino acid.
- **Non-synonymous** mutations result in the different amino acid.
- **Conservative** changes are non-synonymous changes that result in a chemically similar amino acid.
- **Neutral** mutations result in a new genotype that has the same fitness as the genotypes currently fixed in the population.

Rate heterogeneity in protein-coding genes: generalities

- Synonymous changes are often neutral (or close to neutral),
- Third base positions and untranslated regions (introns and other non-coding regions) tend to have high rates because changes to these sites lead to synonymous changes.
- Transitions tend to lead to more synonymous or conservative changes.
- Amino acid residues that are embedded, involved in salt bonding, or part of the active site tend to be more constrained.
- Loops of amino acid residues on the outside of proteins often tolerate a wide range of substitutions (or even indels).

Site-specific rates

- You decide there are 3 classes of sites:
  - 1st positions evolve at relative rate r_1
  - 2nd positions evolve at relative rate r_2
  - 3rd positions evolve at relative rate r_3
- r_1, r_2 and r_3 are *relative* rates, not *actual* rates:
  - their average is 1.0: if each category has the same number of sites, $(r_1 + r_2 + r_3)/3 = 1.0$
  - the actual rates are $r_1 \alpha$ (for 1st positions), $r_2 \alpha$ (for 2nd positions) and $r_3 \alpha$ (for 3rd positions)
  - note that the average substitution rate over all sites is α $(r_1 \alpha + r_2 \alpha + r_3 \alpha)/3 = \alpha\ (1.0) = \alpha$
- Assuming k rate classes adds k-1 parameters to the model

Transition probabilities under the JC69 model

with no rate heterogeneity:

$$\begin{aligned} \Pr(i \to i|\nu) &= \frac{1}{4} + \frac{3}{4}\, e^{\frac{-4\nu}{3}} \\ \Pr(i \to j|\nu) &= \frac{1}{4} - \frac{1}{4}\, e^{\frac{-4\nu}{3}} \end{aligned}$$

Transition probabilities under the JC69 model

First base positions under a *site-specific rates* model:

$$\begin{aligned}
\Pr(i \to i|\nu) &= \frac{1}{4} + \frac{3}{4}\, e^{\frac{-4r_1\nu}{3}} \\
\Pr(i \to j|\nu) &= \frac{1}{4} - \frac{1}{4}\, e^{\frac{-4r_1\nu}{3}}
\end{aligned}$$

Site-specific rates in PAUP*

First, define a character partition that puts each site into one of several mutually exclusive categories (the category names are arbitrary):

```
charpartition codons = one:1-.\3, two:2-.\3, three:3-.\3;
```

Then tell PAUP* that you want site specific rates and provide the partition you defined previously:

```
lset rates=sitespec siterates=partition:codons;
```

Pinvar approach

- Unlike the site-specific rates approach, this approach does not require you to assign sites to rate categories
- Assumes there are only two classes of sites:
  - invariable sites (evolve at relative rate 0)
  - variable sites (evolves at relative rate *r*)
- Remarks:
  - mean of relative rates = $(p_{invar})(0) + (1-p_{invar})(r) = 1$
  - this means that $r = 1/(1-p_{invar})$
  - if all sites are variable, $p_{invar} = 0$ and $r = 1$