# Lecture 26. Phylogeny methods, part 6 (Modeling rate variation among sites)

Joe Felsenstein

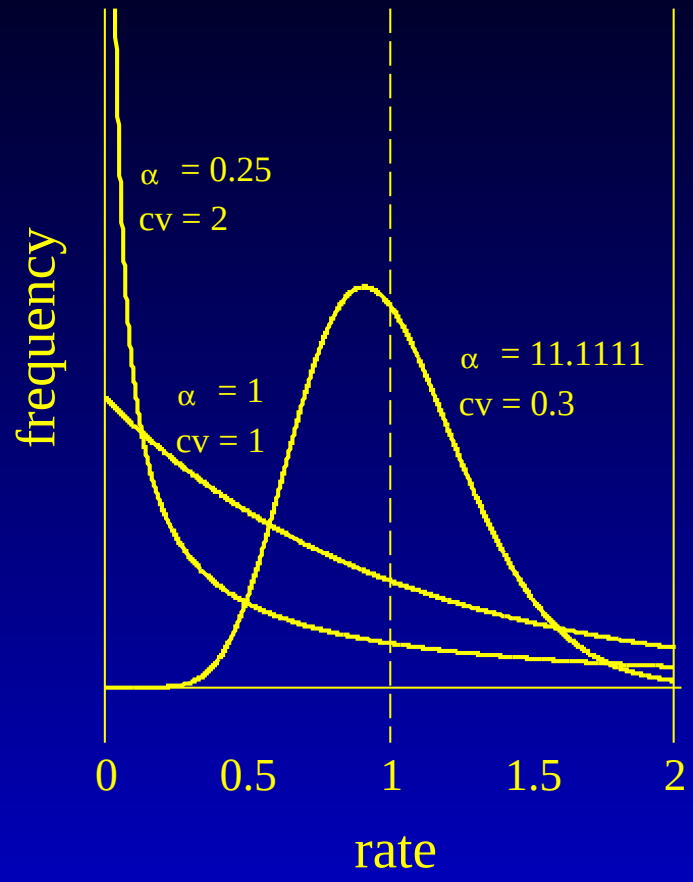
Department of Genome Sciences and Department of Biology

# A model of variation in evolutionary rates among sites

The basic idea is that the rate at each site is drawn independently from a distribution of rates. The most widely used choice is the Gamma distribution, which has density function (if its mean is 1):

$$\mathbf{f(r)} = \frac{\alpha^{\alpha}\ \mathbf{r}^{\alpha-1}\ \mathbf{e}^{-\alpha\ \mathbf{r}}}{\mathbf{\Gamma(\alpha)}}$$

# Gamma distributions




Gamma distributions with mean 1 and different coefficients of variation (standard deviation / mean). $\alpha = 1/\mathrm{CV}^2$ is the "shape parameter" of the Gamma distribution

## Unrealistic aspects of the model:

- There is no reason, aside from mathematical convenience, to assume that the Gamma is the right distribution. A common variation is to assume there is a separate probability $f_0$ of having rate 0.
- Rates at different sites appear to be correlated, which this model does not allow.
- Rates are not constant throughout evolution – they change with time.

## Rates varying among sites

If $\mathbf{L^{(i)}(r_i)}$ is the likelihood of the tree for site i given that the rate of evolution at site i is $\mathbf{r_i}$, we can integrate this over a gamma density

$$\mathbf{L^{(i)}} = \int_0^\infty \mathbf{f(r_i;\alpha)L^{(i)}(r_i)\ dr_i}$$

so that the overall likelihood is

$$\mathbf{L} = \prod_{i=1}^{m} \left[ \int_0^\infty \mathbf{f(r_i;\alpha)L^{(i)}(r_i)\ dr_i} \right]$$

Unfortunately these integrals cannot be evaluated for trees with more than a few tips as the quantities $\mathbf{L^{(i)}(r_i)}$ complicated.

## Hidden Markov Models

These are the most widely used models allowing rate variation to be correlated along the sequence.

We assume:

- There are a finite number of rates, $\mathbf{m}$. Rate $i$ is $\mathbf{r_i}$.
- There are probabilities $\mathbf{p_i}$ of a site having rate $\mathbf{i}$.
- A process not visible to us ("hidden") assigns rates to sites. It is a Markov process working along the sequence. For example it might have transition probability $\mathrm{Prob}\ (\mathbf{j}|\mathbf{i})$ of changing to rate $\mathbf{j}$ in the next site, given that it is at rate $\mathbf{i}$ in this site.
- The probability of our seeing some data are to be obtained by summing over all possible combinations of rates, weighting approproately by their probabilities of occurrence.

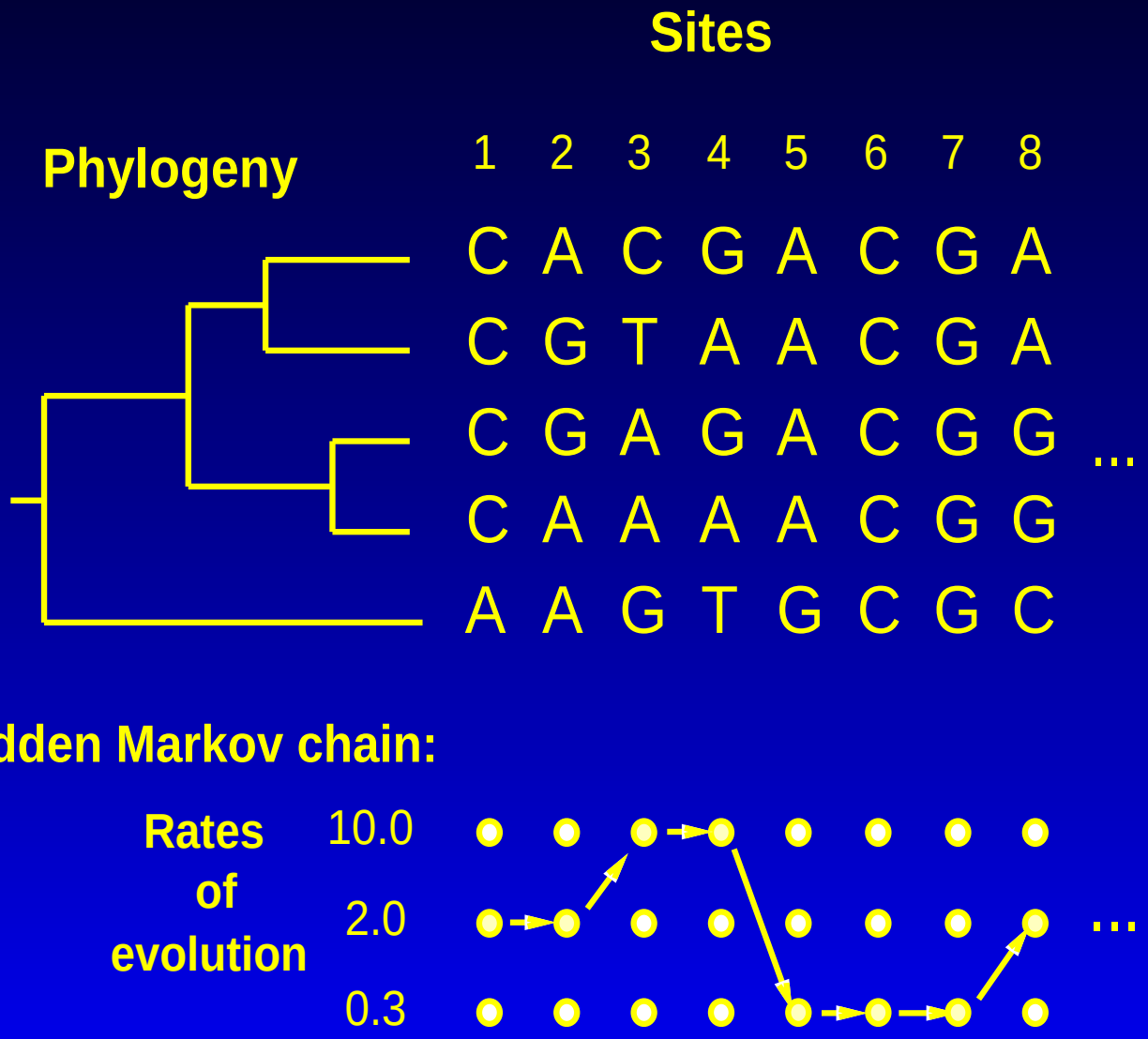
## Likelihood with a[n] HMM

Suppose that we have a way of calculating, for each possible rate at each possible site, the probability of the data at that site given that rate. This is

$$\text{Prob}\left(\mathbf{D^{(i)}} \mid \mathbf{r_j}\right)$$

To get the overall probability of all data, sum over all possible paths through the array of sites $\times$ rates, weighting each combination of rates by its probability:

# A Hidden Markov Model for rates in a phylogeny



## Hidden Markov Models

If there are a number of hidden rate states, with state $i$ having rate $r_i$

$$\text{Prob}\left(\mathbf{D} \mid \mathbf{T}\right) = \sum_{\mathbf{i_1}} \sum_{\mathbf{i_2}} \ldots \sum_{\mathbf{i_p}} \text{Prob}\left(\mathbf{r_{i_1}}, \mathbf{r_{i_2}}, \ldots \mathbf{r_{i_p}}\right)$$

$$\times \text{Prob}\left(\mathbf{D} \mid \mathbf{T}, \mathbf{r_{i_1}}, \mathbf{r_{i_2}}, \ldots \mathbf{r_{i_m}}\right)$$

Evolution is independent once each site has had its rate specified

$$\text{Prob}\left(\mathbf{D} \mid \mathbf{T}, \mathbf{r_1}, \mathbf{r_2}, \ldots, \mathbf{r_p}\right) =$$

$$\prod_{\mathbf{i=1}}^{\mathbf{p}} \text{Prob}\left(\mathbf{D^{(i)}} \mid \mathbf{T}, \mathbf{r_i}\right).$$

## Seems impossible ...

Evolution is independent once each site has had its rate specified

$$\text{Prob}\,(\mathbf{D}|\mathbf{T},\mathbf{r_1},\mathbf{r_2},\ldots,\mathbf{r_m}) \;=\; \prod_{\mathbf{i=1}}^{\mathbf{m}} \text{Prob}\,(\mathbf{D^{(i)}}|\mathbf{T},\mathbf{r_i}).$$

To compute the likelihood we sum over all ways rate states could be assigned to sites:

$$\begin{aligned}
\mathbf{L} &= \text{Prob}\,(\mathbf{D} \mid \mathbf{T}) \\
&= \sum_{\mathbf{i_1=1}}^{\mathbf{m}} \sum_{\mathbf{i_2=1}}^{\mathbf{m}} \ldots \sum_{\mathbf{i_p=1}}^{\mathbf{m}} \text{Prob}\,\left(\mathbf{r_{i_1}}, \mathbf{r_{i_2}}, \ldots, \mathbf{r_{i_p}}\right) \\
&\quad \times \text{Prob}\,\left(\mathbf{D^{(1)}} \mid \mathbf{r_{i_1}}\right) \text{Prob}\,\left(\mathbf{D^{(2)}} \mid \mathbf{r_{i_2}}\right) \ldots \text{Prob}\,\left(\mathbf{D^{(n)}} \mid \mathbf{r_{i_p}}\right)
\end{aligned}$$

Problem: The number of rate combinations is very large. With 100 sites and 3 rates at each, it is $\mathbf{3^{100} \simeq 5 \times 10^{47}}$. This makes the summation impractical.

# Factorization and the algorithm

Fortunately, the terms can be reordered:

$$
\begin{aligned}
\mathbf{L} &= \text{Prob}\left(\mathbf{D} \mid \mathbf{T}\right) \\
&= \sum_{\mathbf{i_1}=1}^{\mathbf{m}} \sum_{\mathbf{i_2}=1}^{\mathbf{m}} \cdots \sum_{\mathbf{i_p}=1}^{\mathbf{m}} \text{Prob}\left(\mathbf{i_1}\right) \text{Prob}\left(\mathbf{D^{(1)}} \mid \mathbf{r_{i_1}}\right) \\
&\quad \times \text{Prob}\left(\mathbf{i_2} \mid \mathbf{i_1}\right) \text{Prob}\left(\mathbf{D^{(2)}} \mid \mathbf{r_{i_2}}\right) \\
&\quad \times \text{Prob}\left(\mathbf{i_3} \mid \mathbf{i_2}\right) \text{Prob}\left(\mathbf{D^{(3)}} \mid \mathbf{r_{i_3}}\right) \\
&\quad \vdots \\
&\quad \times \text{Prob}\left(\mathbf{i_p} \mid \mathbf{i_{p-1}}\right) \text{Prob}\left(\mathbf{D^{(p)}} \mid \mathbf{r_{i_p}}\right)
\end{aligned}
$$

## Using Horner's Rule

and the summations can be moved each as far rightwards as it can go:

$$
\begin{aligned}
\mathbf{L} \quad = \quad & \sum_{\mathbf{i_1}=1}^{\mathbf{m}} \text{Prob}\left(\mathbf{i_1}\right) \text{Prob}\ \left(\mathbf{D^{(1)}} \mid \mathbf{r_{i_1}}\right) \\
& \sum_{\mathbf{i_2}=1}^{\mathbf{m}} \text{Prob}\left(\mathbf{i_2} \mid \mathbf{i_1}\right) \text{Prob}\ \left(\mathbf{D^{(2)}} \mid \mathbf{r_{i_2}}\right) \\
& \sum_{\mathbf{i_3}=1}^{\mathbf{m}} \text{Prob}\left(\mathbf{i_3} \mid \mathbf{i_2}\right) \text{Prob}\ \left(\mathbf{D^{(3)}} \mid \mathbf{r_{i_3}}\right) \\
& \quad \vdots \\
& \sum_{\mathbf{i_p}=1}^{\mathbf{m}} \text{Prob}\left(\mathbf{i_p} \mid \mathbf{i_{p-1}}\right) \text{Prob}\ \left(\mathbf{D^{(p)}} \mid \mathbf{r_{i_p}}\right)
\end{aligned}
$$

## Recursive calculation of HMM likelihoods

The summations can be evaluated innermost-outwards. The same summations appear in multiple terms. We can then evaluate them only once. A huge saving results. The result is this algorithm:
Define $\mathcal{P}_{\mathbf{i}}(\mathbf{j})$ as the probability of everything at or to the right of site $\mathbf{i}$, given that site $\mathbf{i}$ has the $\mathbf{j}$-th rate.
Now we can immediately see for the last site that for each possible rate category $\mathbf{i_p}$

$$\mathcal{P}_{\mathbf{p}}(\mathbf{i_p}) \;=\; \mathrm{Prob}\,\left(\mathbf{D}^{(\mathbf{p})} \mid \mathbf{r}_{\mathbf{i_p}}\right)$$

(as “at or to the right of" simply means “at" for that site).

## Recursive calculation

More generally, for site $\ell < \mathbf{p}$ and its rates $\mathbf{i}_\ell$

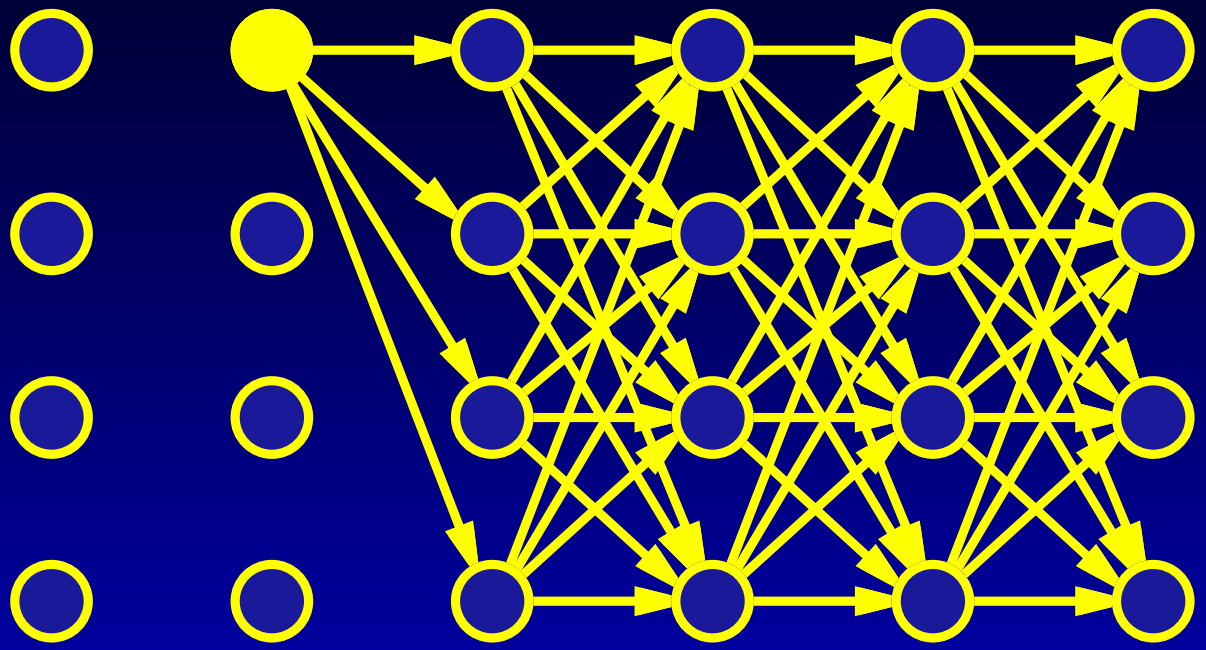
$$\mathcal{P}_\ell(\mathbf{i}_\ell) \;=\; \text{Prob}\left(\mathbf{D}^{(\ell)} \mid \mathbf{r}_{\mathbf{i}_\ell}\right) \sum_{\mathbf{i}_\ell=\mathbf{1}}^{\mathbf{m}} \text{Prob}\left(\mathbf{i}_{\ell+\mathbf{1}} \mid \mathbf{i}_\ell\right) \mathcal{P}_{\ell+\mathbf{1}}(\mathbf{i}_{\ell+\mathbf{1}})$$

We can compute the $\mathcal{P}$'s recursively using this, starting with the last site and moving leftwards down the sequence. Finally we have the $\mathcal{P}_\mathbf{1}(\mathbf{i}_\mathbf{1})$ for all $\mathbf{m}$ states. These are simply weighted by the equilibrium probabilities of the Markov chain of rate categories:

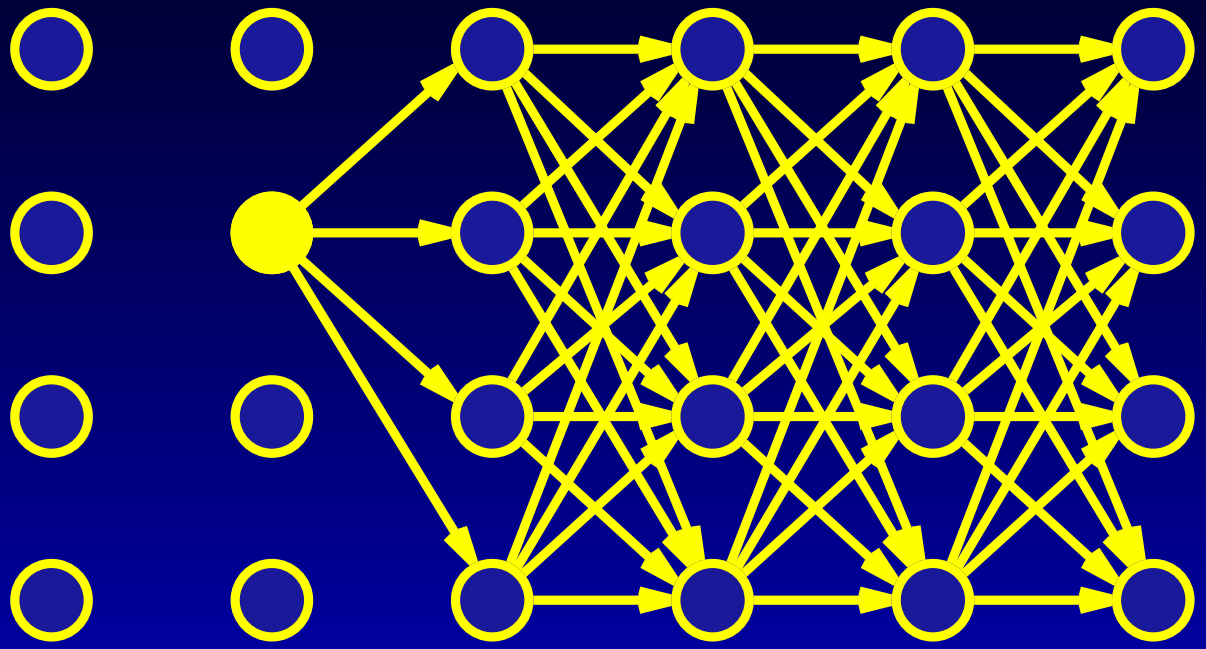
$$\mathbf{L} \;=\; \text{Prob}\left(\mathbf{D} \mid \mathbf{T}\right) \;=\; \sum_{\mathbf{i}_\mathbf{1}=\mathbf{1}}^{\mathbf{m}} \text{Prob}\left(\mathbf{i}_\mathbf{1}\right) \mathcal{P}_\mathbf{1}(\mathbf{i}_\mathbf{1})$$

An entirely similar calculation can be done from left to right, remembering that the transition probabilities $\text{Prob}\left(\mathbf{i_k}|\mathbf{i_{k+1}}\right)$ would be different in that case.

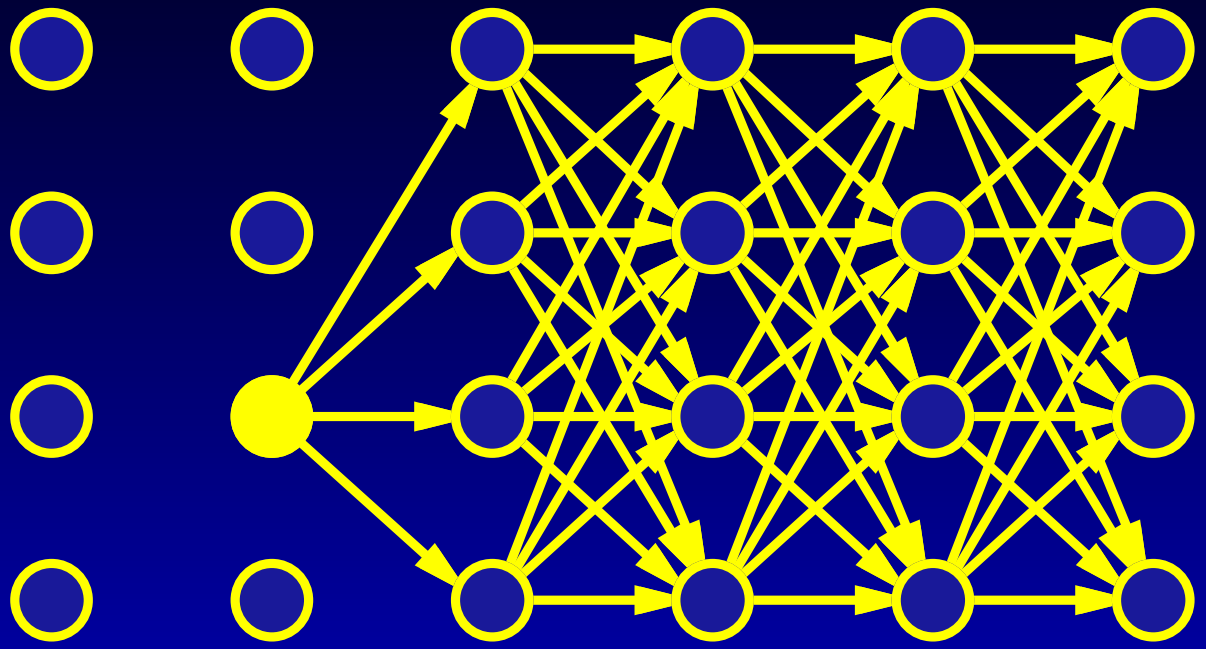
# Computing all possible paths from one node.



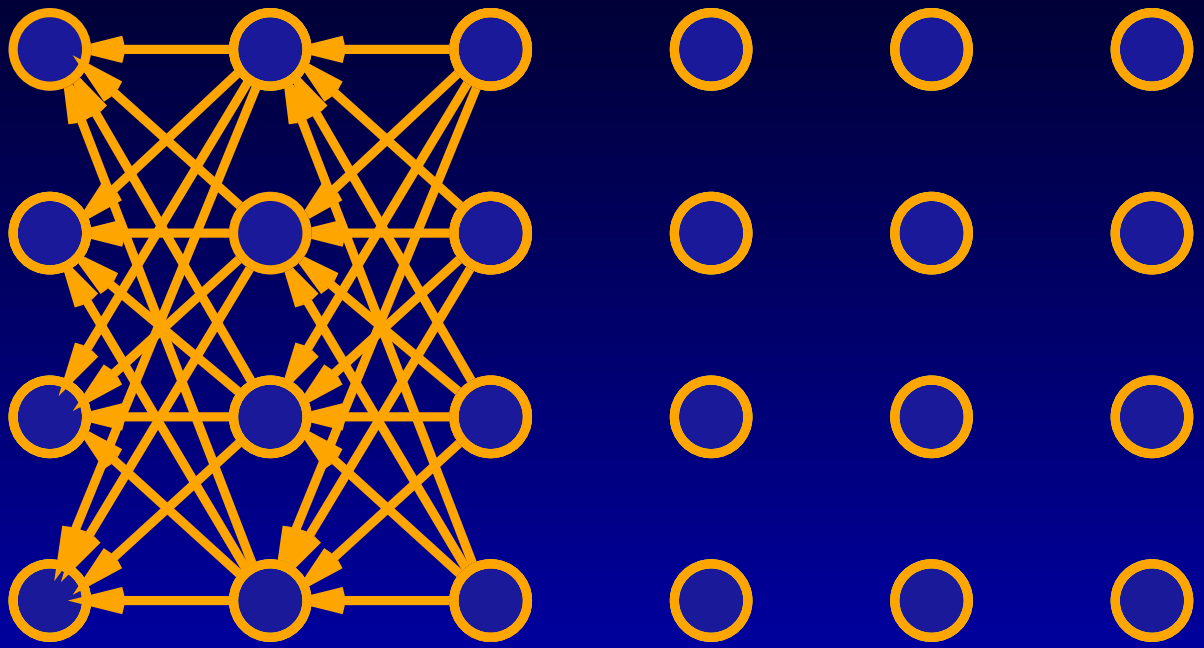
# Computing all possible paths starting at another node



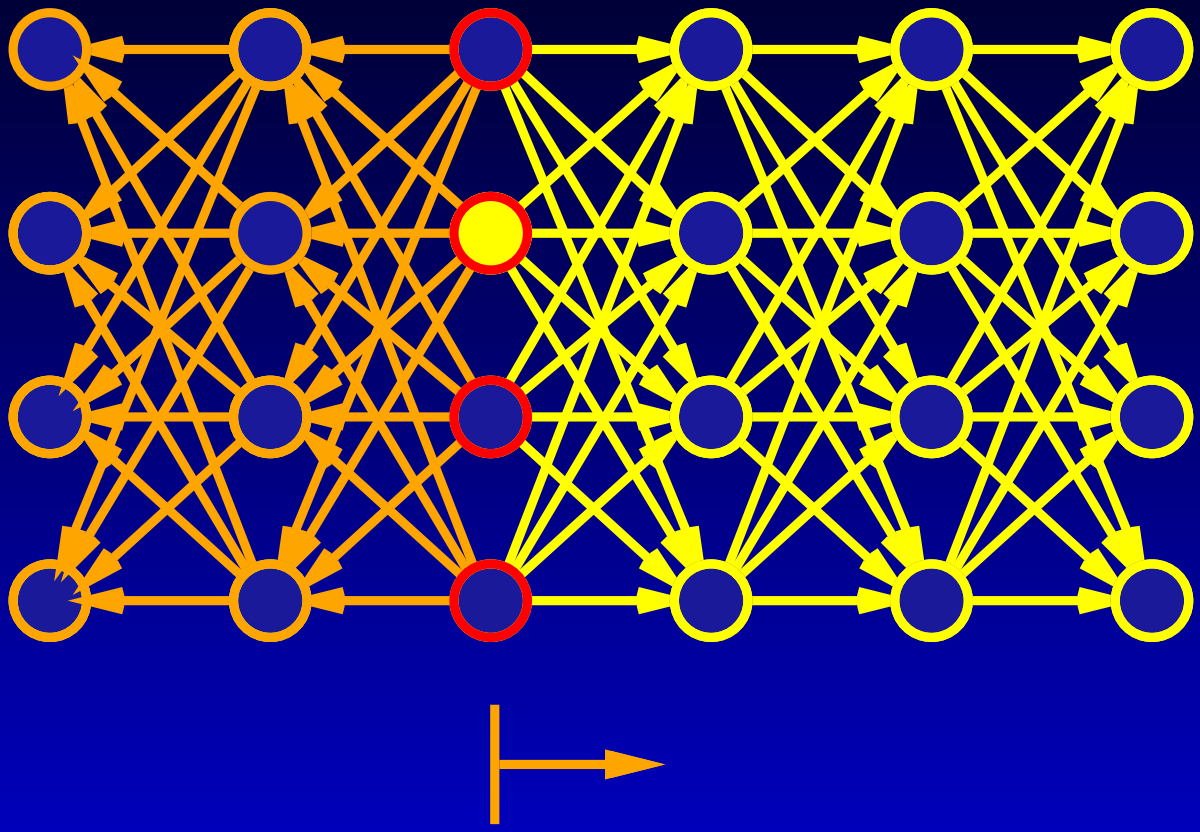
## Note that they re-use the same terms.



# The same algorithm can work forwards.



# The forwards-backwards algorithm.



## References


Churchill, G.A. 1989. Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology* **51:** 79-94. **[First paper using HMMs on sequences, without trees]**

Felsenstein, J. and G. A. Churchill. 1996. A hidden Markov model approach to variation among sites in rate of evolution *Molecular Biology and Evolution* **13:** 93-104. **[HMMs for rates in ML trees]**

Jin, L. and M. Nei. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. *Molecular Biology and Evolution* **7:** 82-102. **[Gamma distributed rates in a distance]**

Olsen, G. J. 1987. Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. *Cold Spring Harbor Symposia on Quantitative Biology* **52:** 825-837. **[Lognormal rates in a distance]**

## references, cont'd


Waddell, P. J. and M. A. Steel. 1997. General time-reversible distances with unequal rates across sites: mixing Γ and inverse Gaussian distributions with invariant sites. *Molecular Phylogenies and Evolution* **8:** 398-414. **[More generalized rate distributions in a more generalized distance]**

Yang, Z. 1994. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution* **10:** 1396-1401. **[Rates varying in gamma distribution in an ML tree method for few species]**

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* **39:** 306-314. **[First paper on HMMs for ML trees, used to approximate Gamma distributions for more species]**

Yang, Z. 1995. A space-time process model for the evolution of DNA sequences. *Genetics* **139:** 993-1005. **[Also allowing for correlated rates along the molecule]**

## How it was done

This projection produced as a PDF, not a PowerPoint file, and viewed using the Full Screen mode (in the View menu of Adobe Acrobat Reader):

- using the `prosper` style in LaTeX,
- using Latex to make a `.dvi` file,
- using `dvips` to turn this into a Postscript file,
- using `ps2pdf` to mill it into a PDF file, and
- displaying the slides in Adobe Acrobat Reader.

Result: nice slides using freeware.