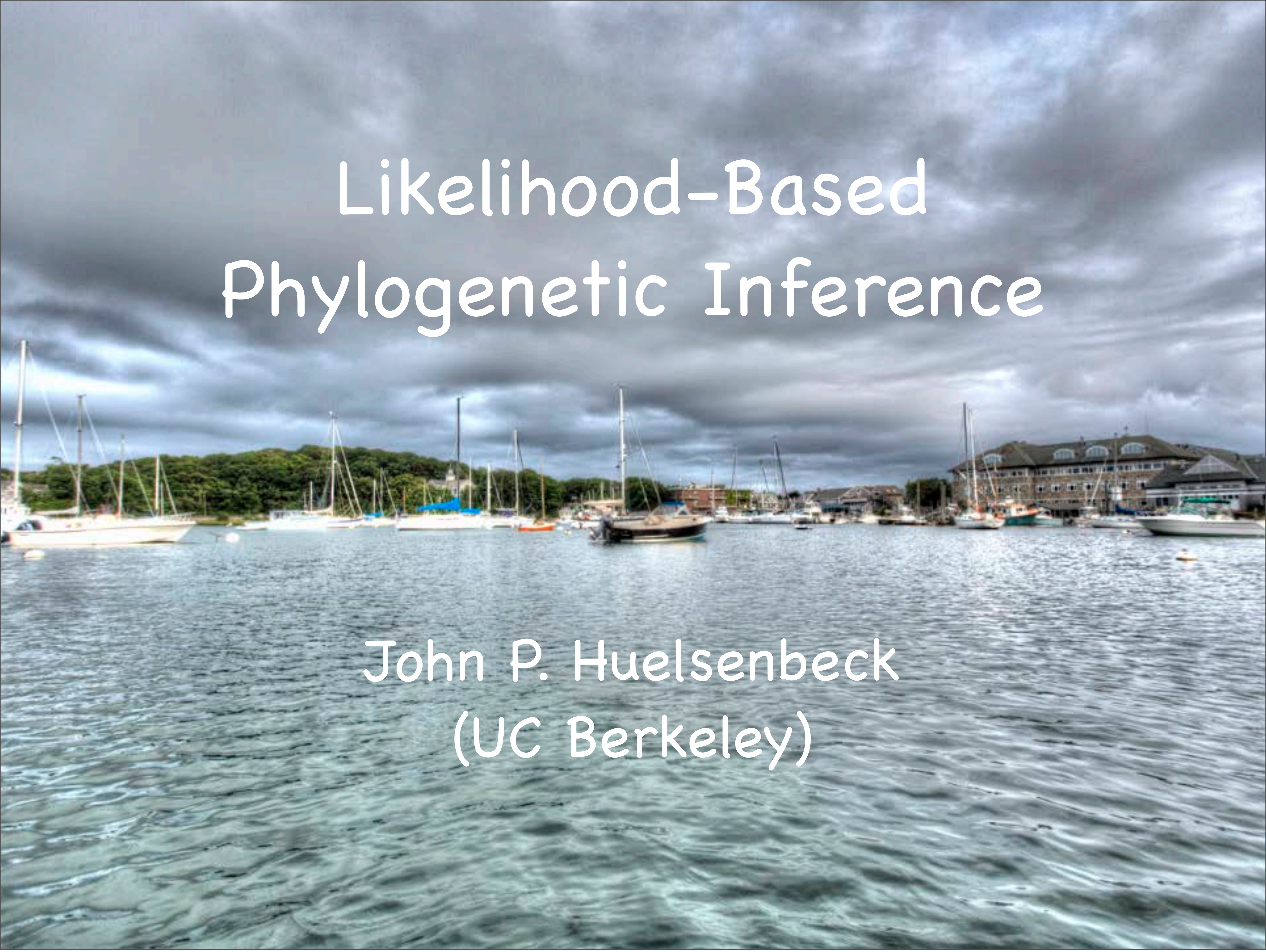


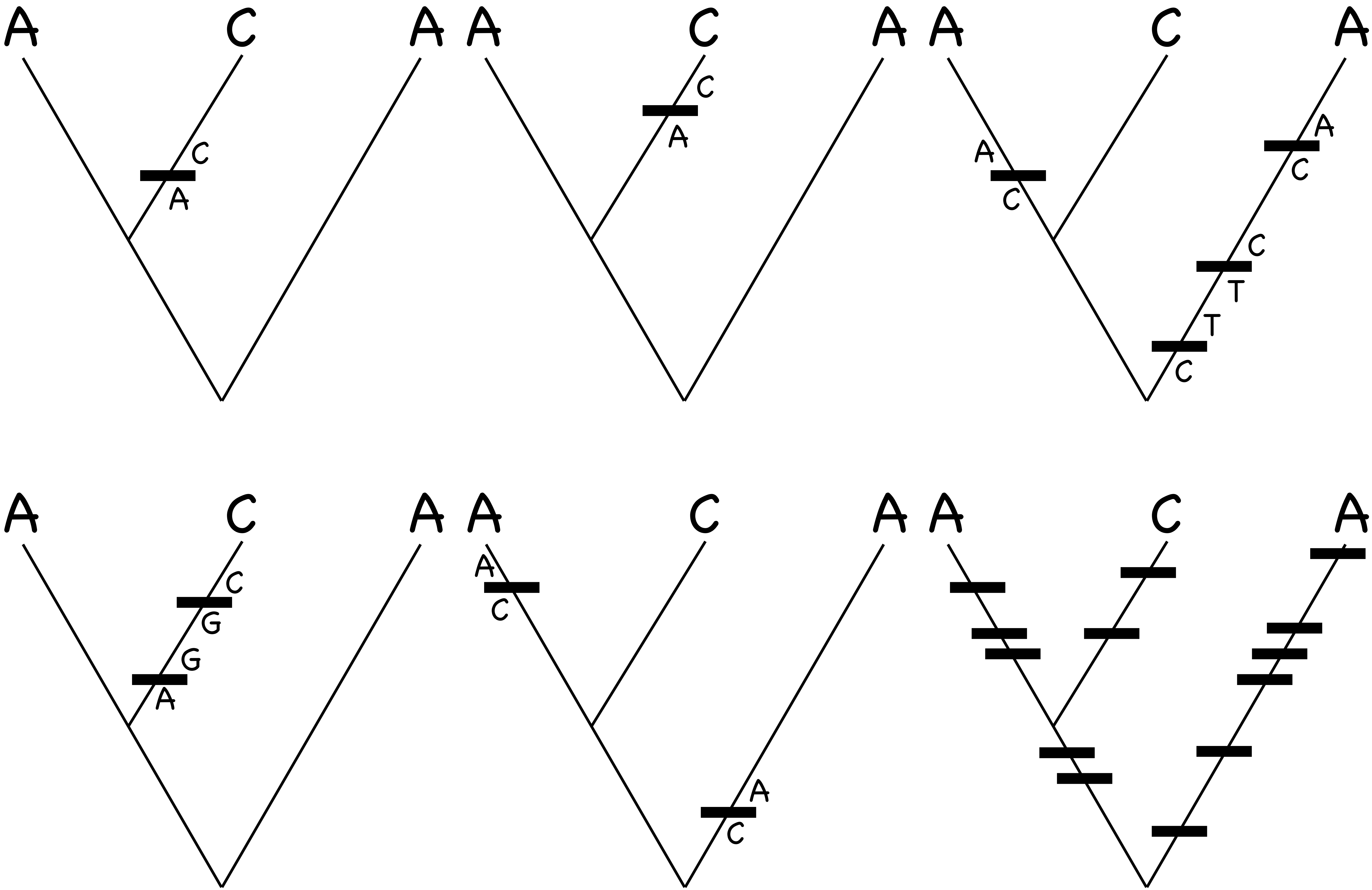
Likelihood-Based Phylogenetic Inference

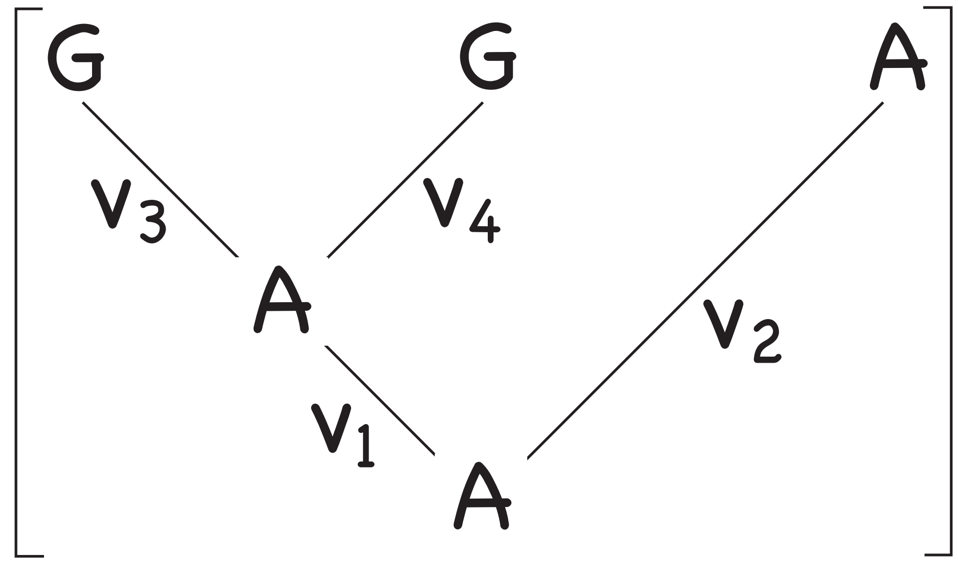
John P. Huelsenbeck
(UC Berkeley)

```
#NEXUS

begin data;
	dimensions ntax=5 nchar=895;
	format gap=- datatype=dna;
	matrix
	Human      AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGACTT.....AACCCAAACAACCCAGCTCTCCCTAAGCTT
	Chimpanzee AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTT.....AACCCAAACAACCCAGCTCTCCCTAAGCTT
	Gorilla    AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTT.....AACCCAAACAATTCAACTCTCCCTAAGCTT
	Orangutan  AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTC.....CACCCAGACACTACAACTCTCACTAAGCTT
	Gibbon     AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTA.....AACCCAAACGCTAGAACTCTCCCTAAGCTT
	;
end;
```

Some Possible Character Histories



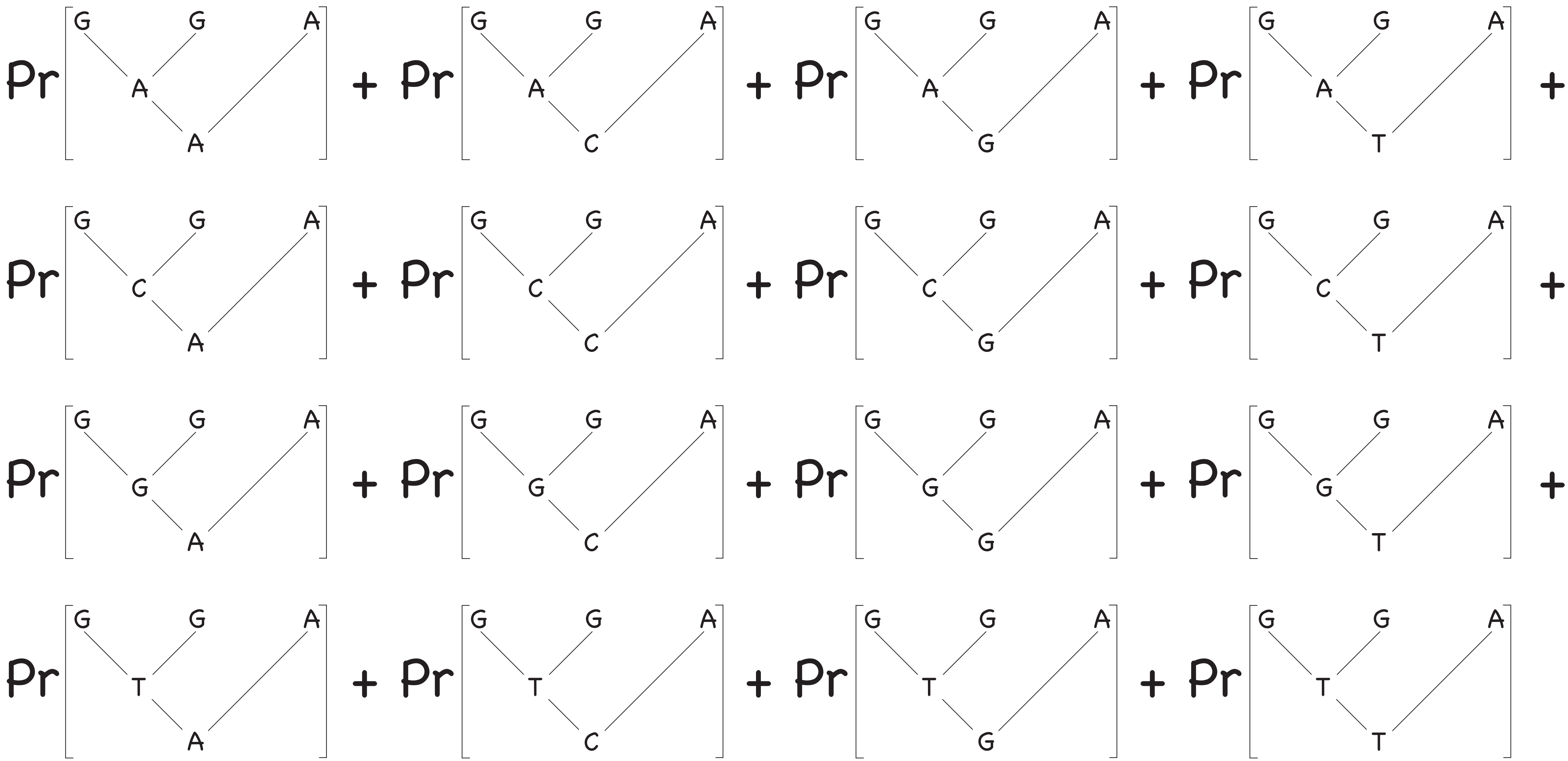
Pr 

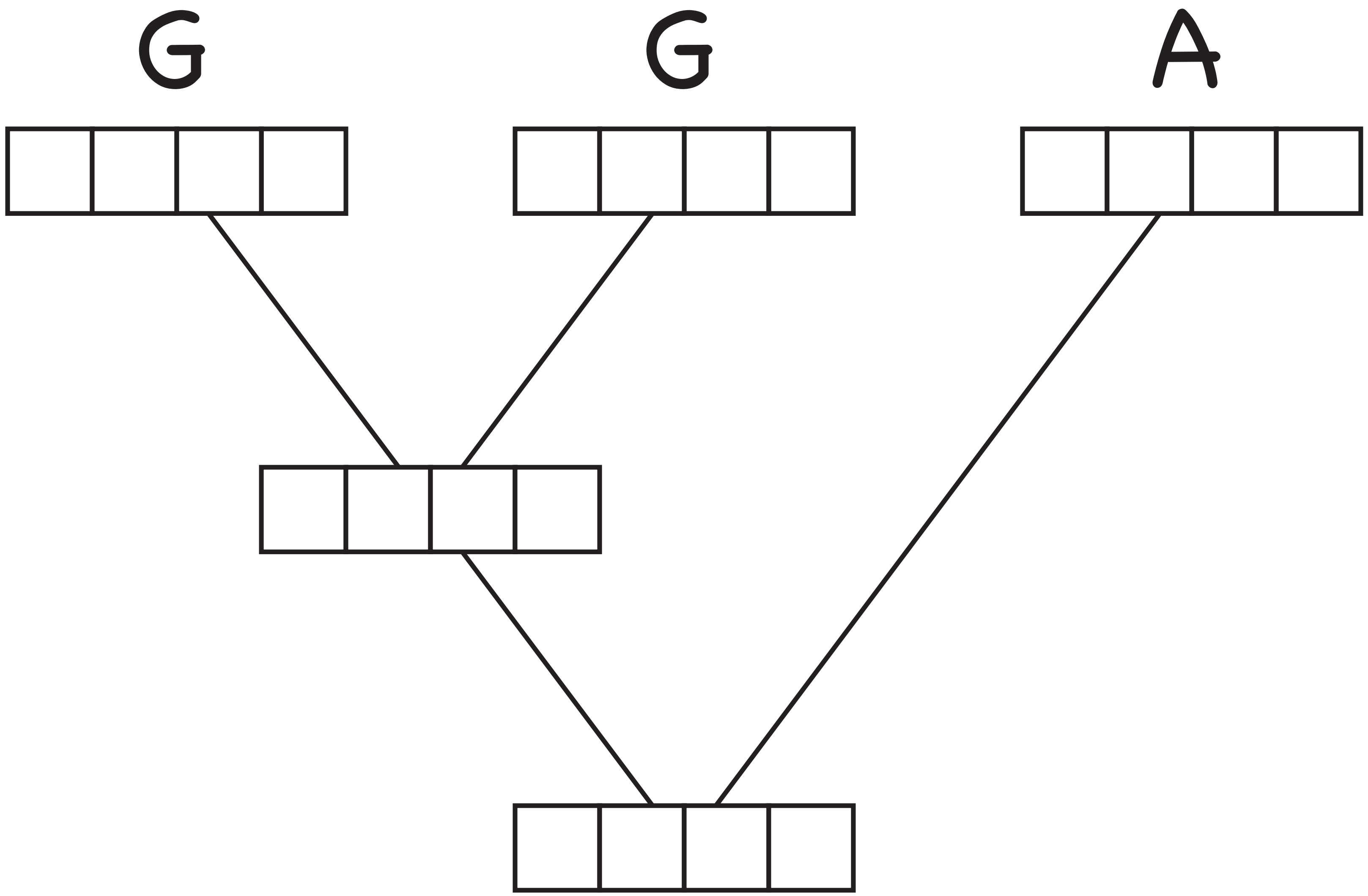
=

$$\pi_A \times p_{AA}(v_1) \times p_{AA}(v_2) \times p_{AG}(v_3) \times p_{AG}(v_4)$$

π_i – Stationary frequencies

$p_{ij}(v)$ – Transition probabilities

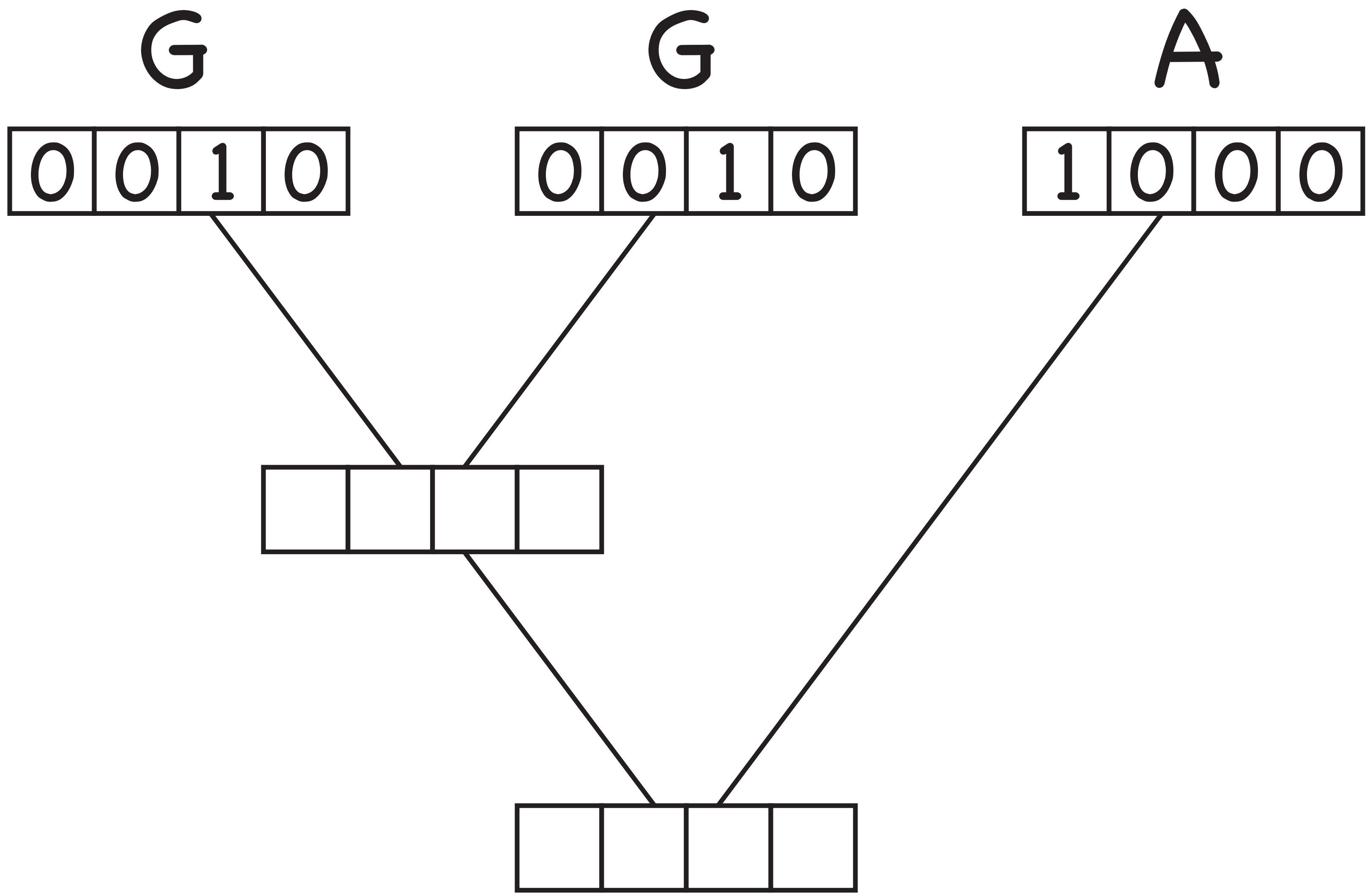

Pr G G A A A + Pr G G A A C + Pr G G A A G + Pr G G A A T +
Pr G G A C A + Pr G G A C C + Pr G G A C G + Pr G G A C T +
Pr G G A G A + Pr G G A G C + Pr G G A G G + Pr G G A G T +
Pr G G A T A + Pr G G A T C + Pr G G A T G + Pr G G A T T

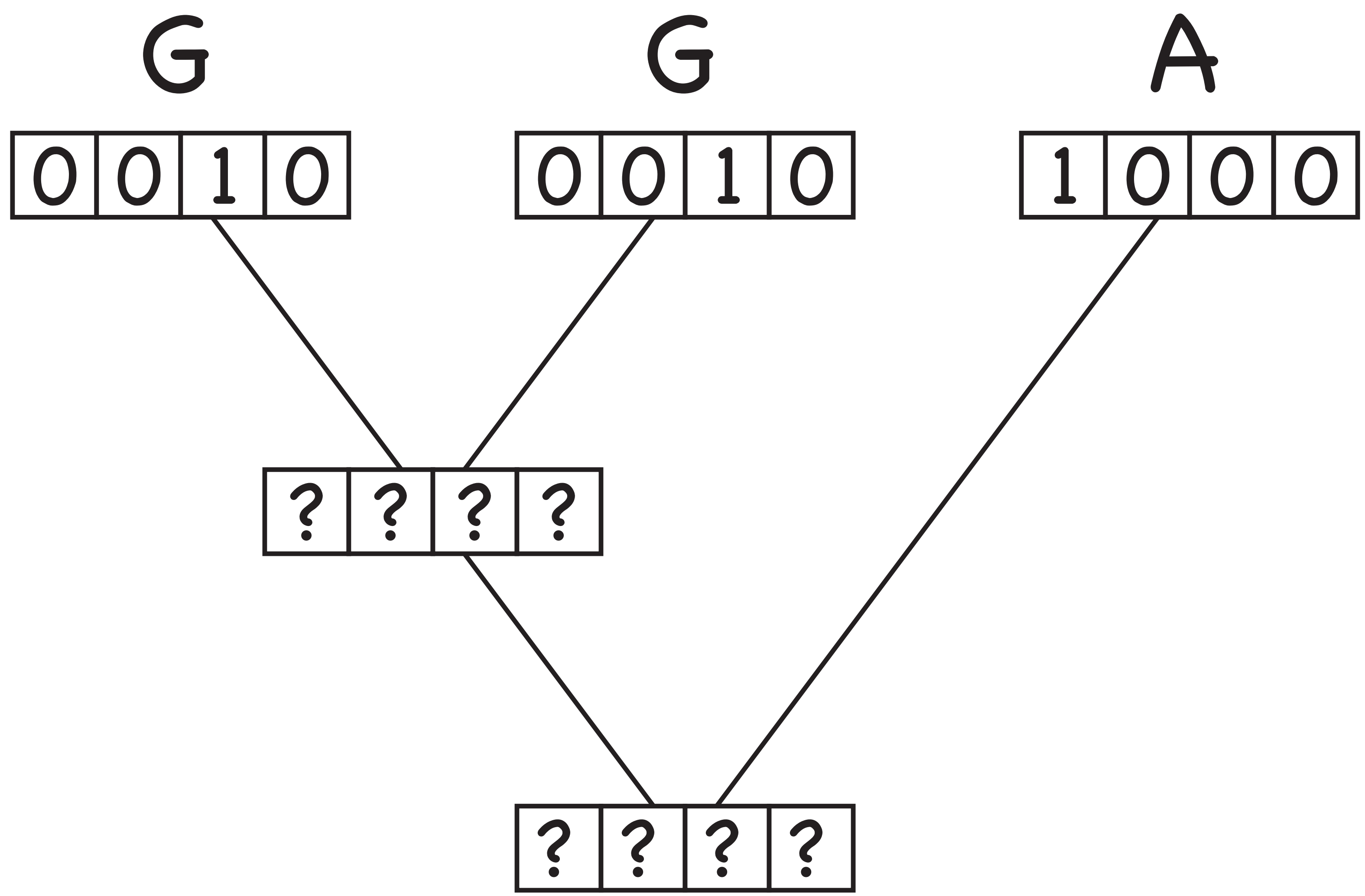


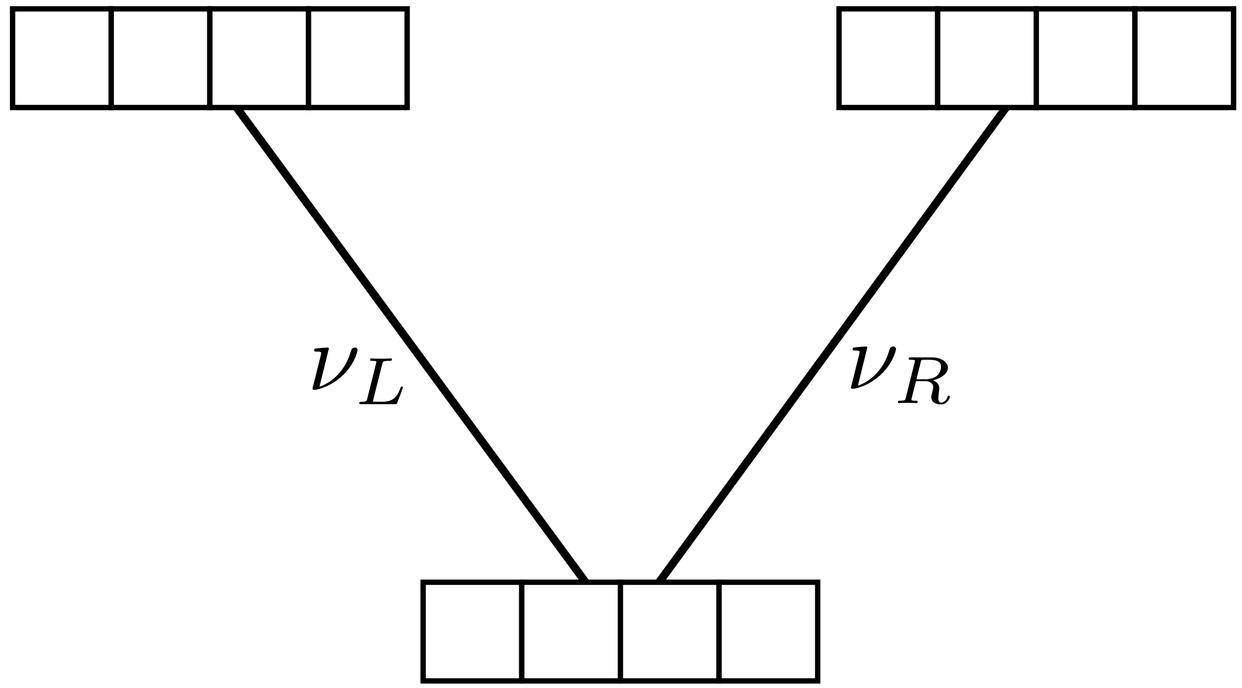
Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

Gallager, R. G. 1962. Low-density parity-check codes. IRE Trans. Inform. Theory 8:21–28.

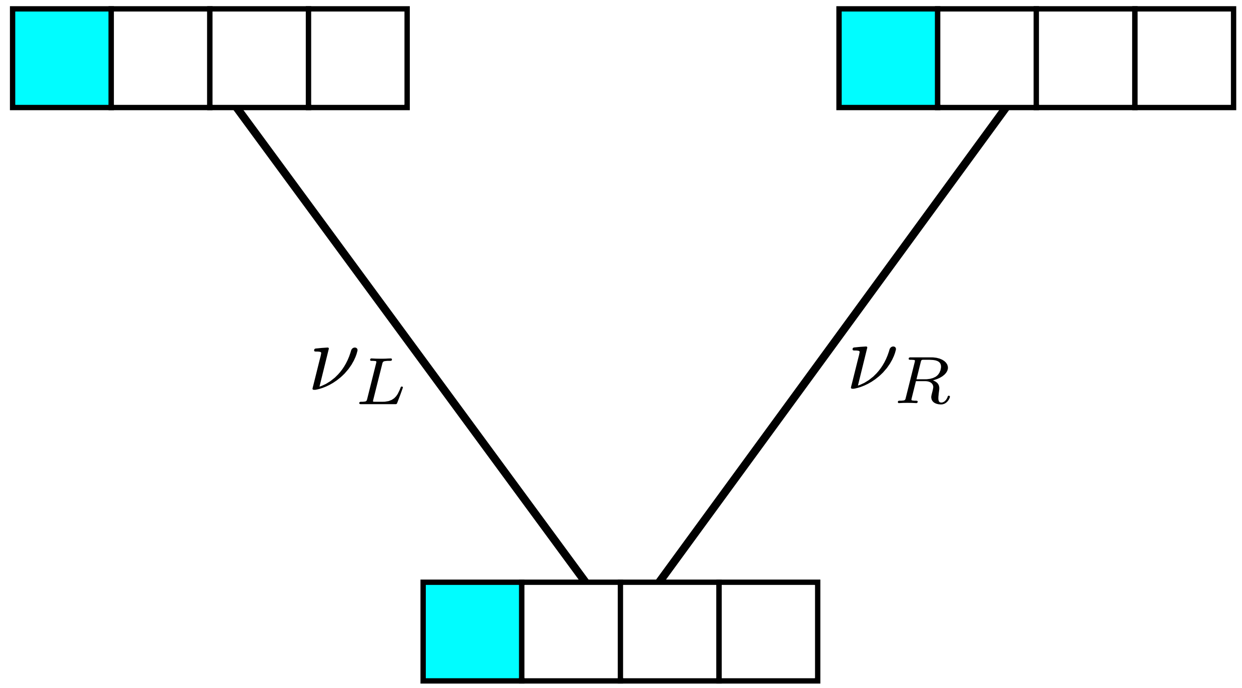
Gallager, R. G. 1963. Low-density parity-check codes. MIT Press, Cambridge, Mass.


G
G
A
0 0 1 0
0 0 1 0
1 0 0 0

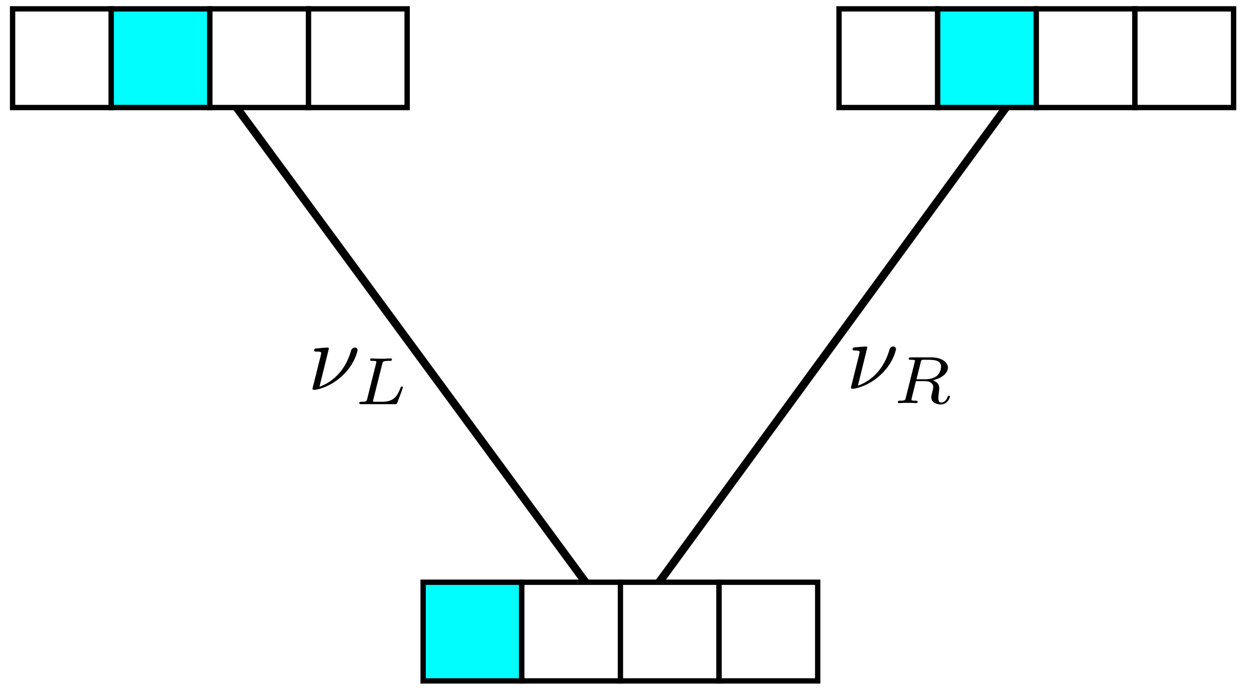

G
G
A
0 0 1 0
0 0 1 0
1 0 0 0
? ? ? ?
? ? ? ?



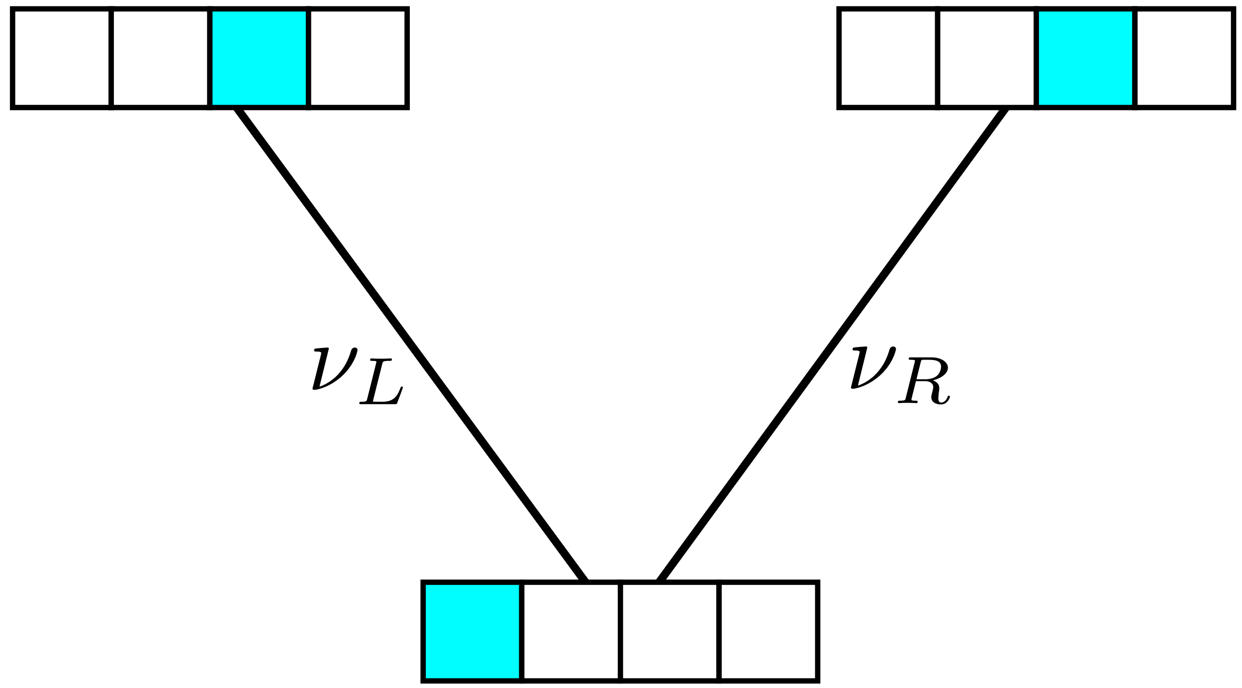
$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$



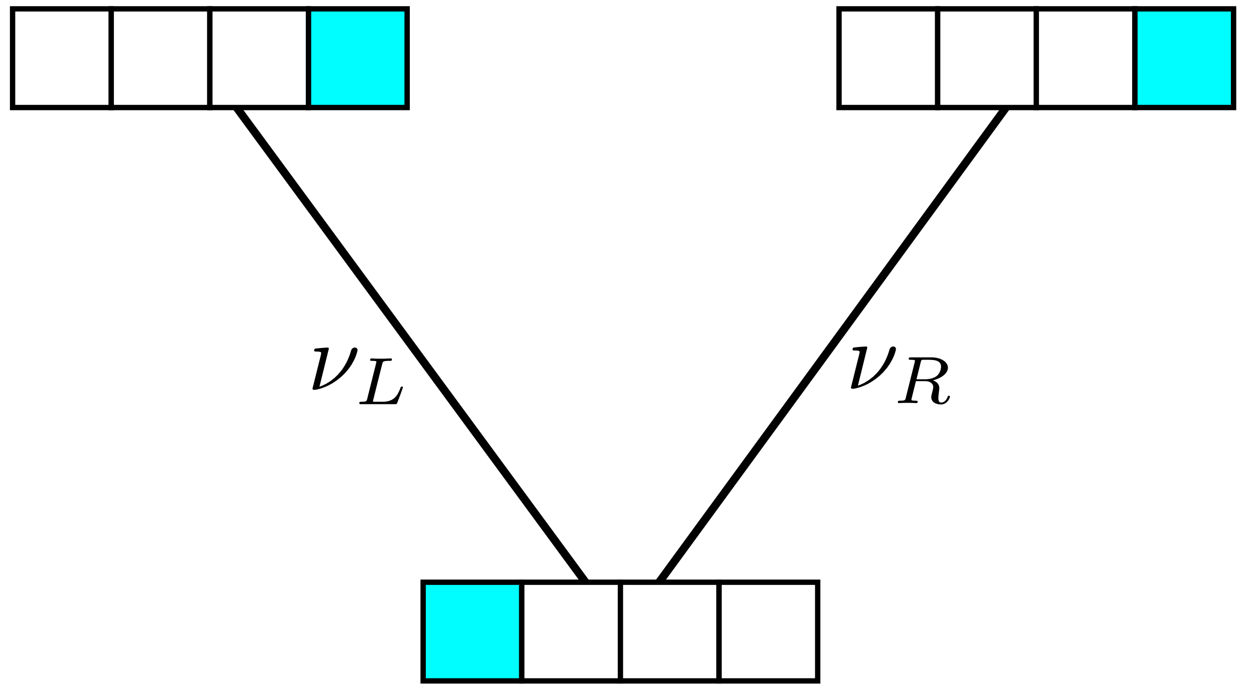
$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$



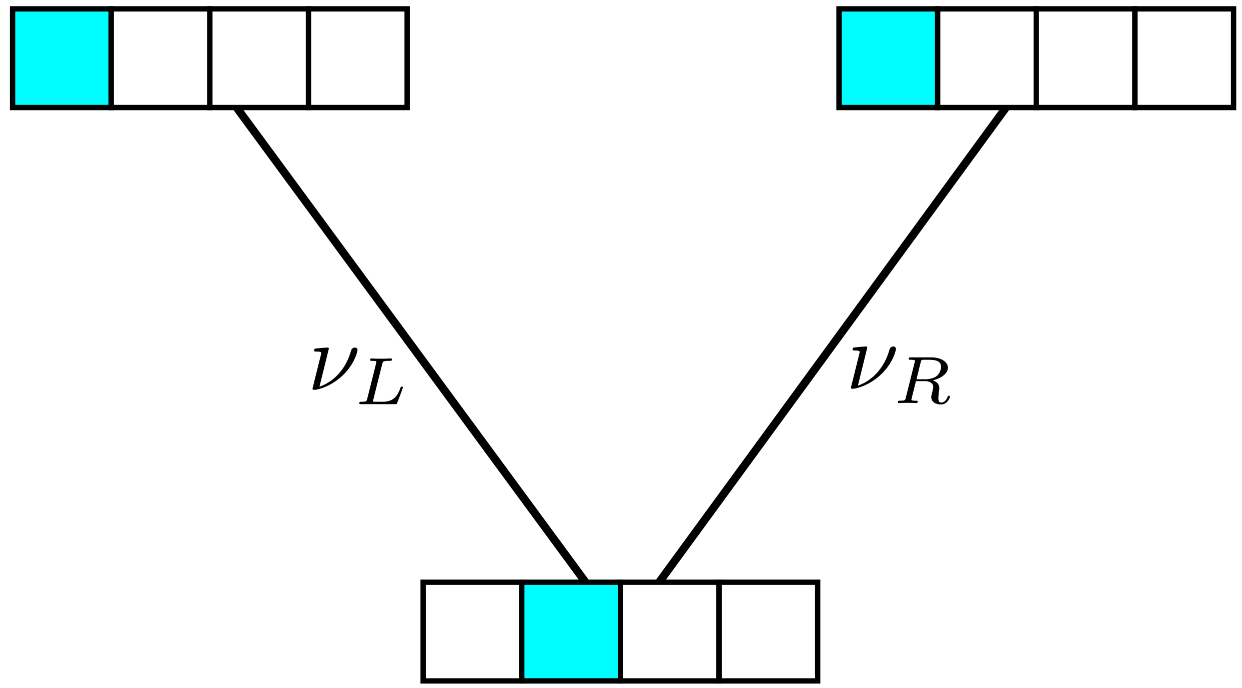
$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$



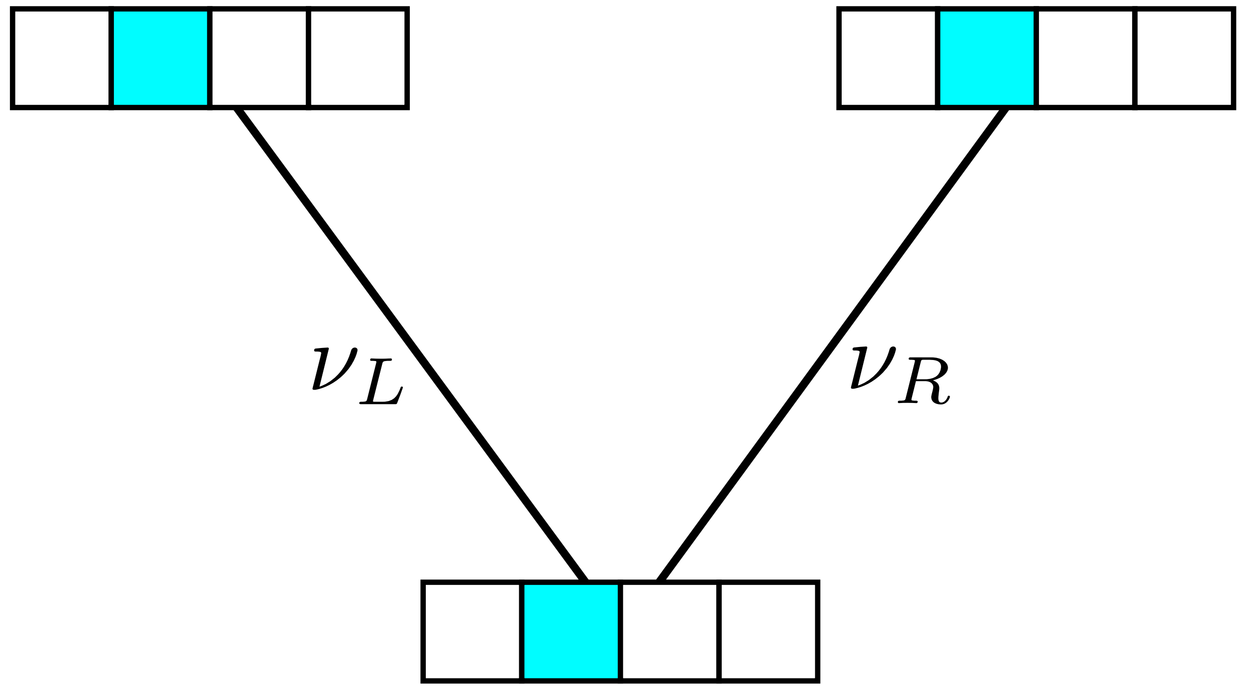
$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$



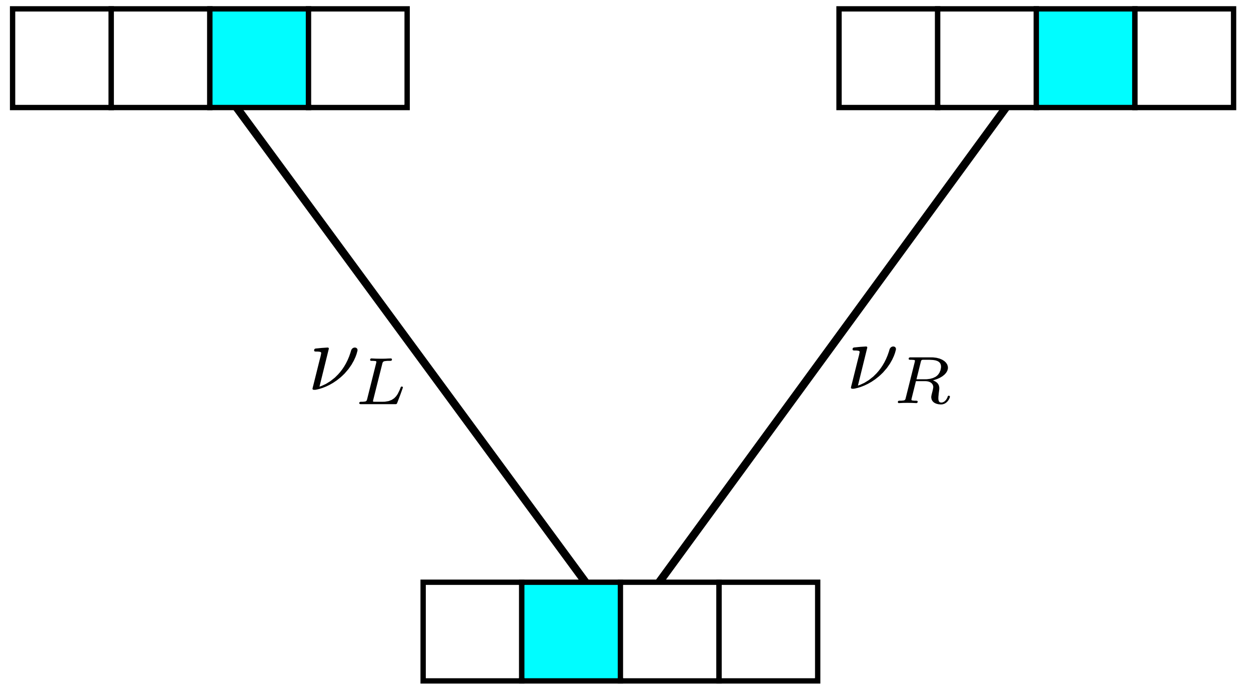
$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$



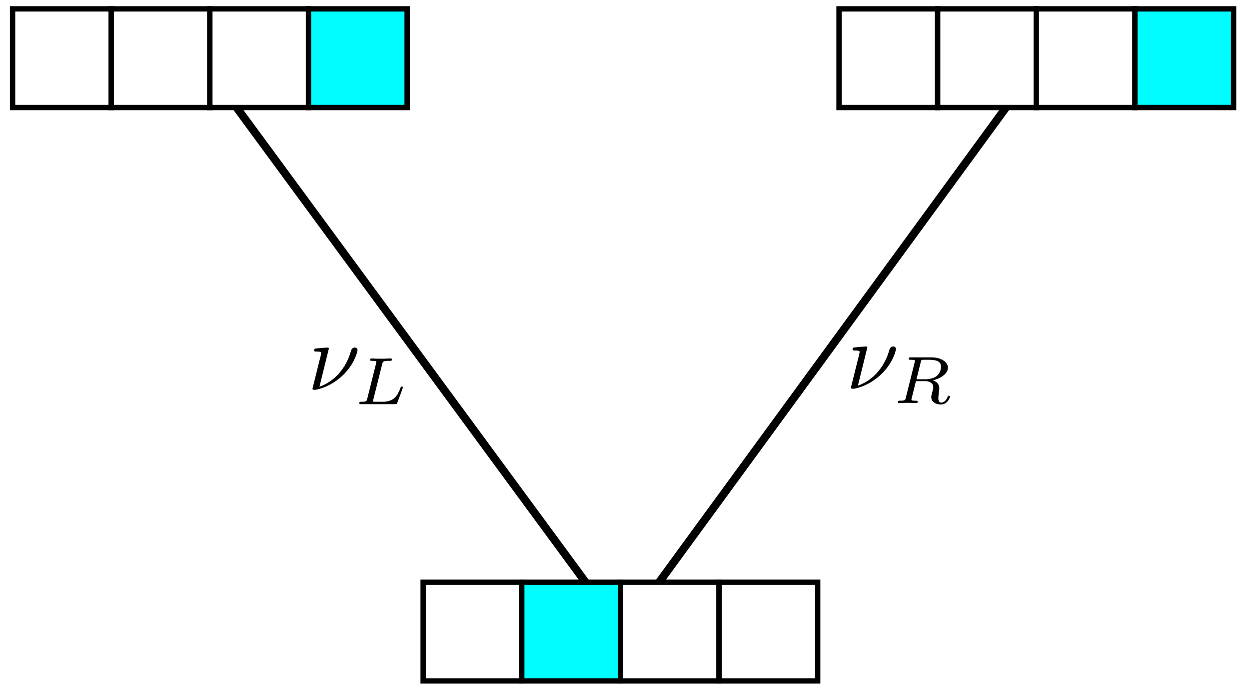
$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$



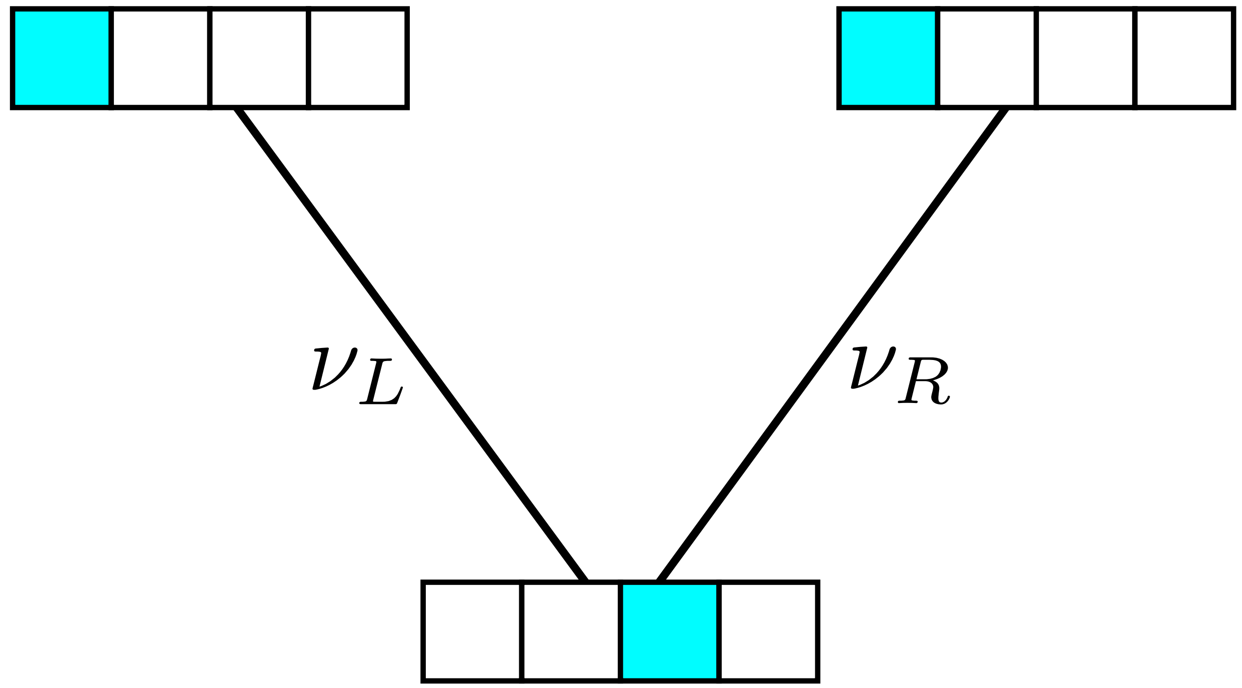
$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$



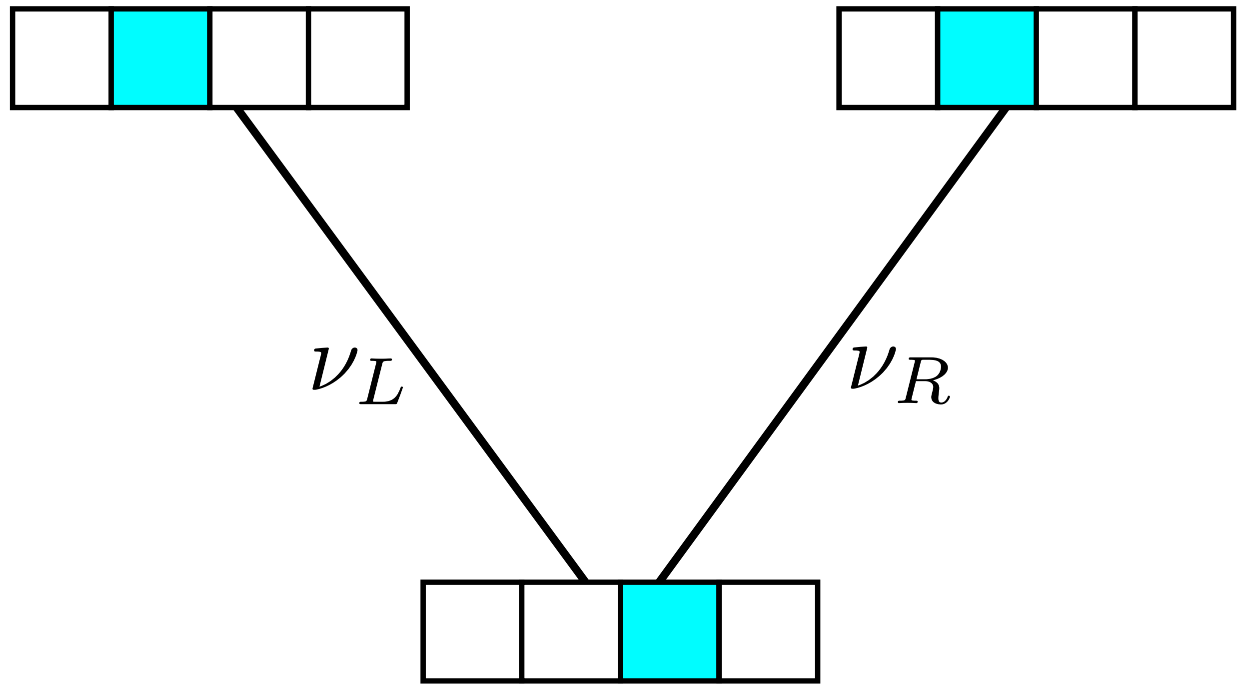
$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$



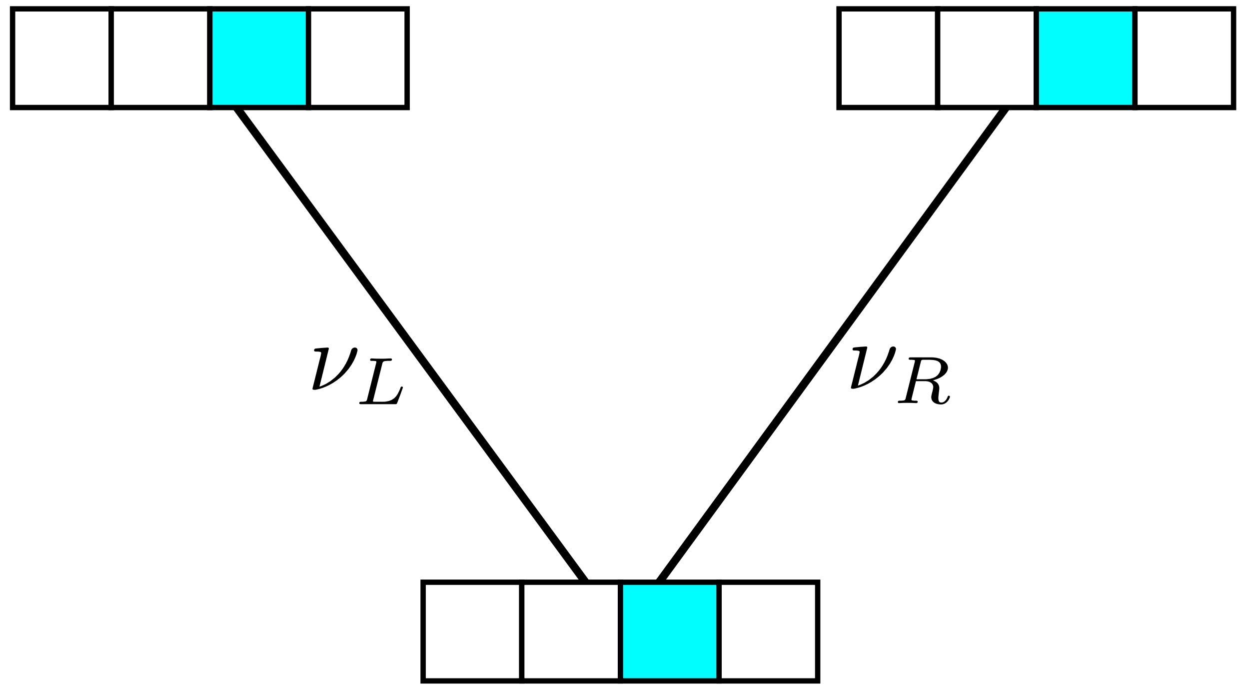
$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$



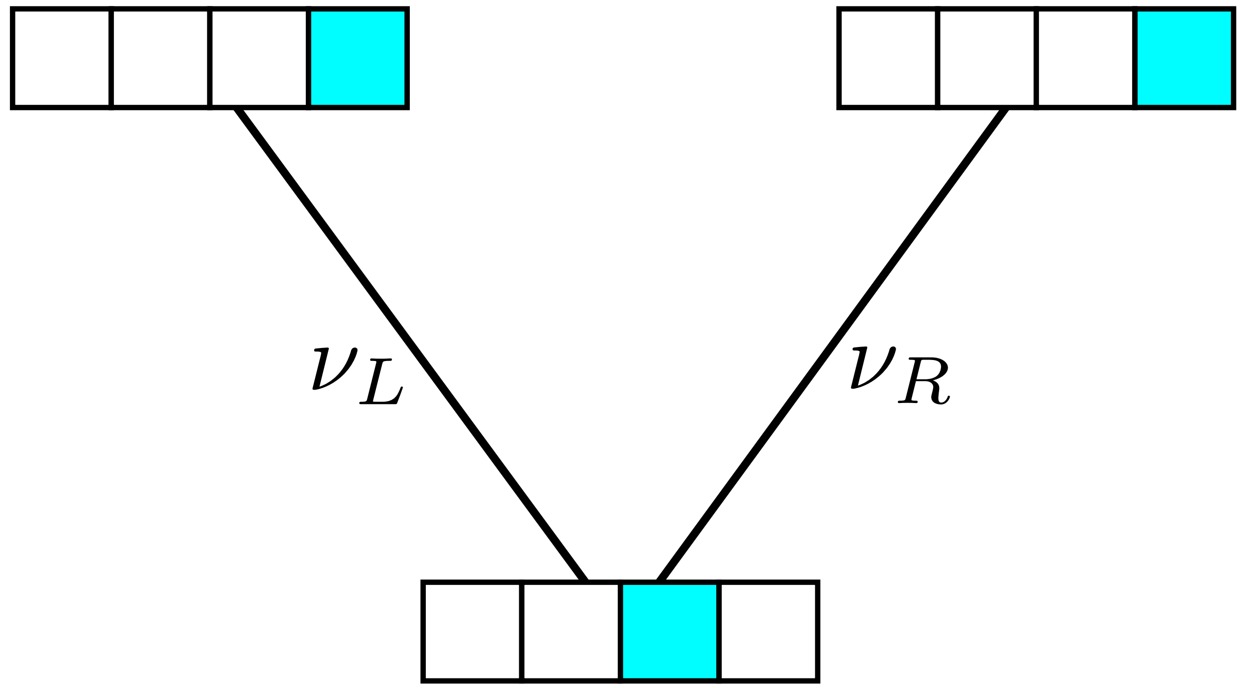
$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$



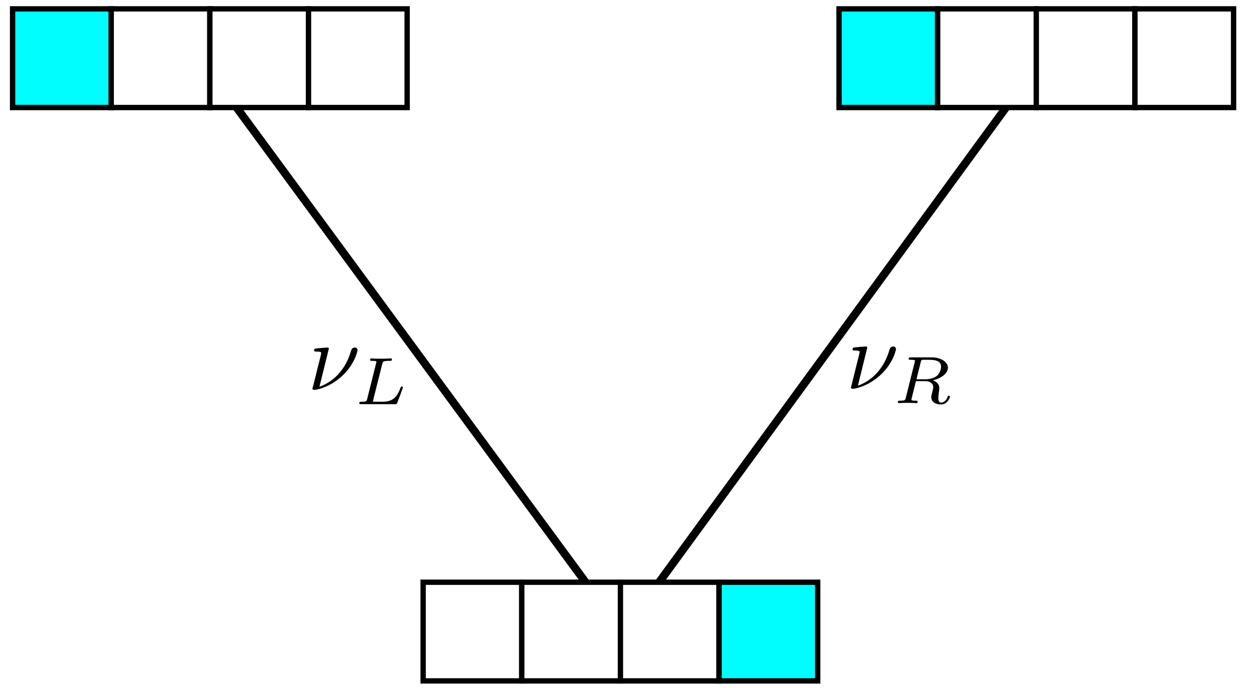
$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$



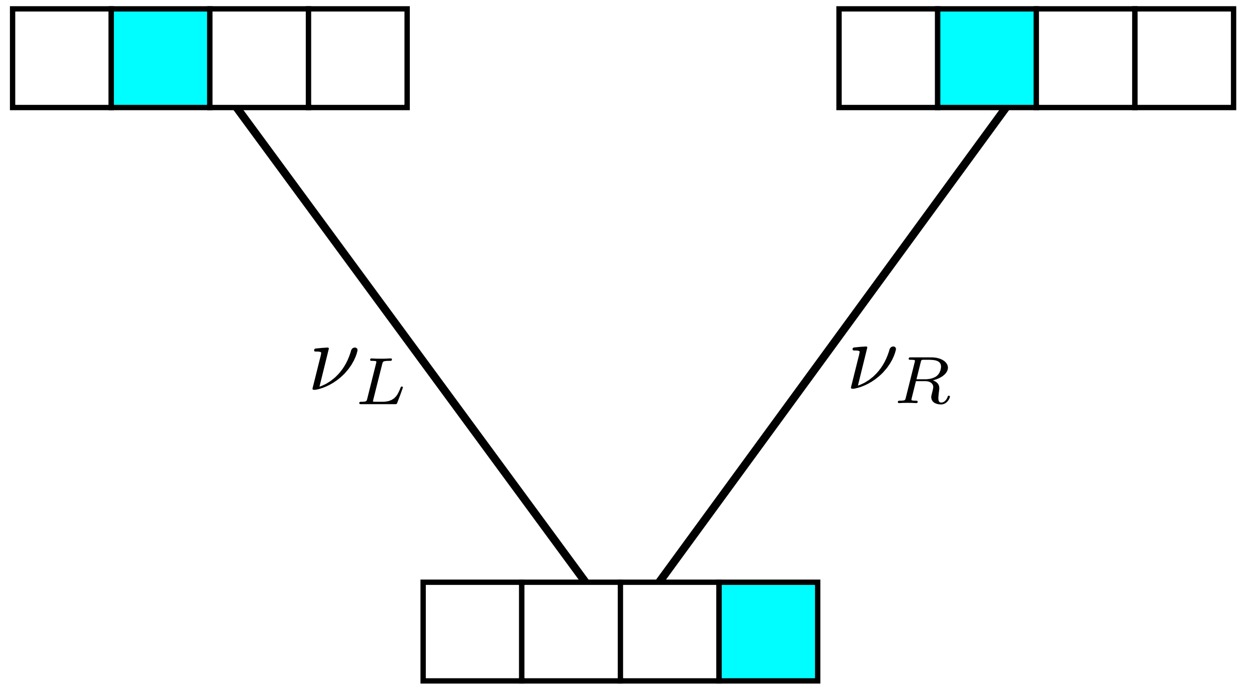
$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$



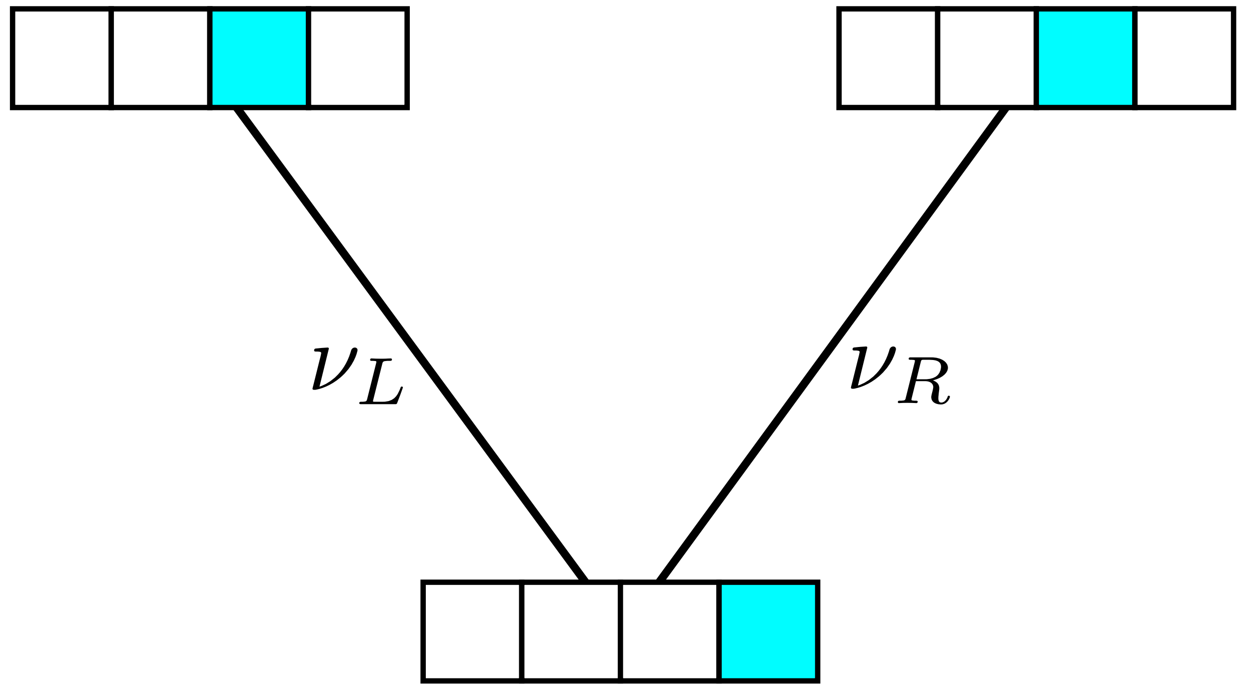
$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$



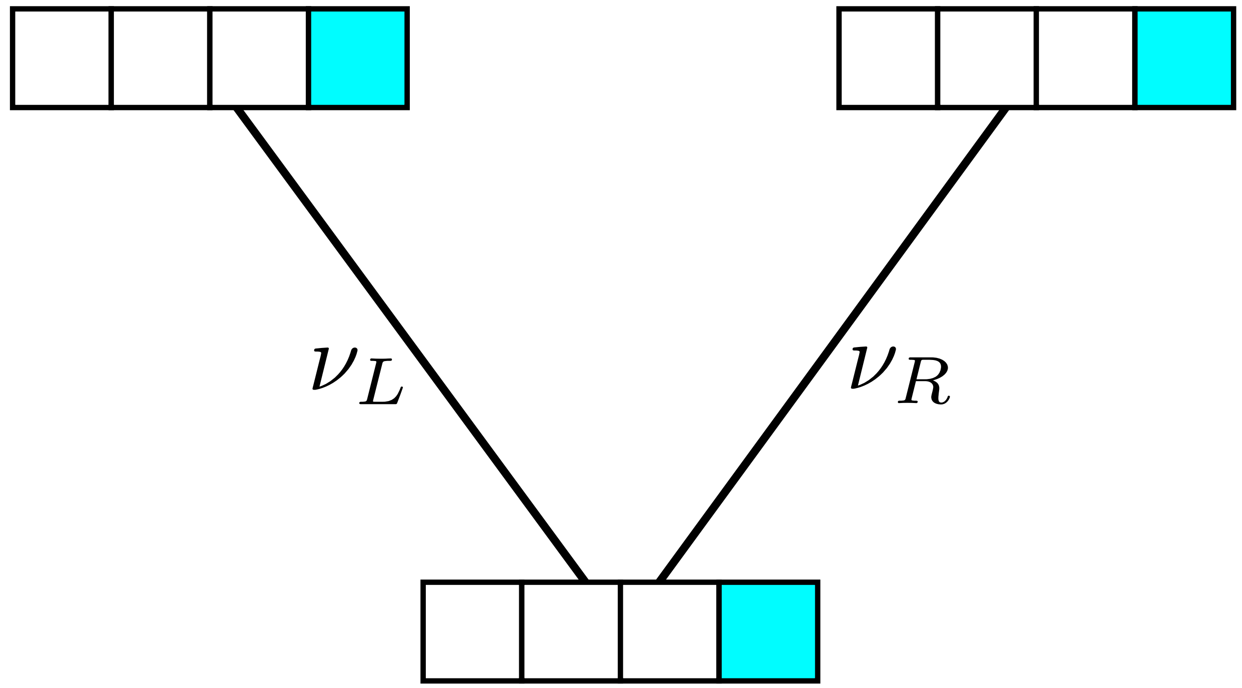
$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$



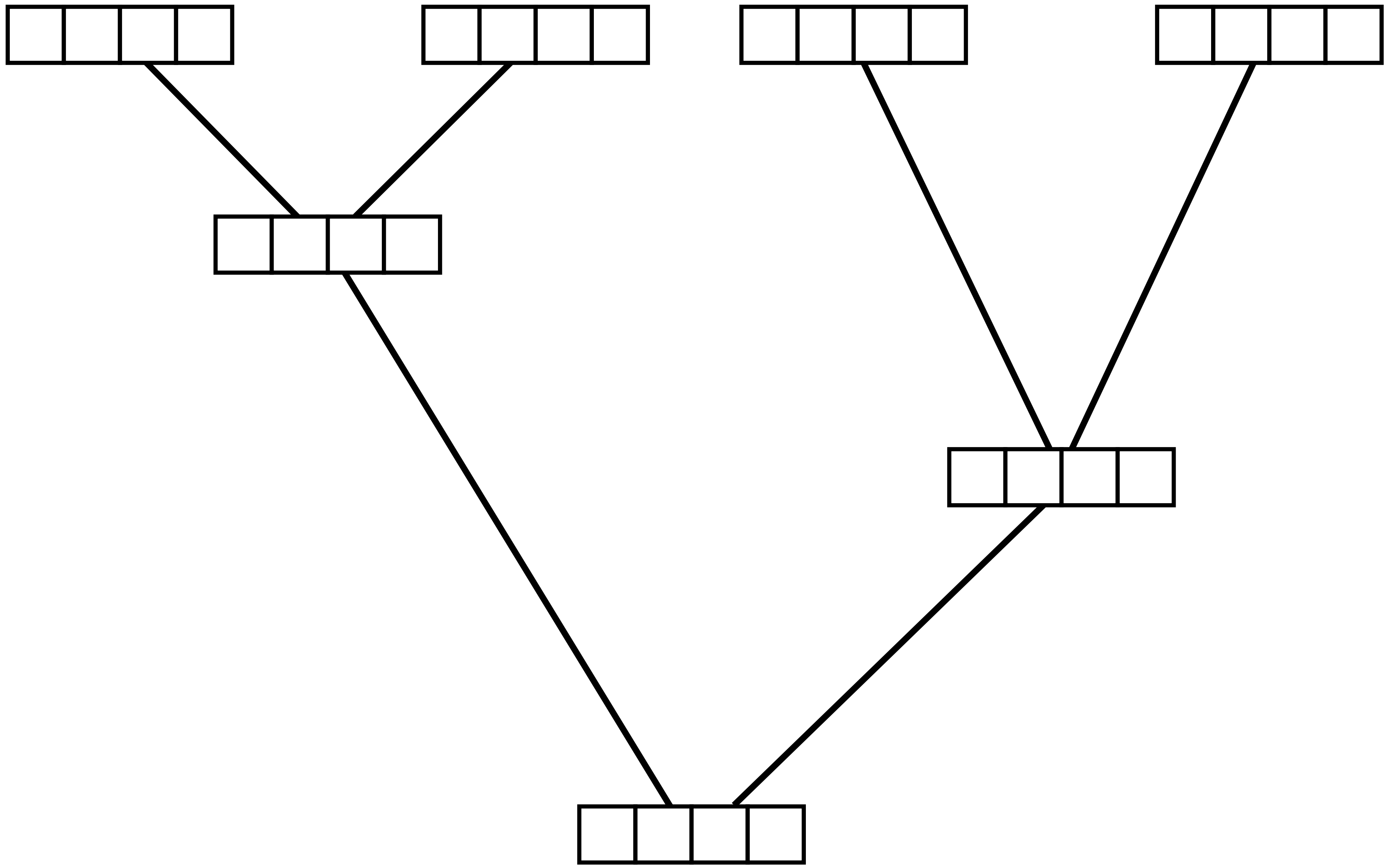
$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$

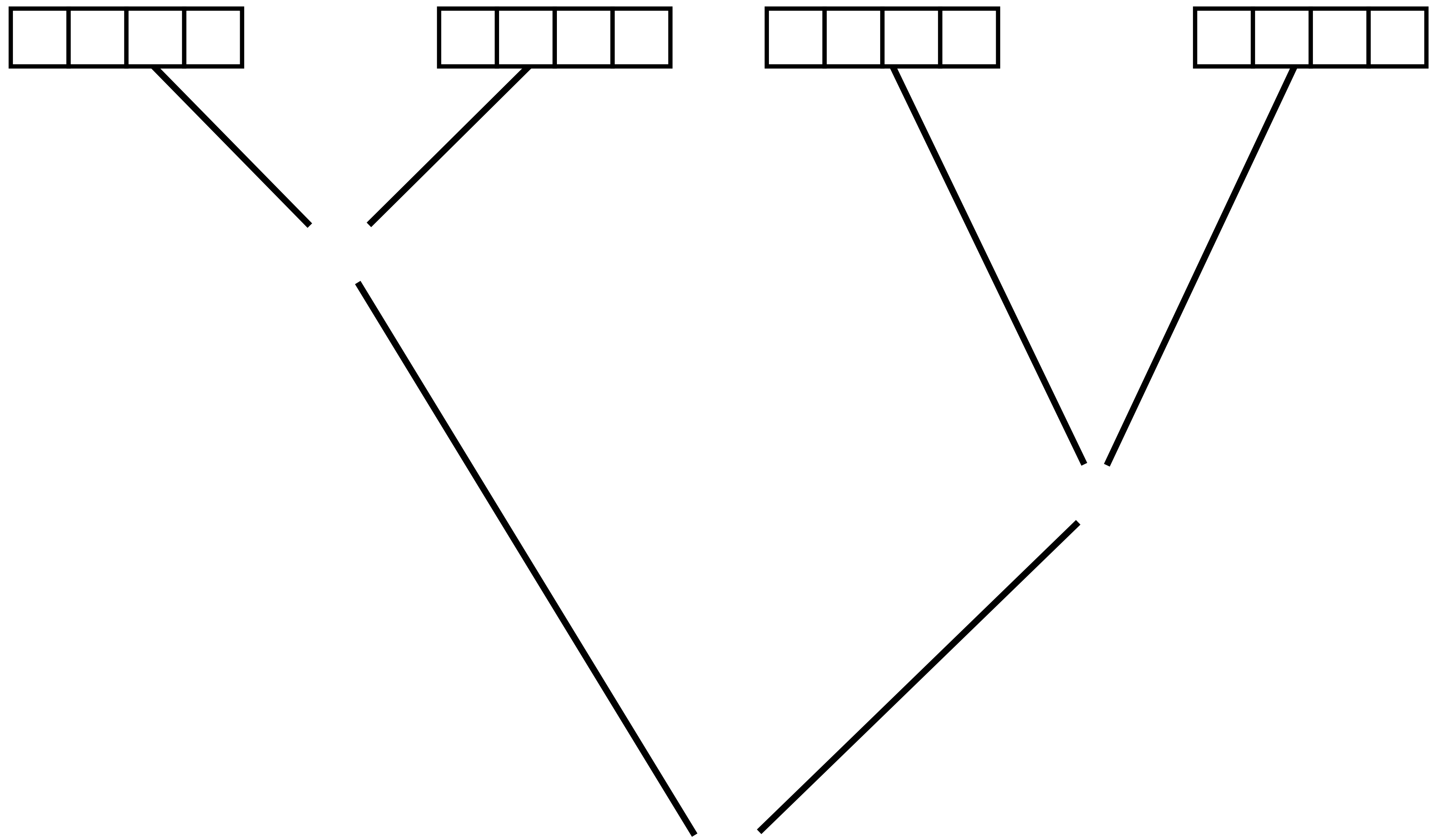


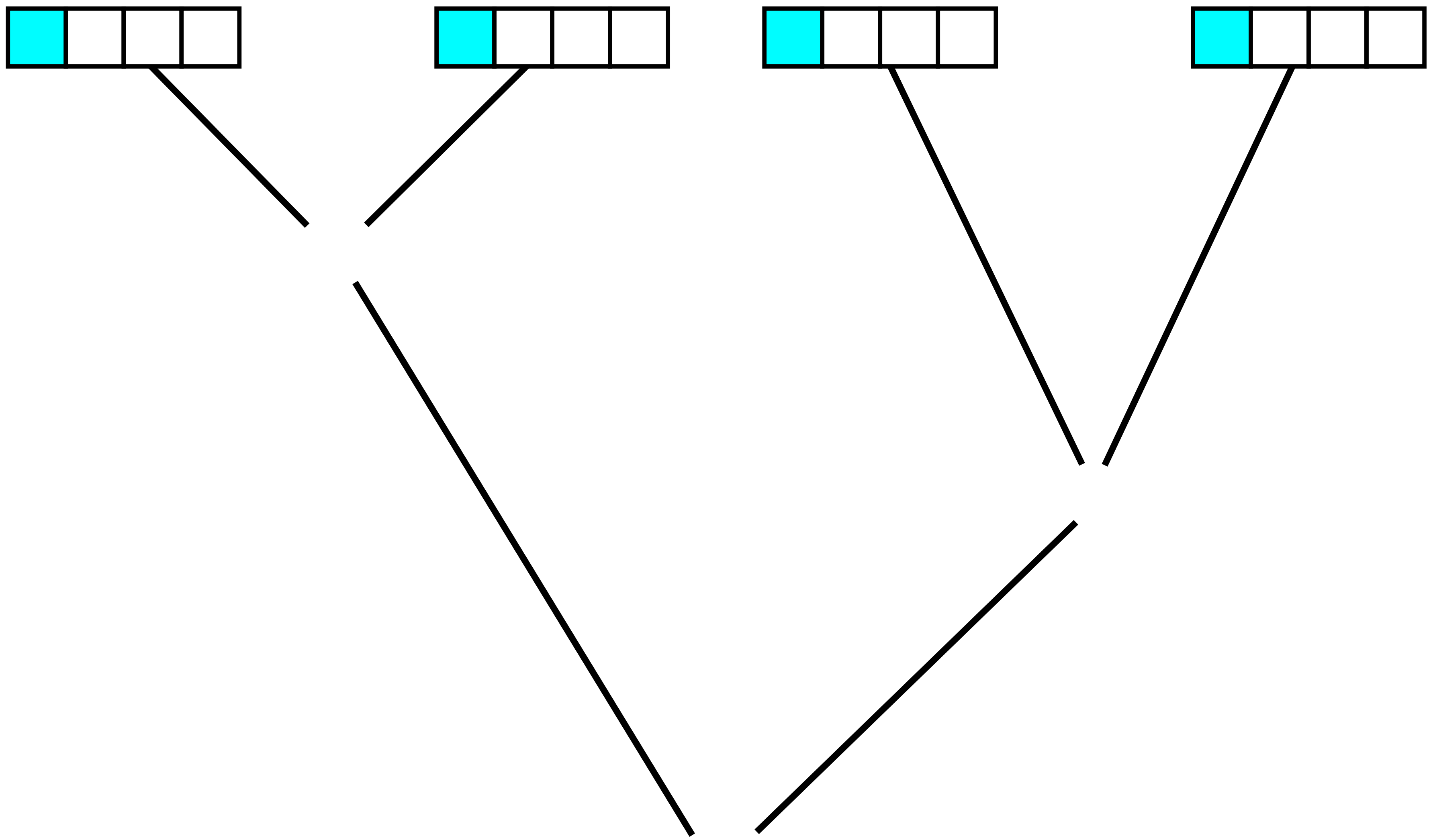
$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$

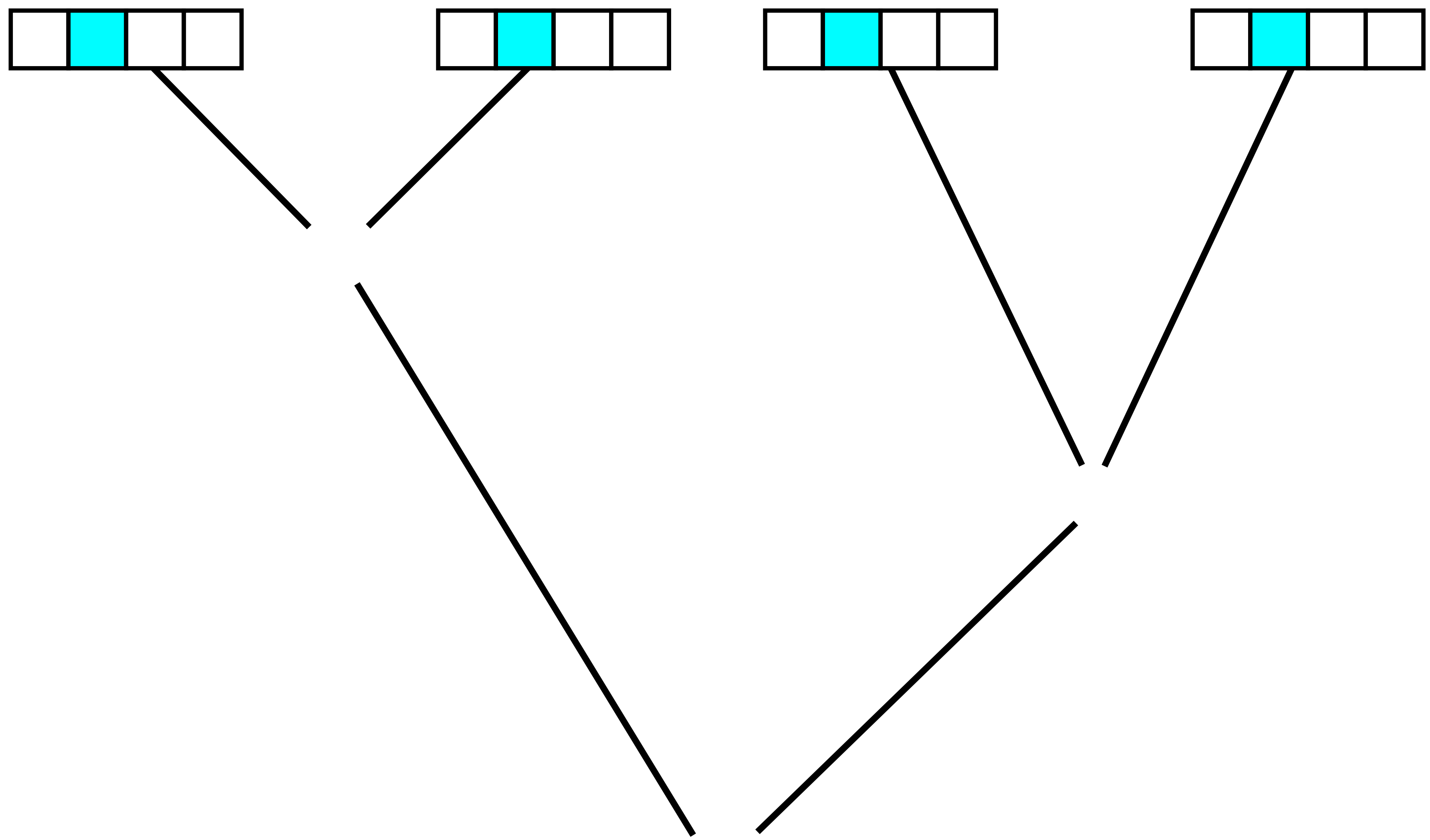


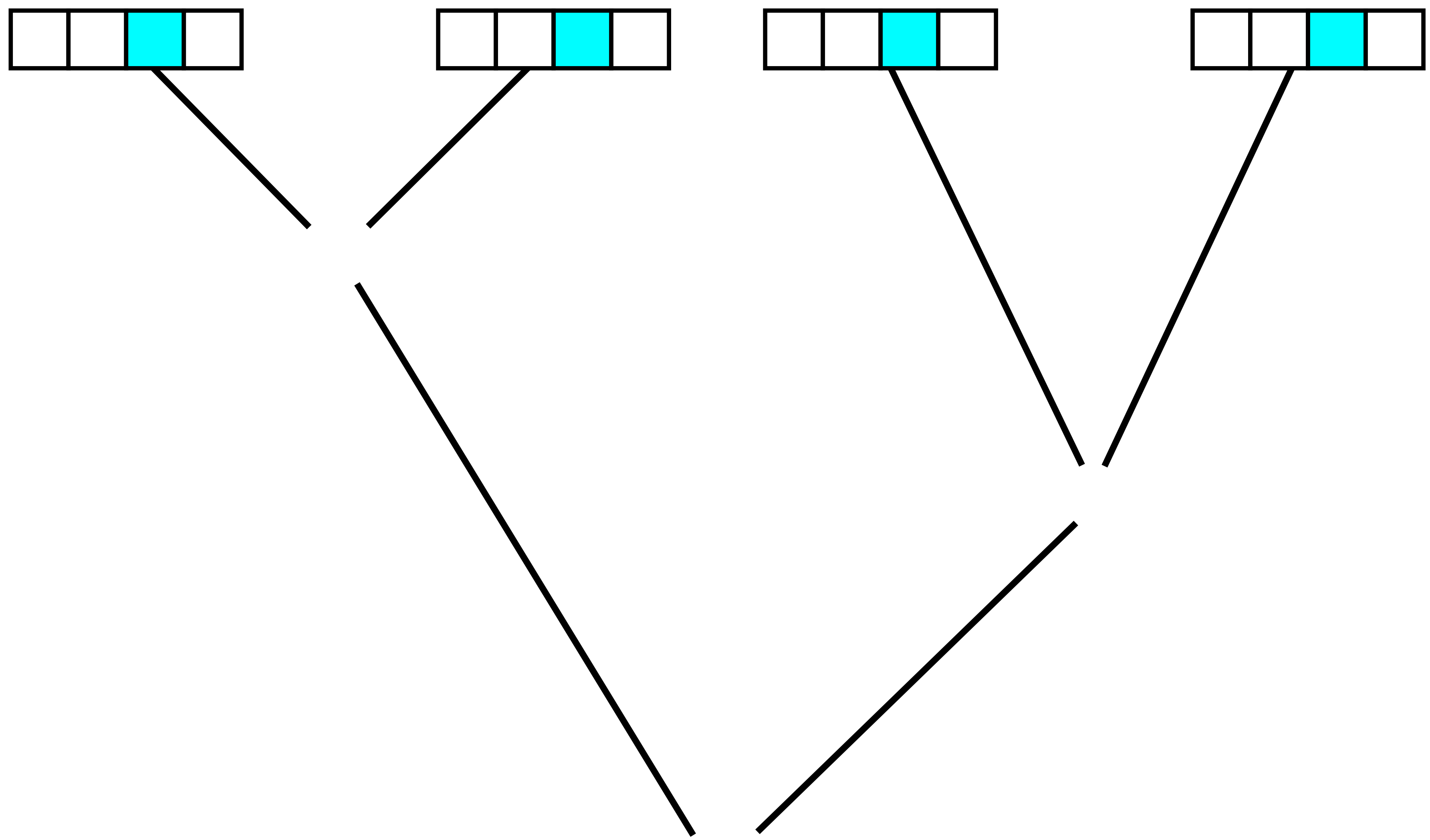
$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$

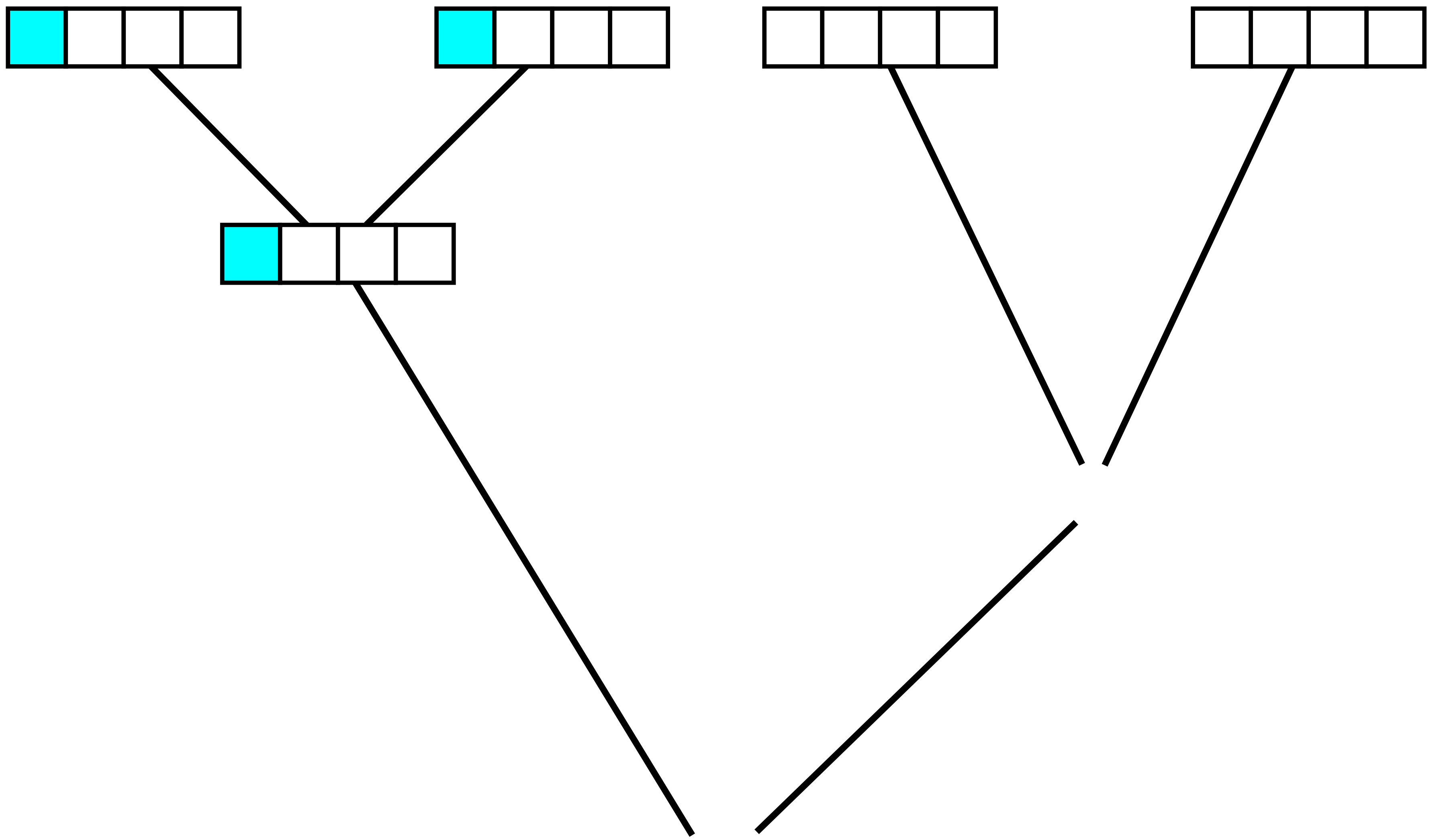


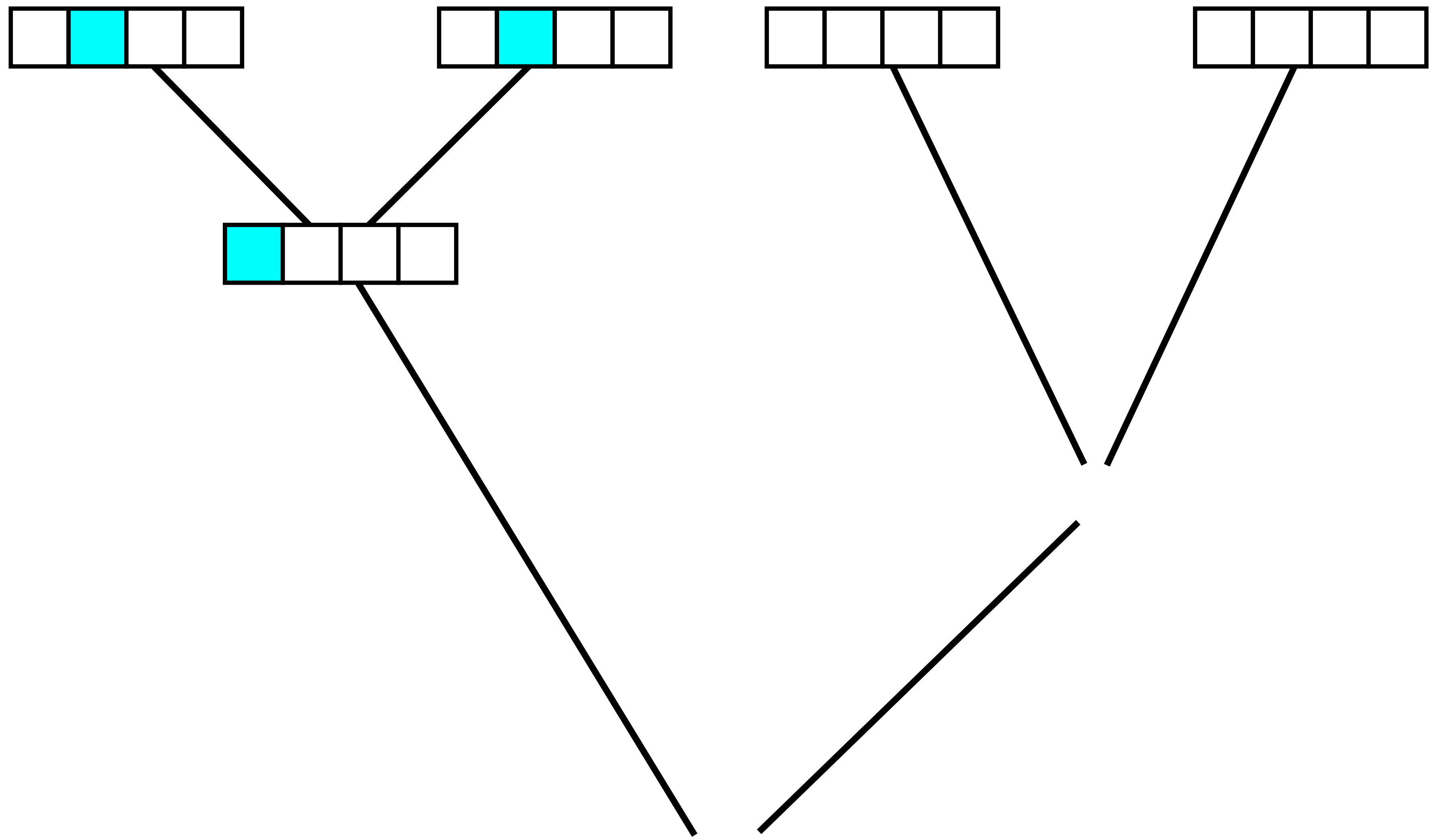


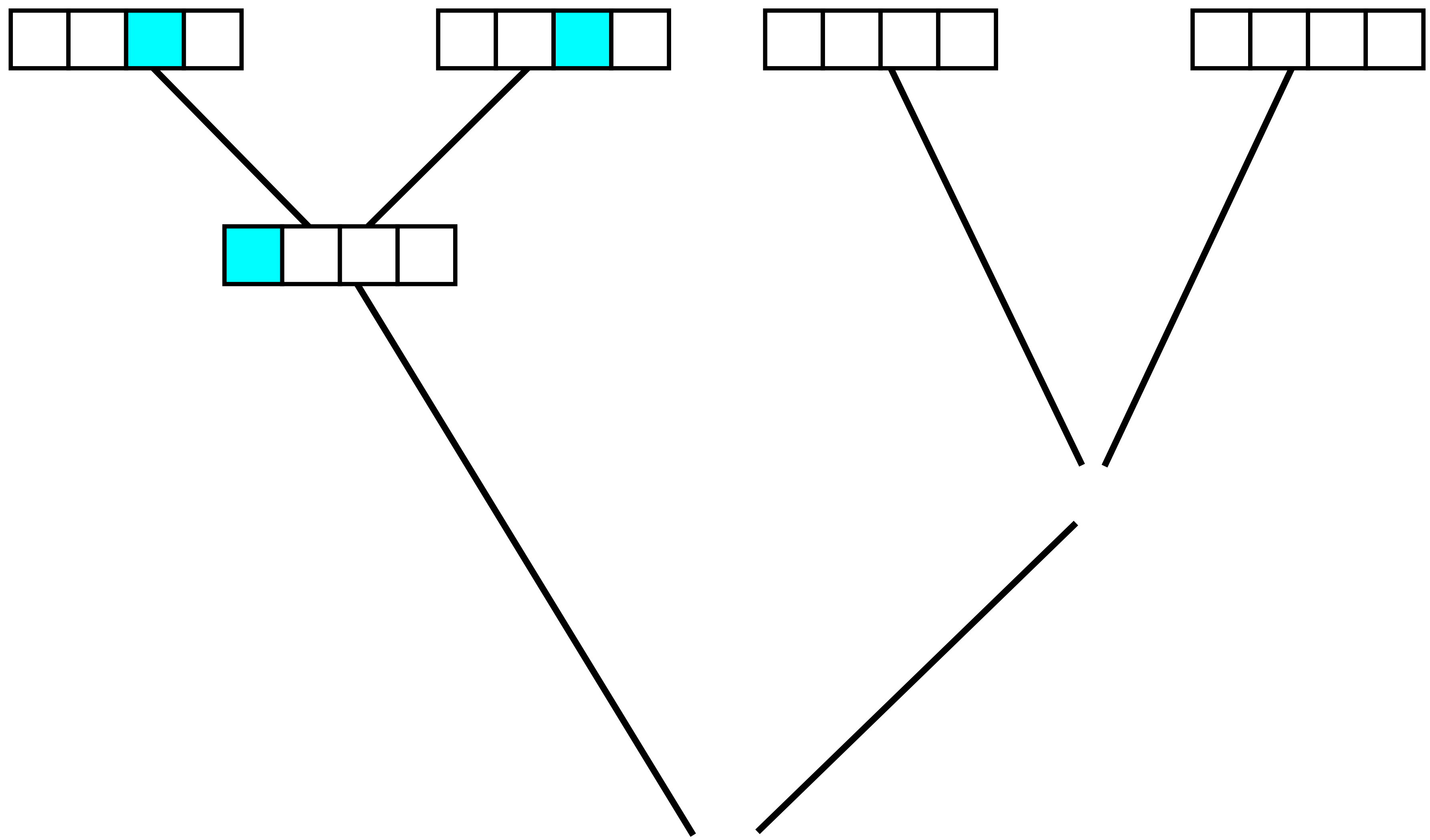


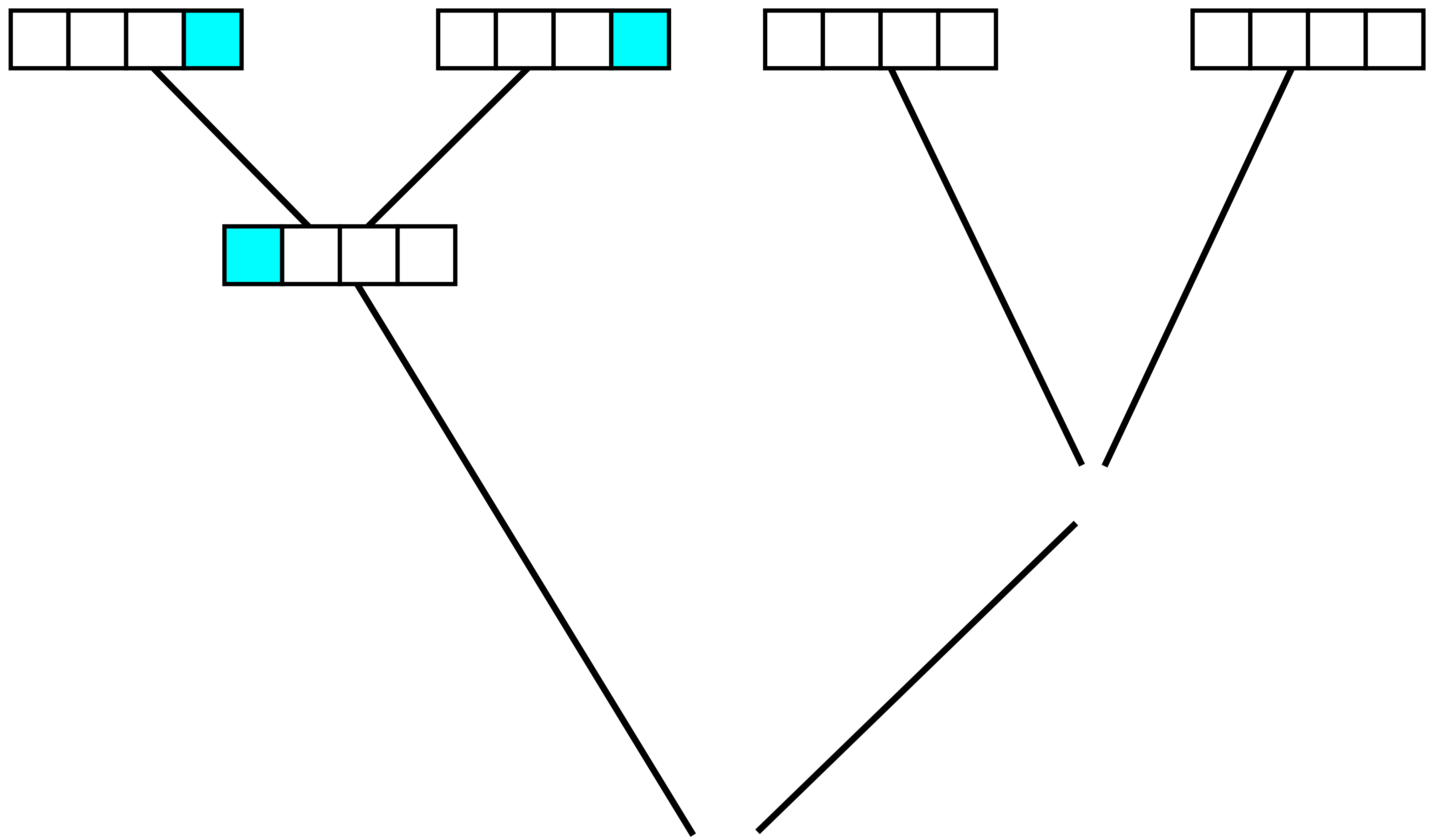


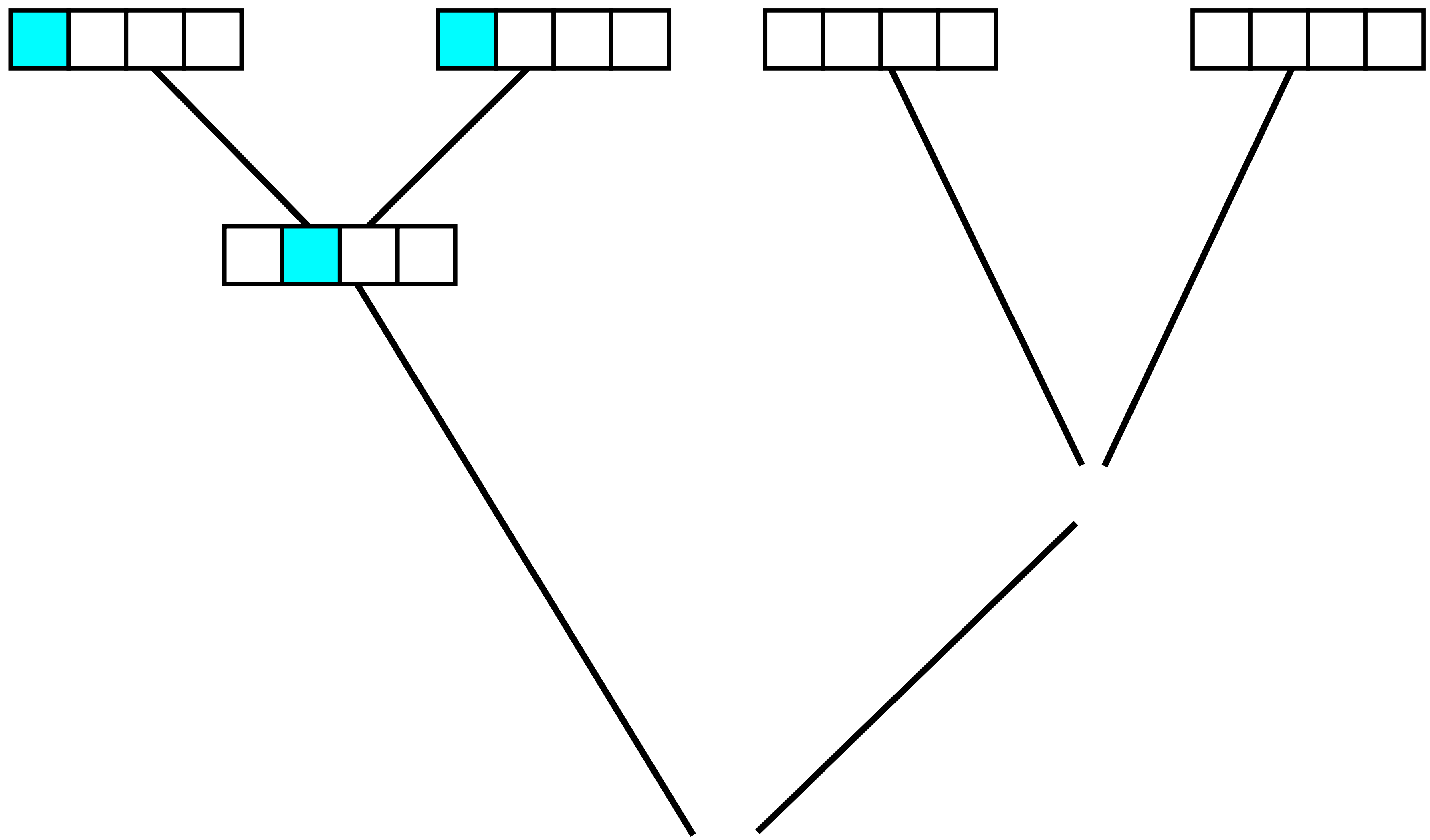


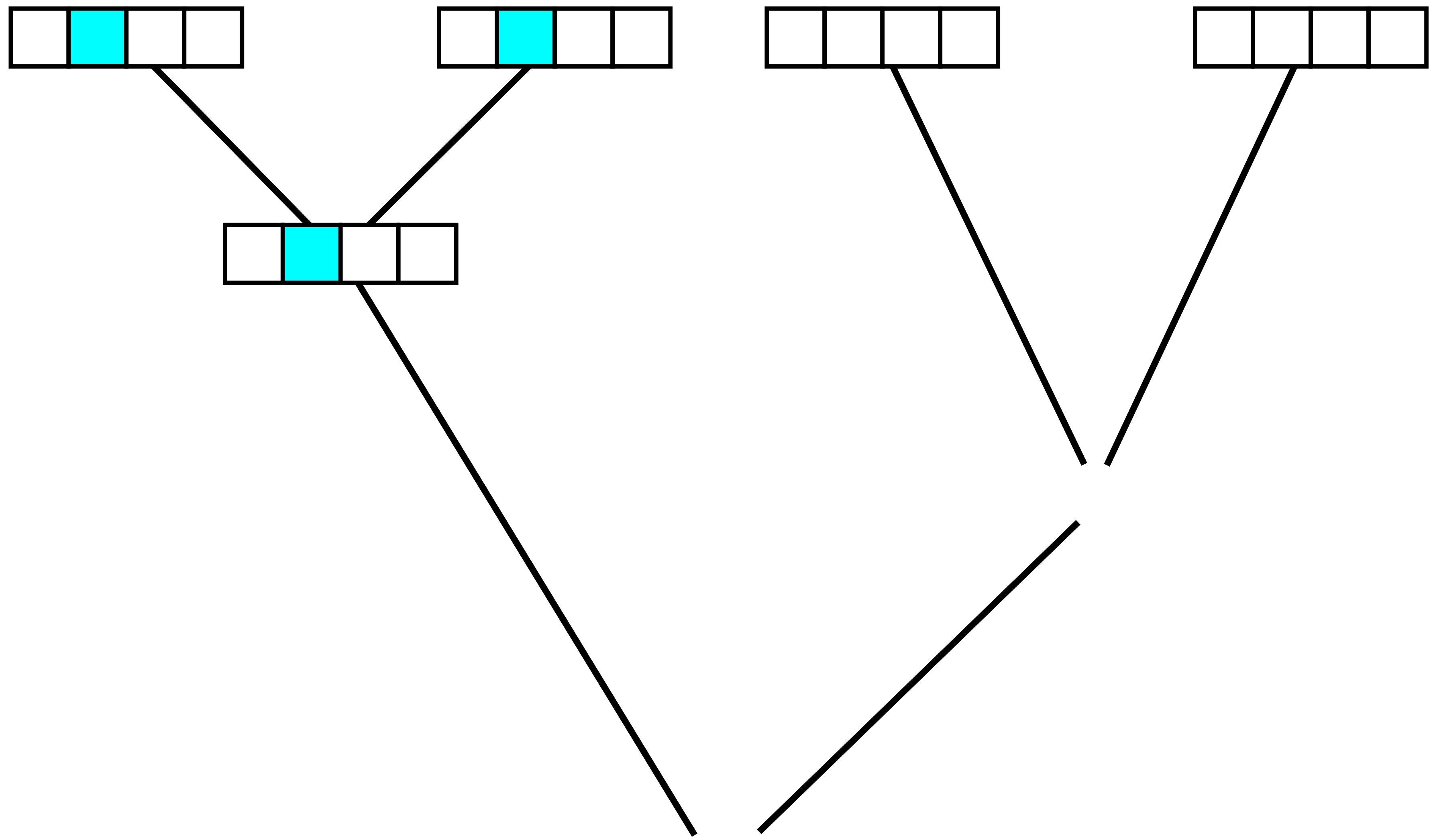


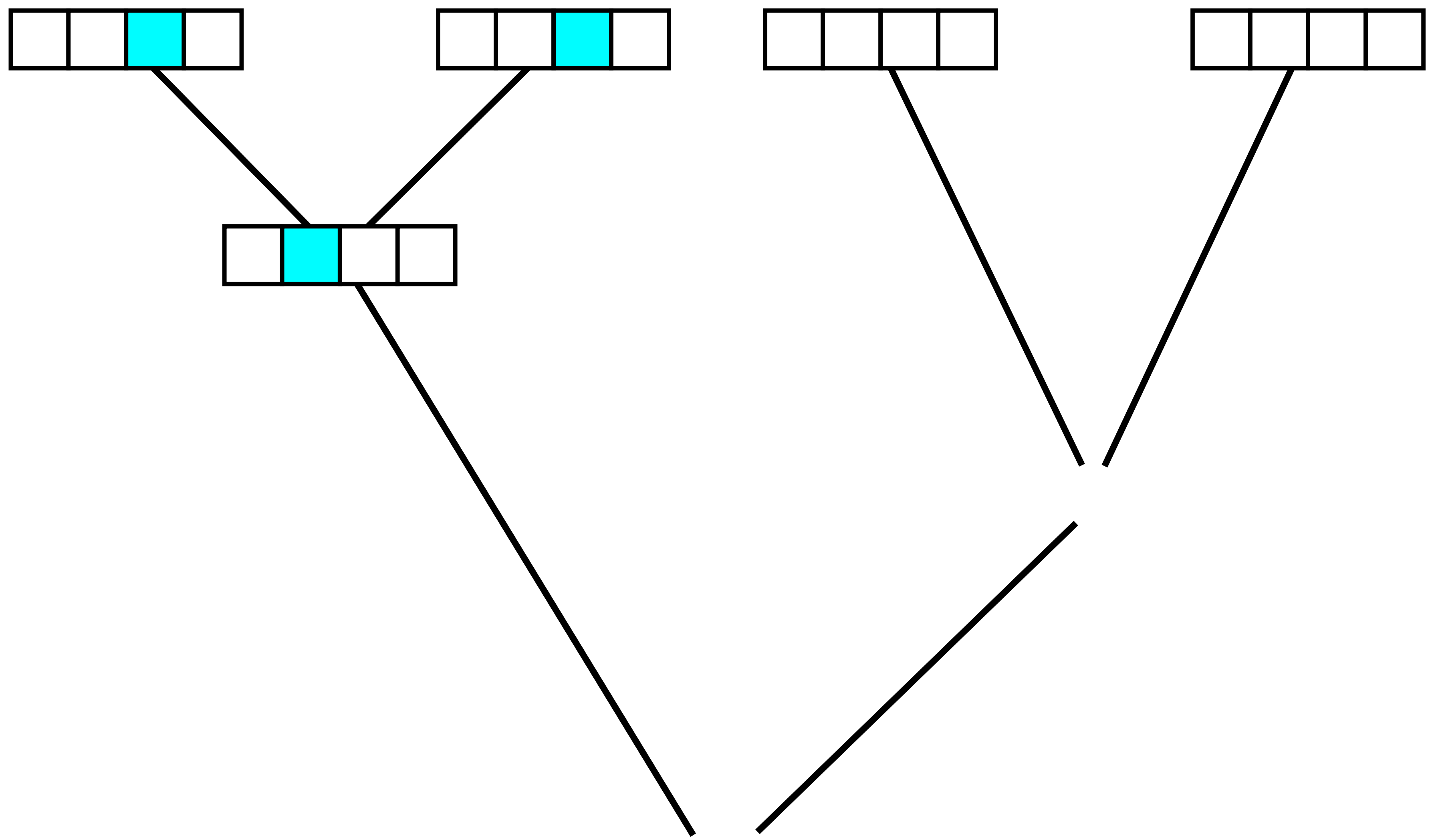


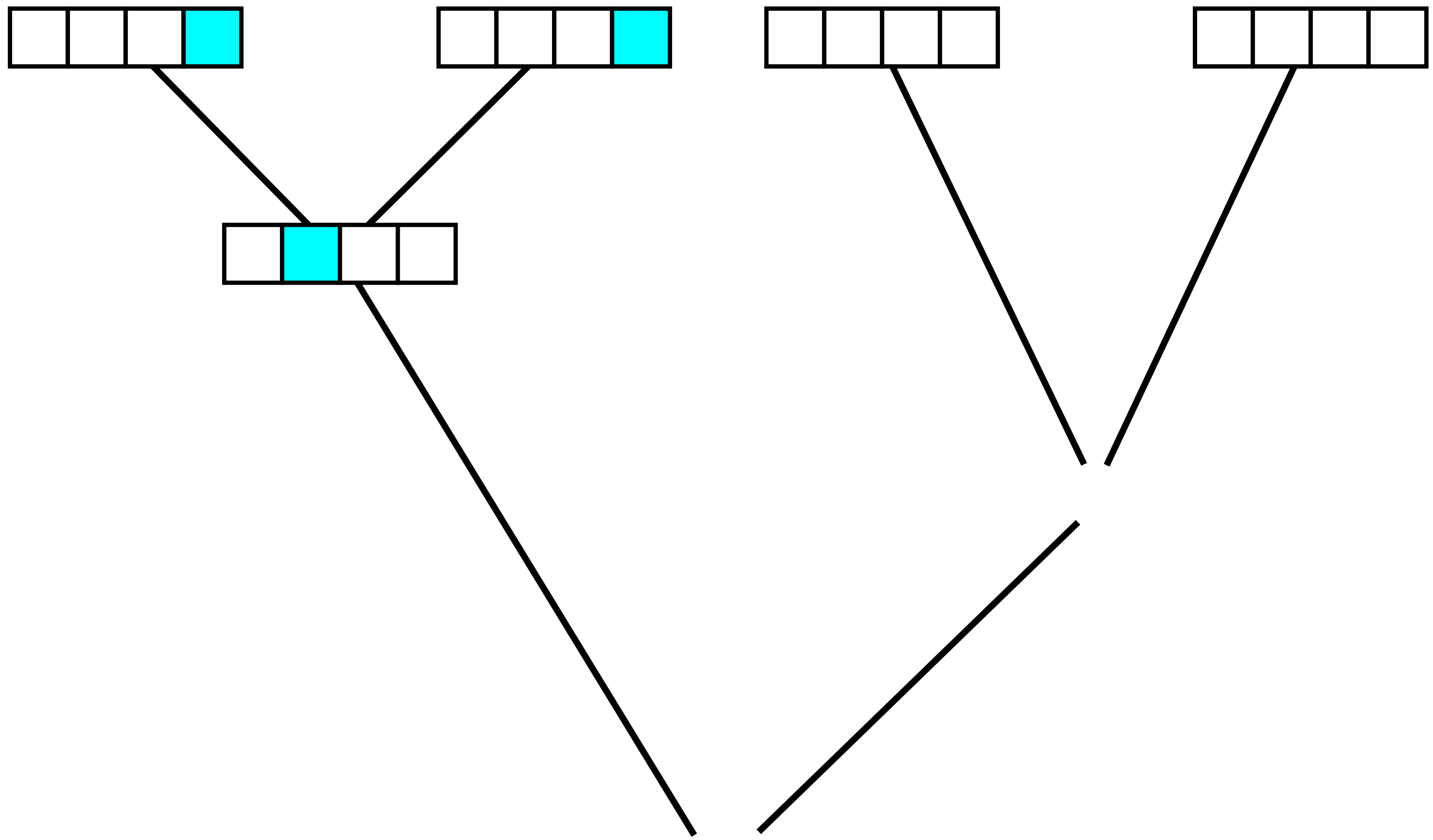


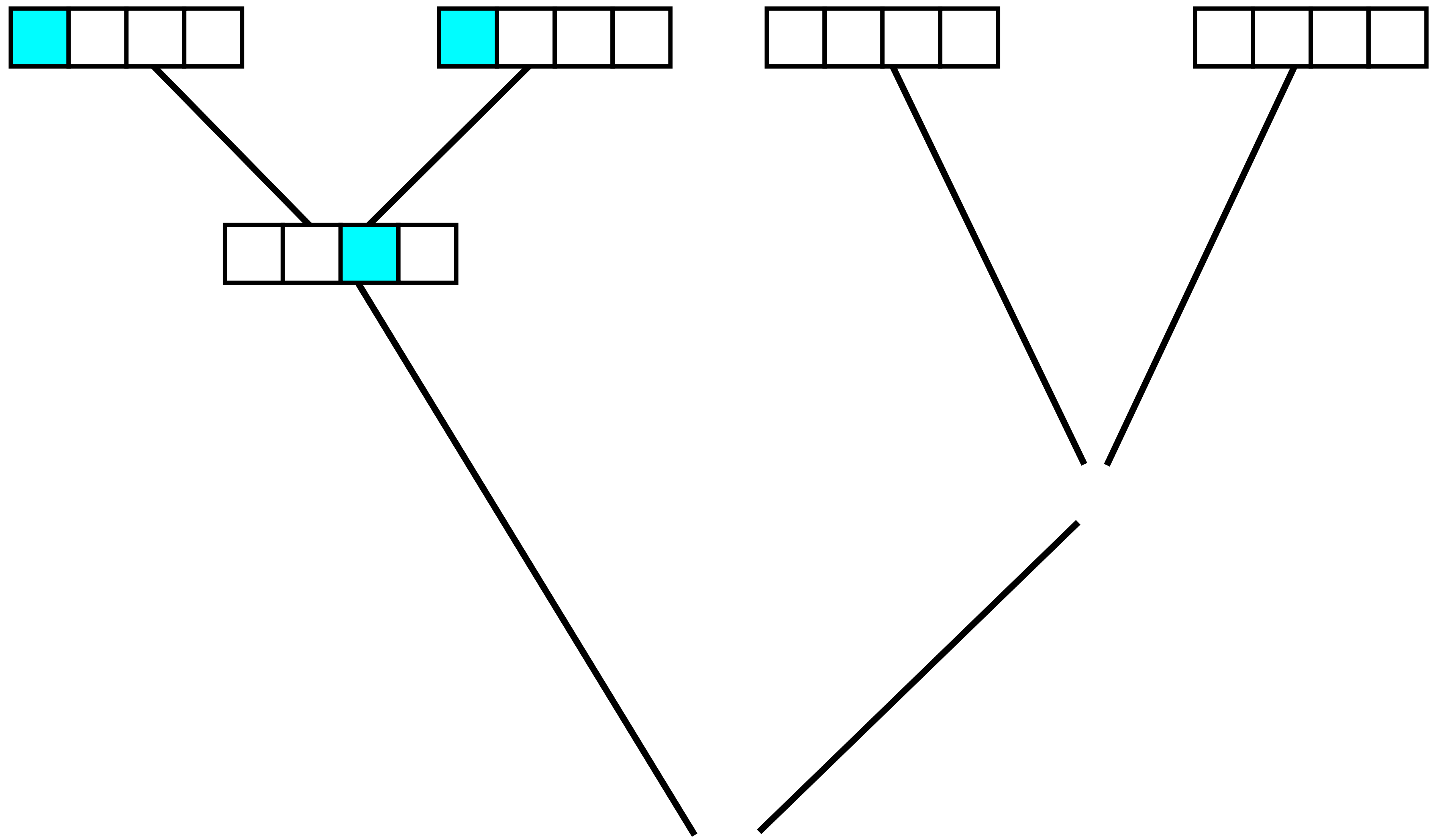


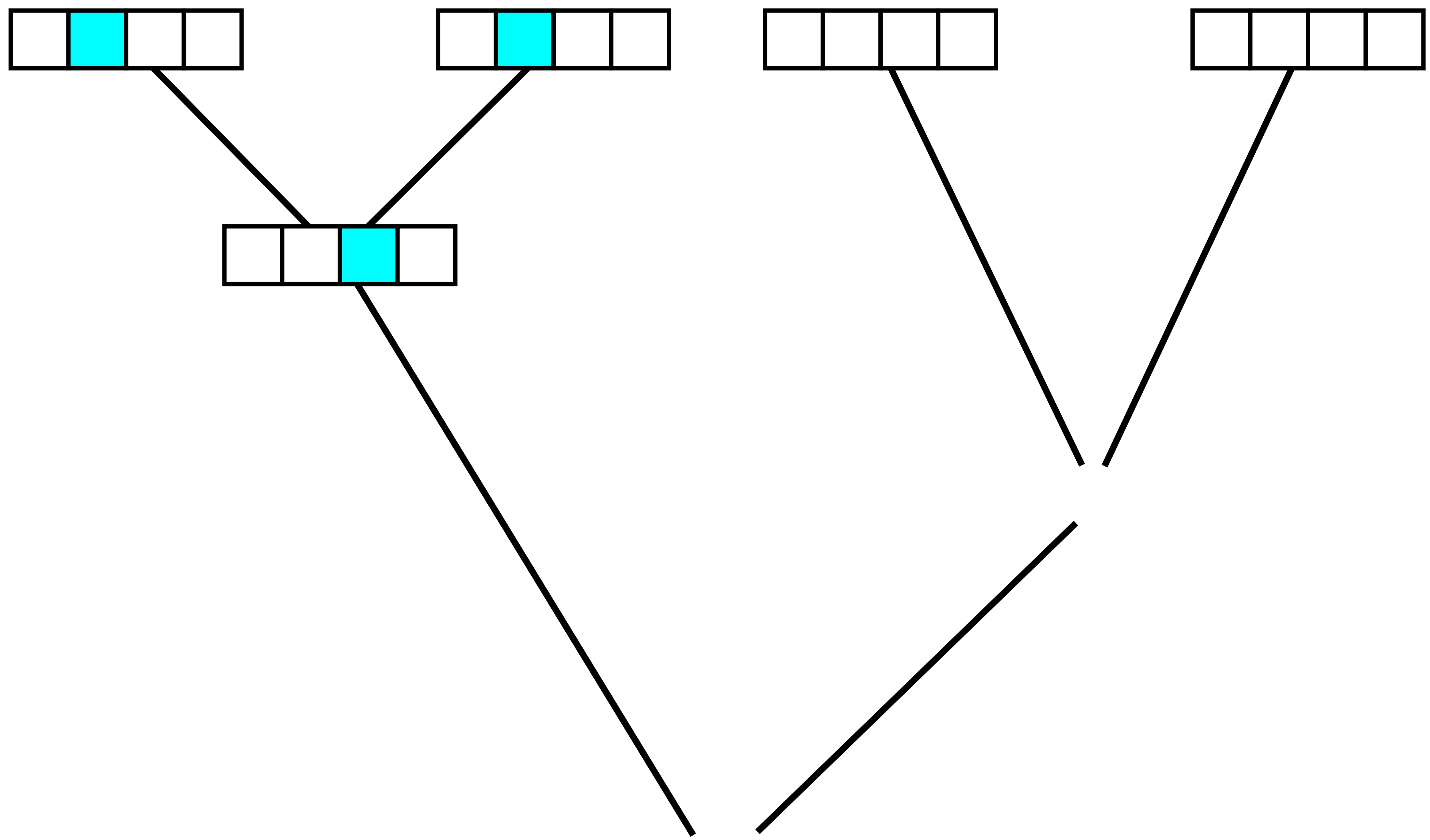


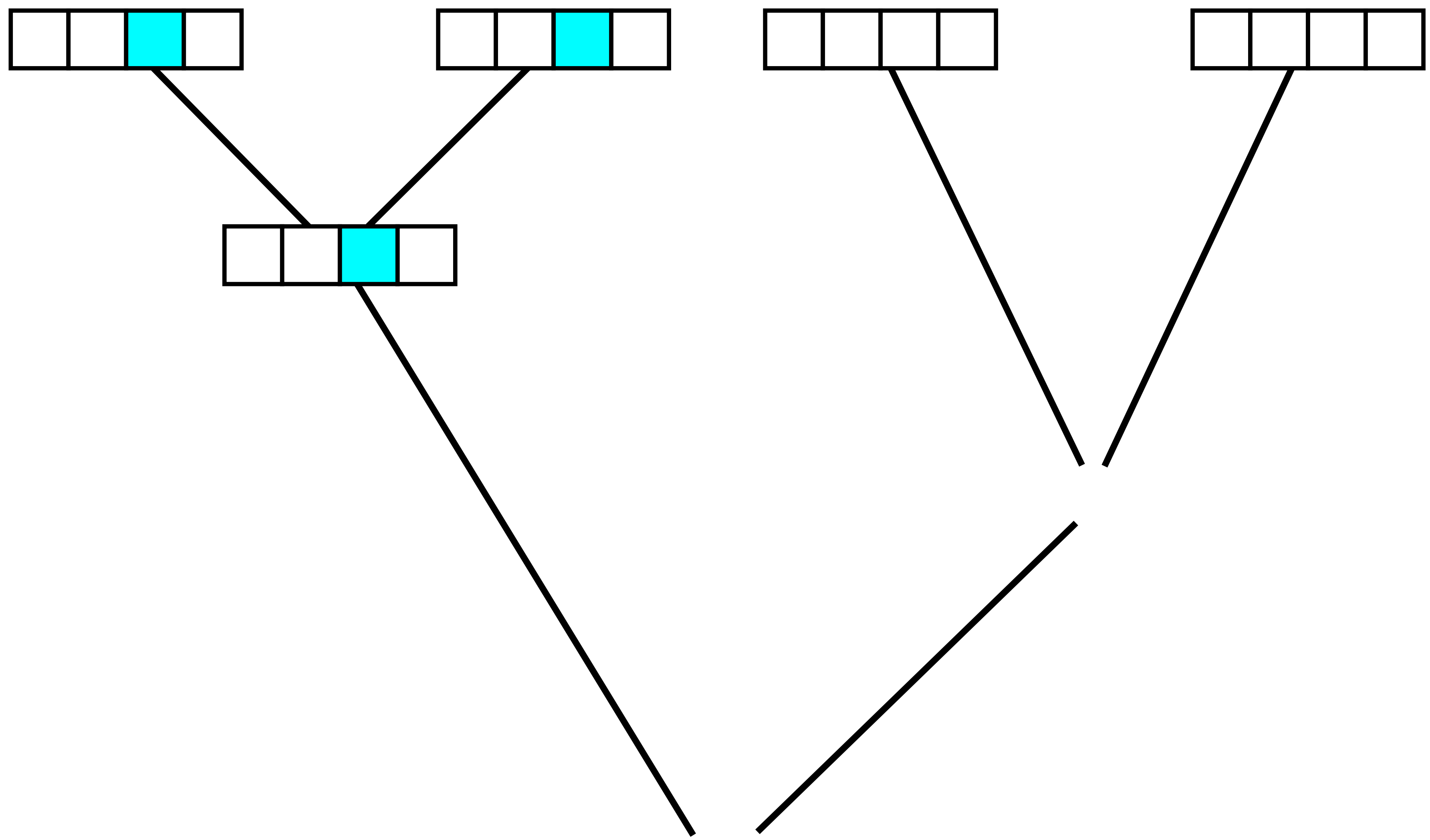


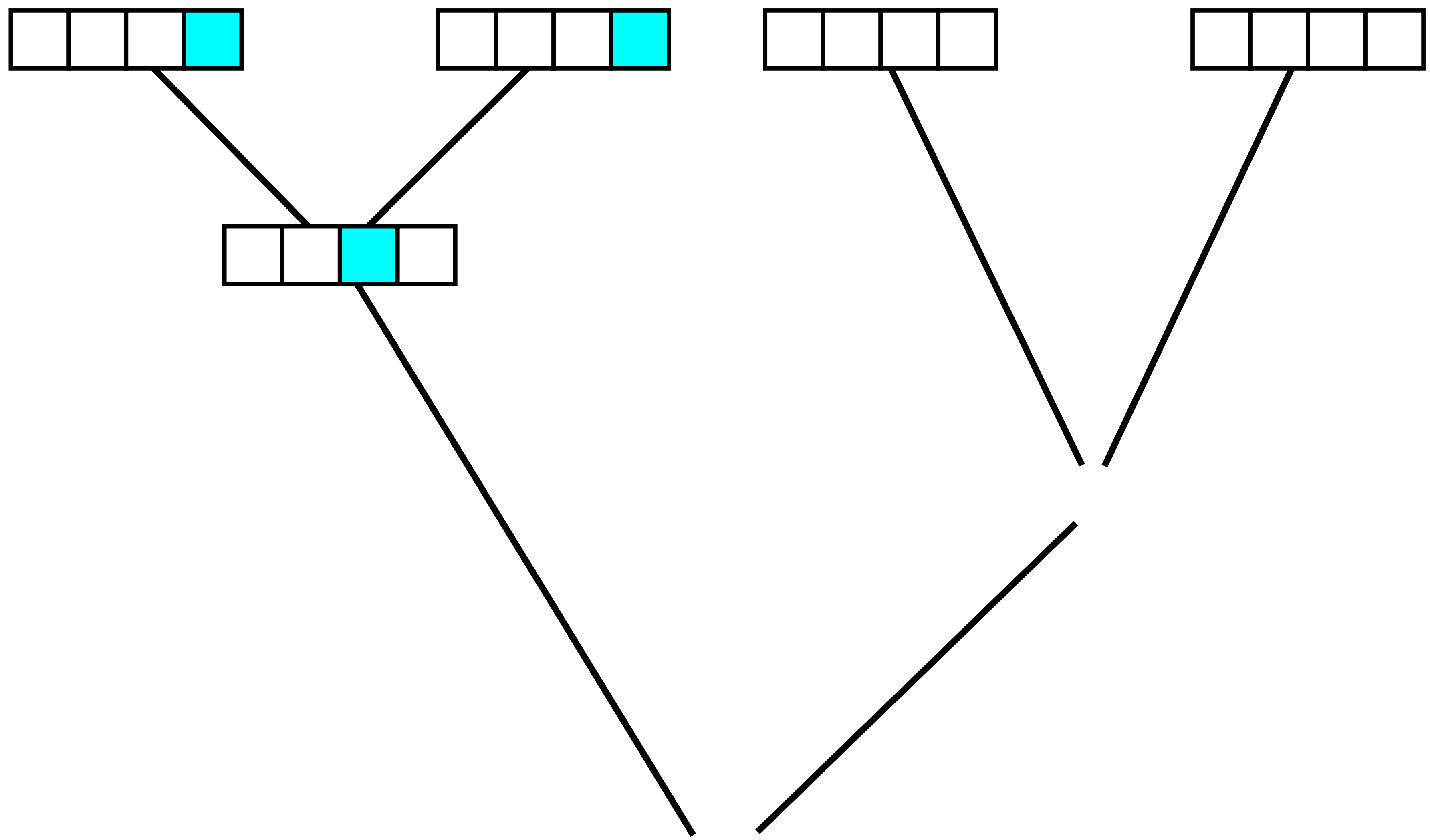


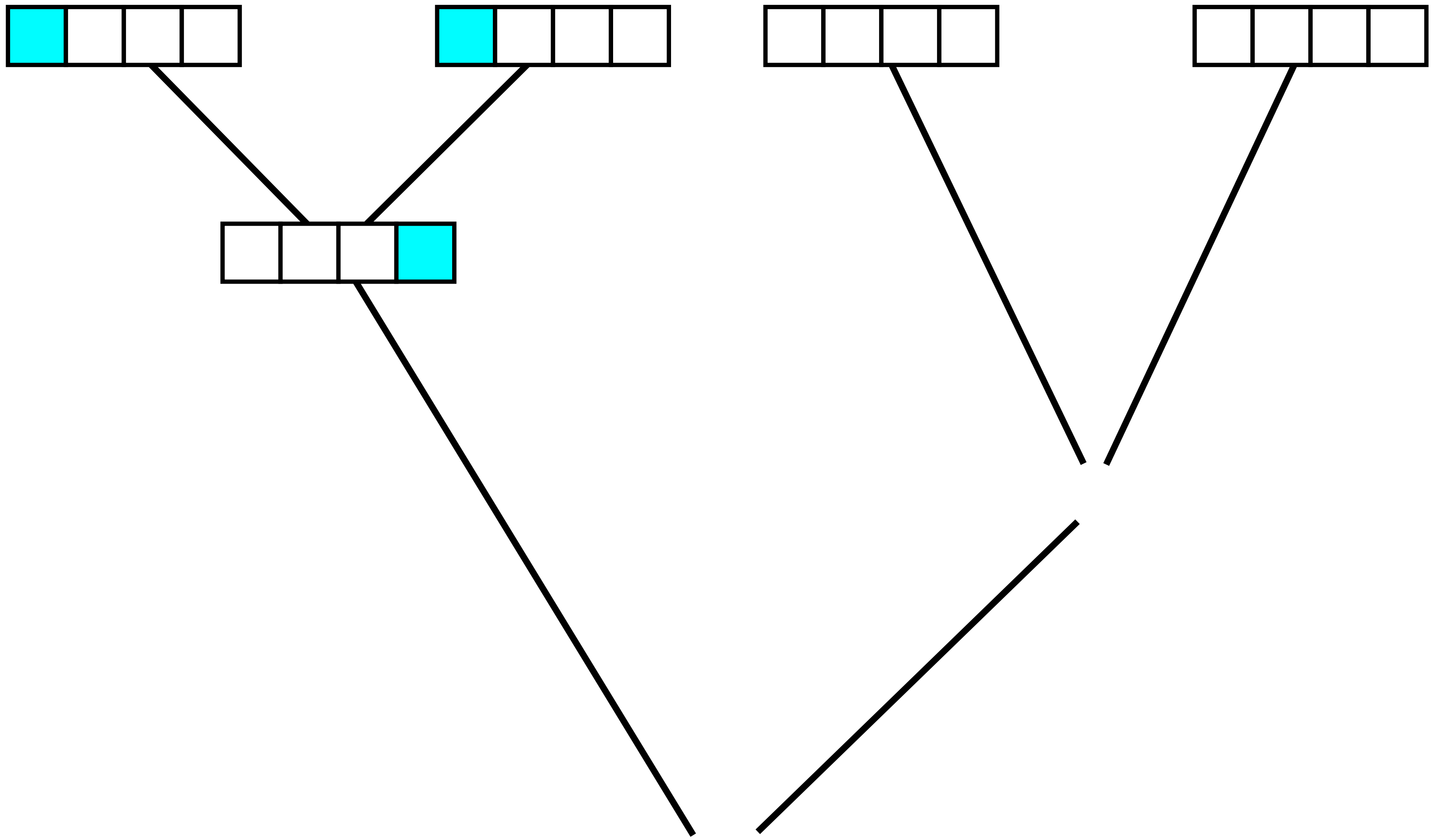


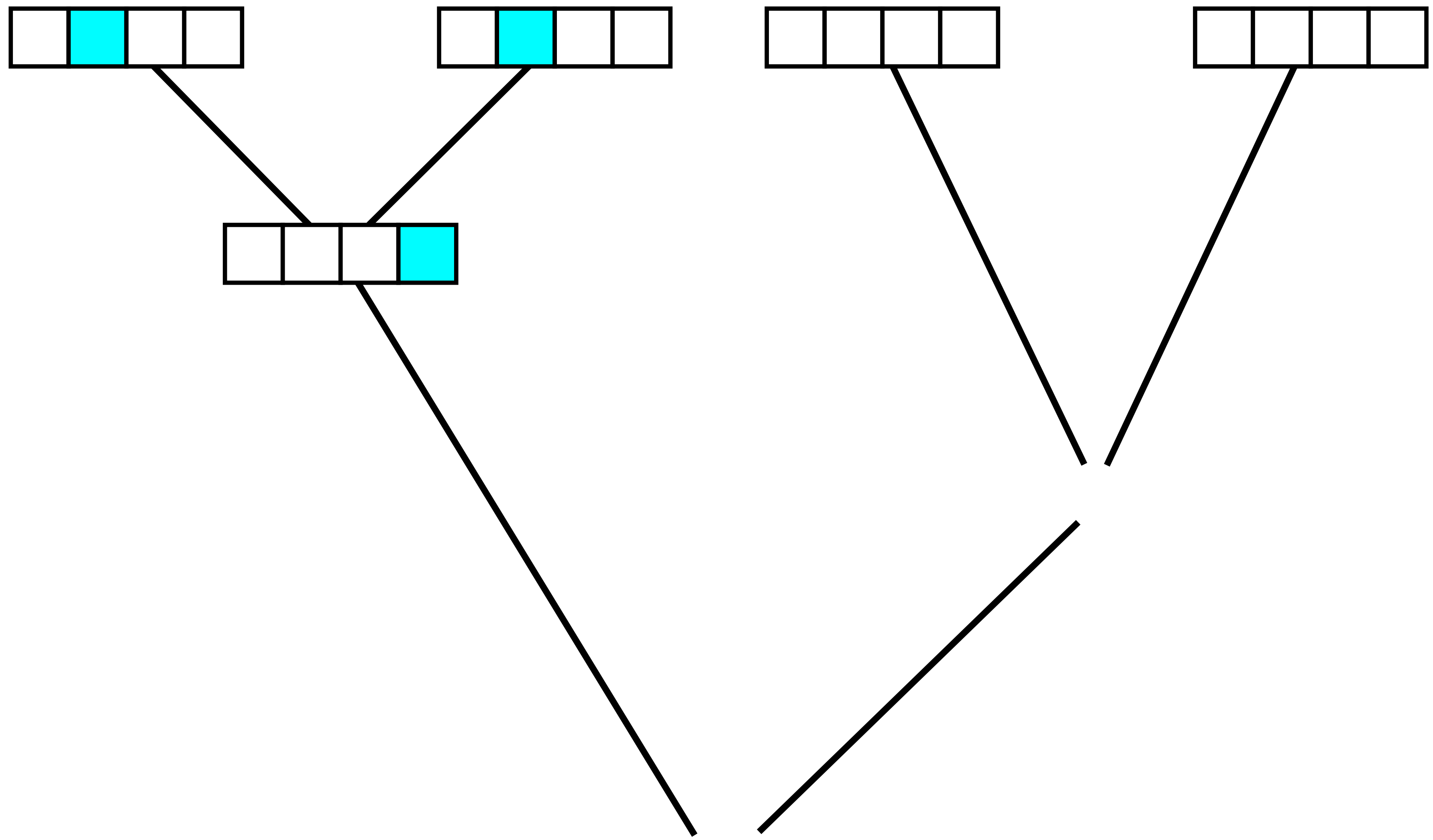


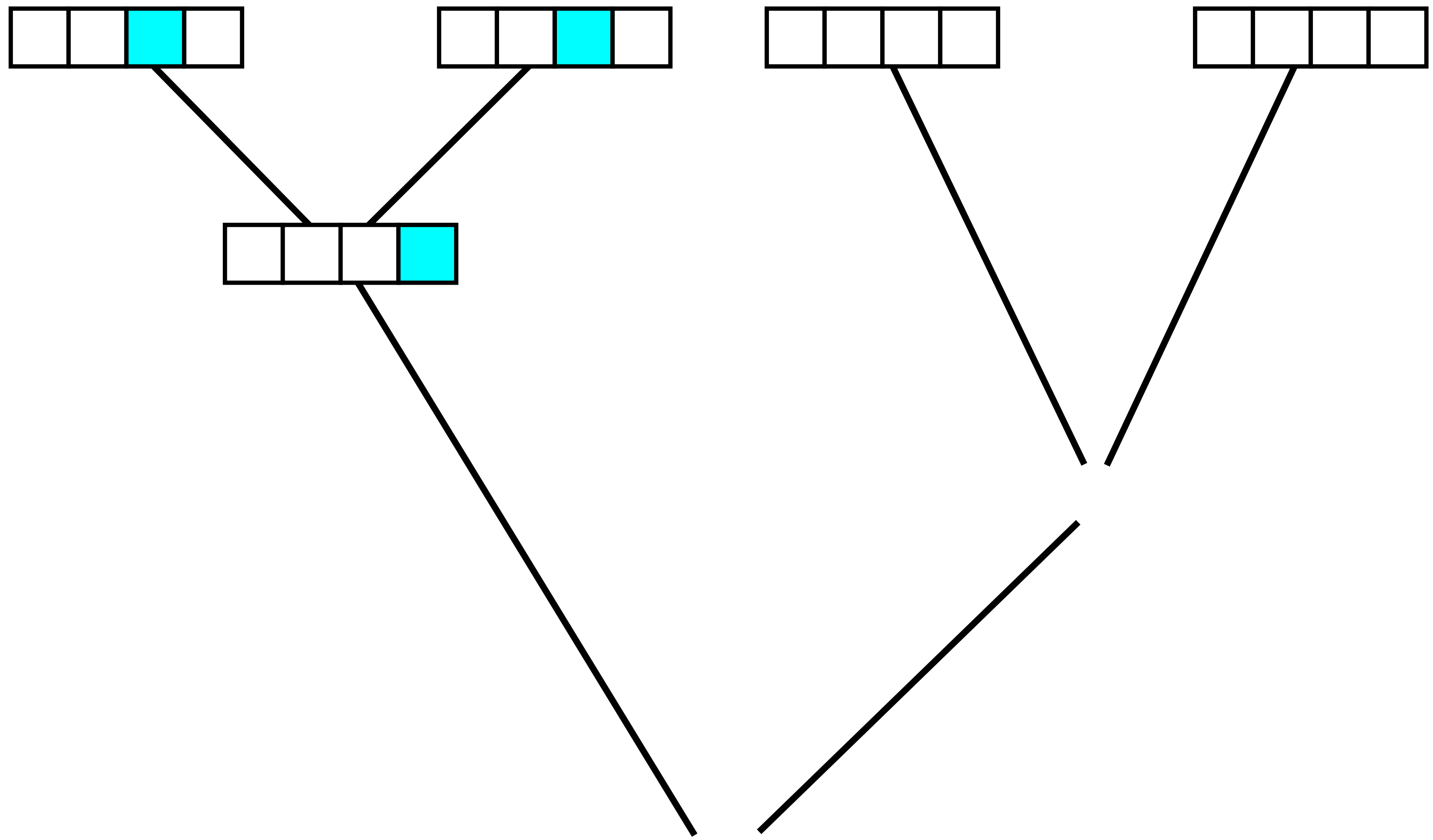


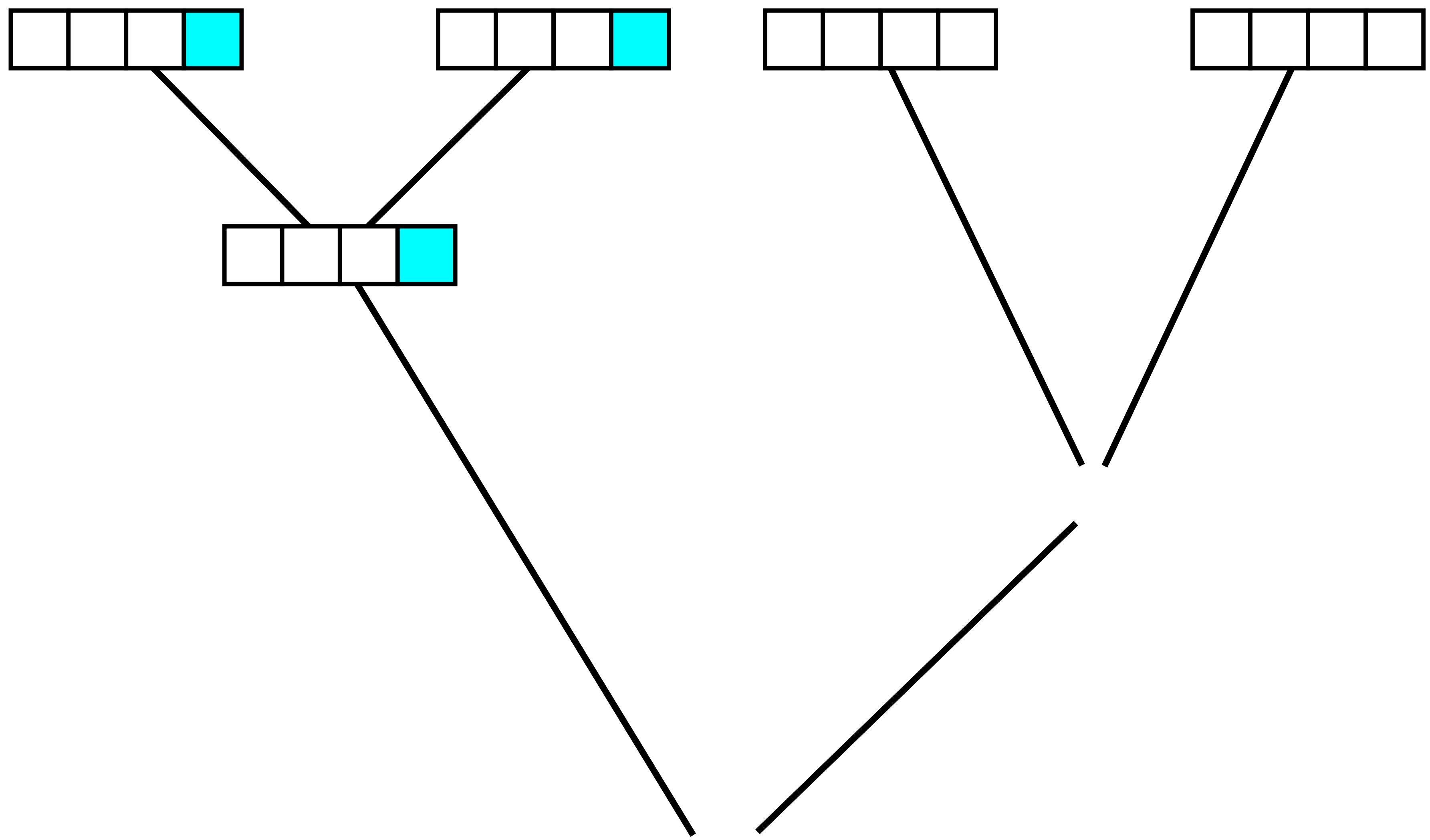


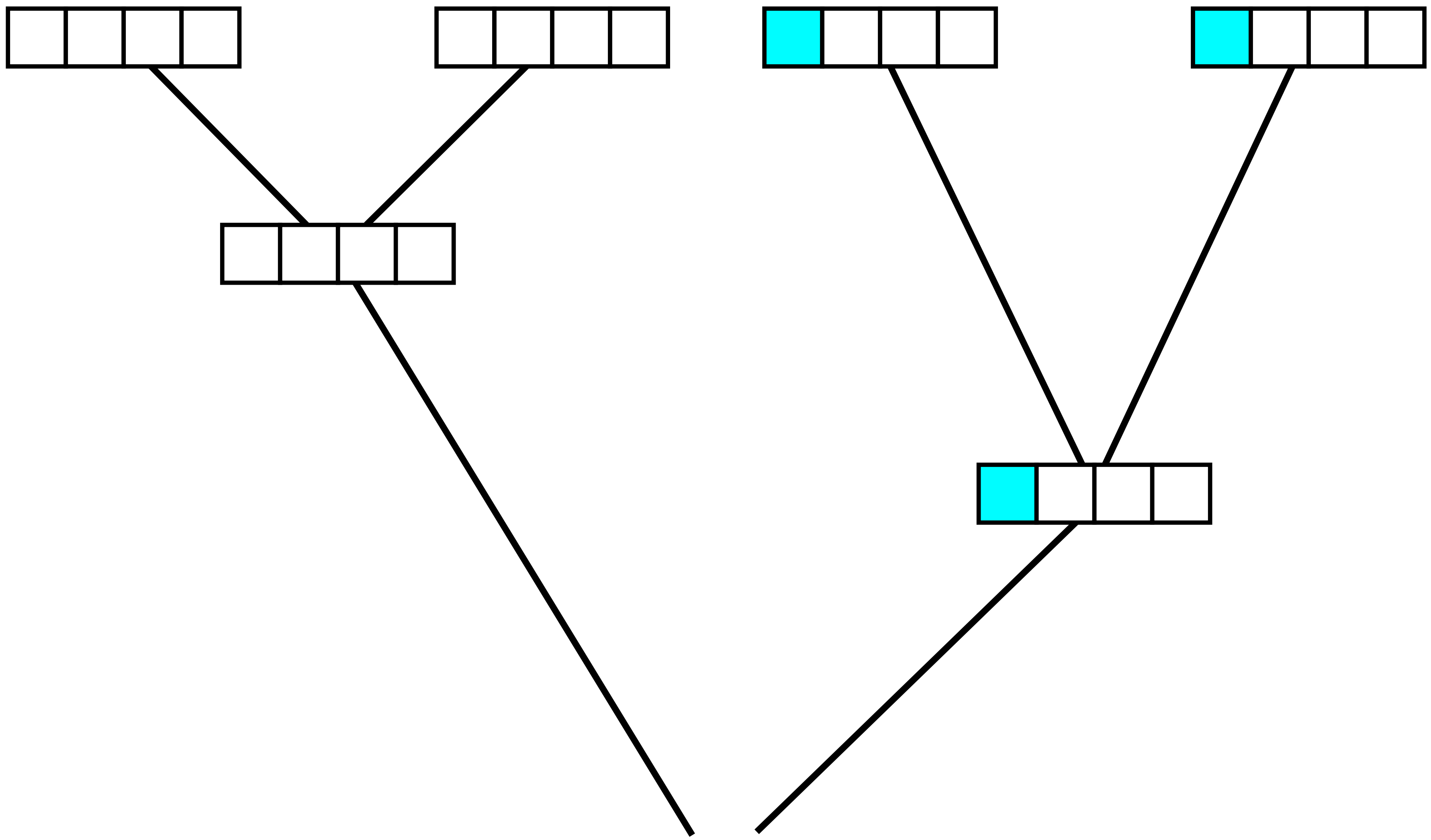


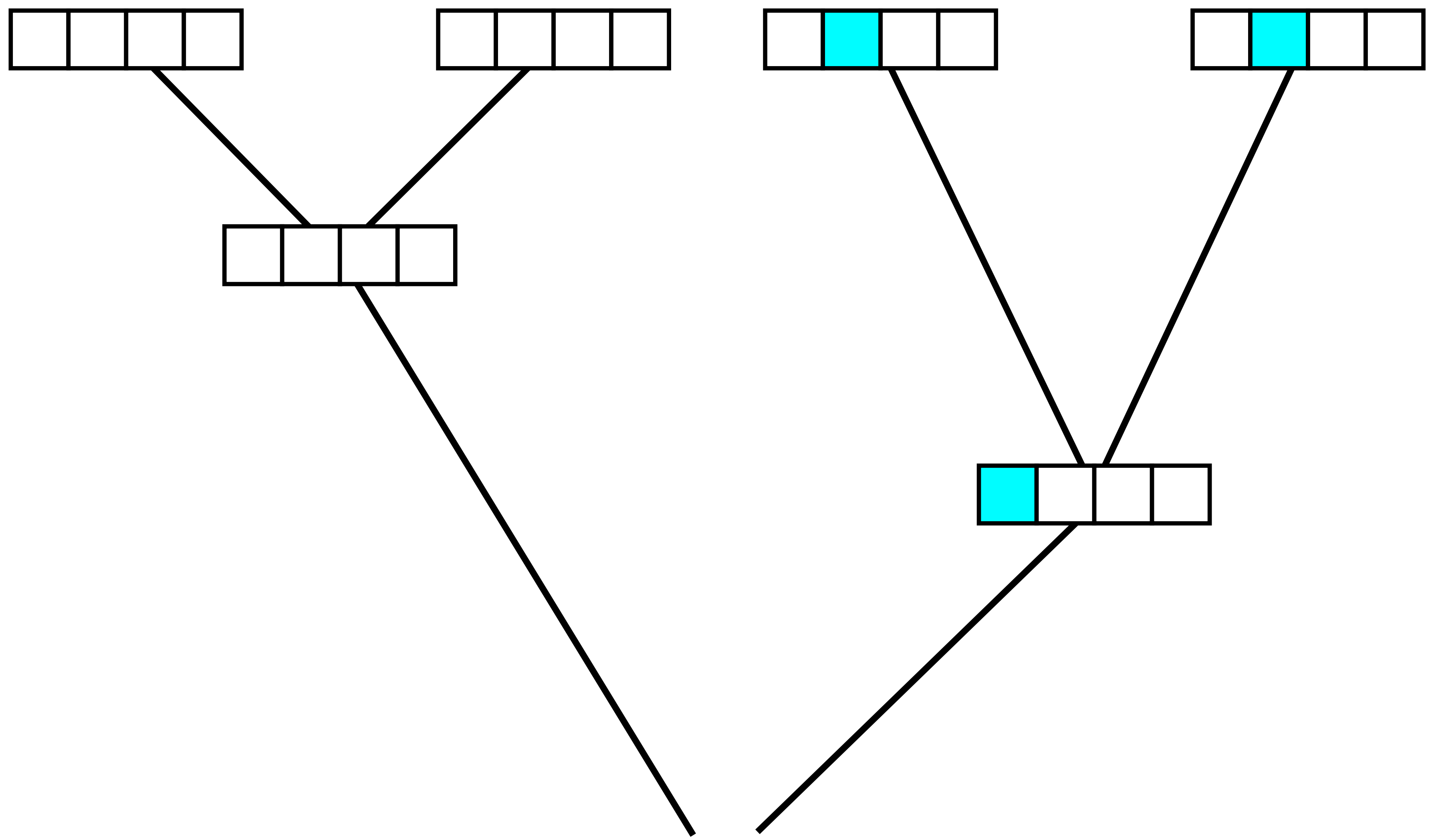


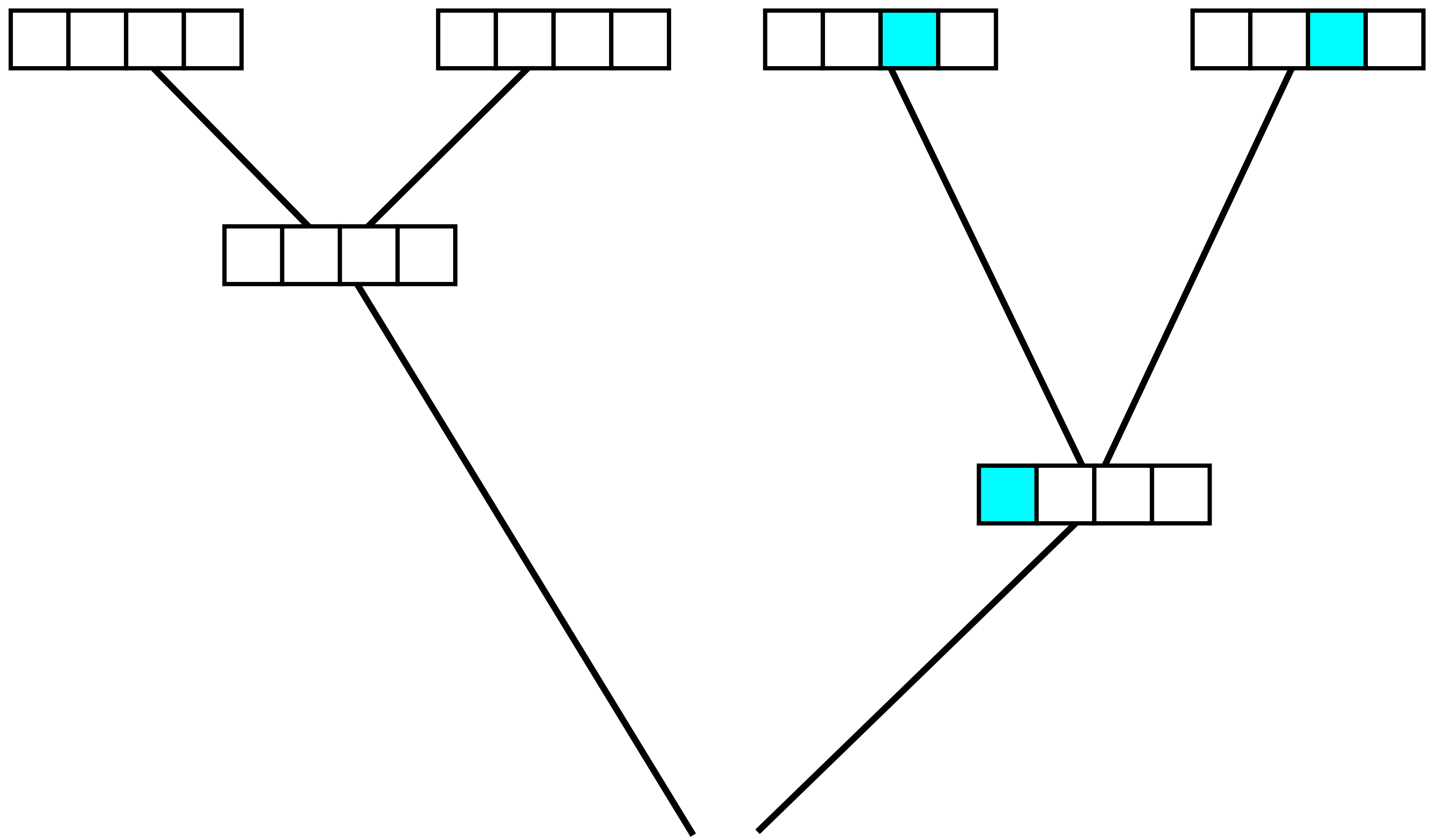


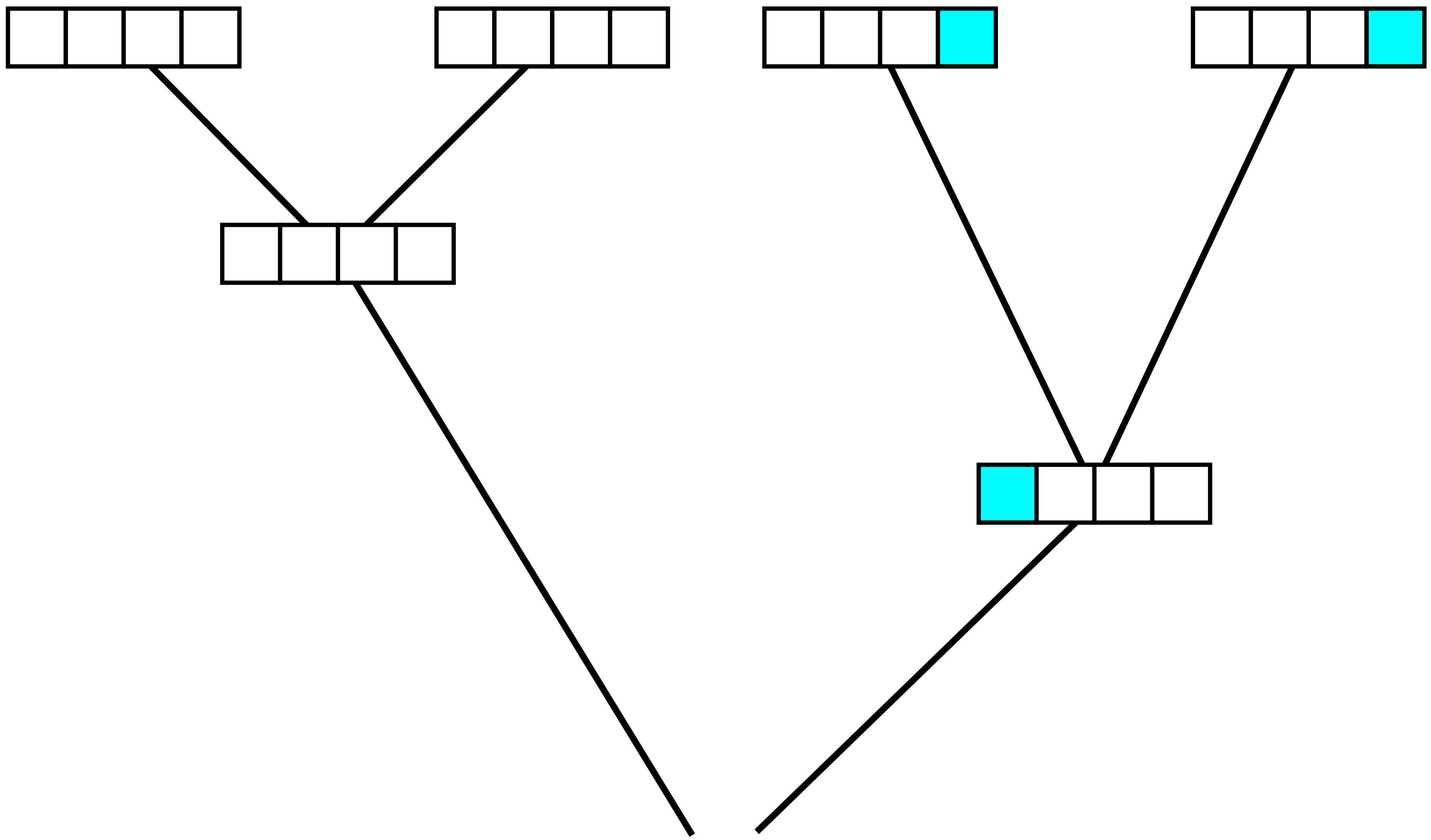


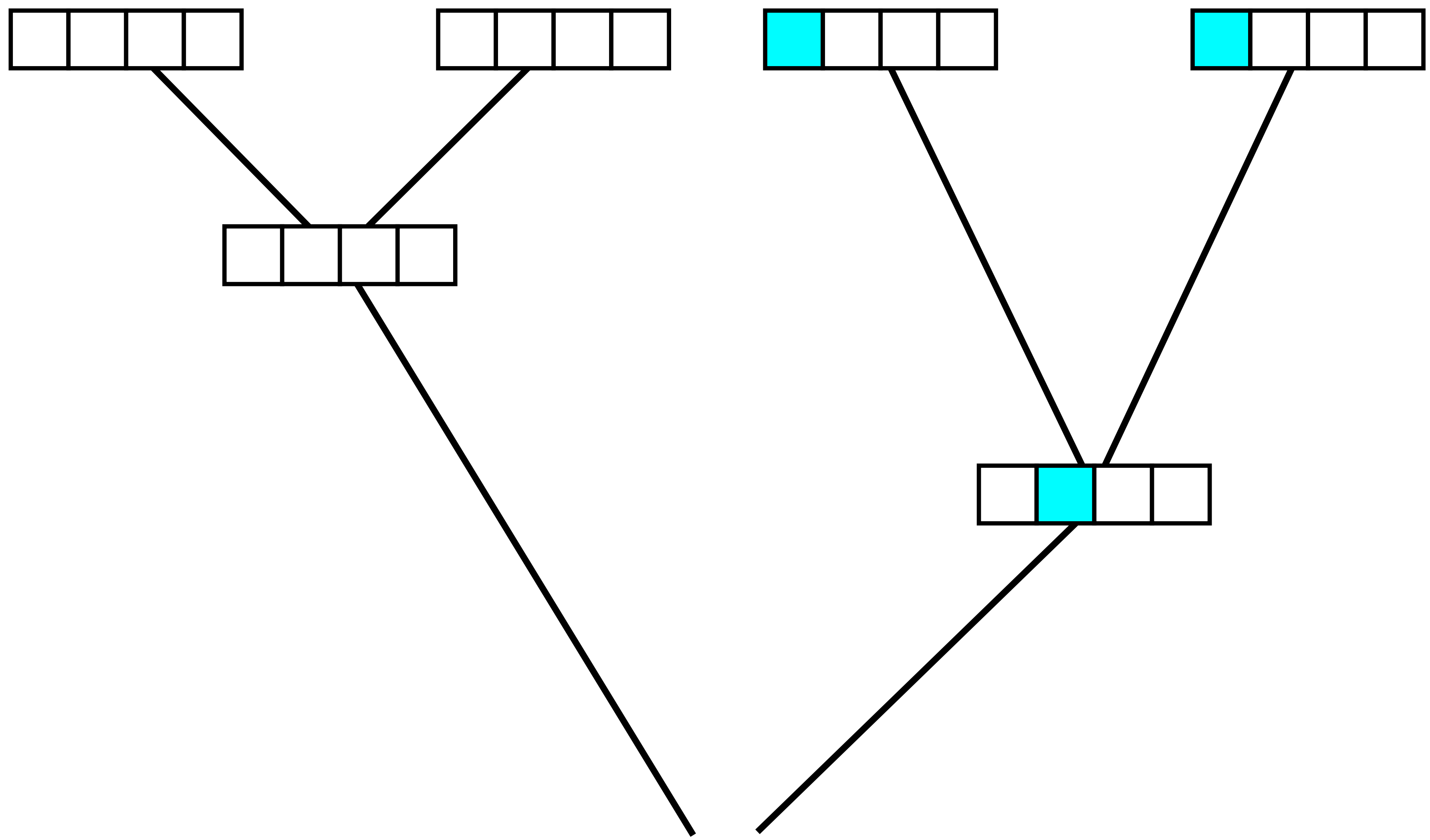


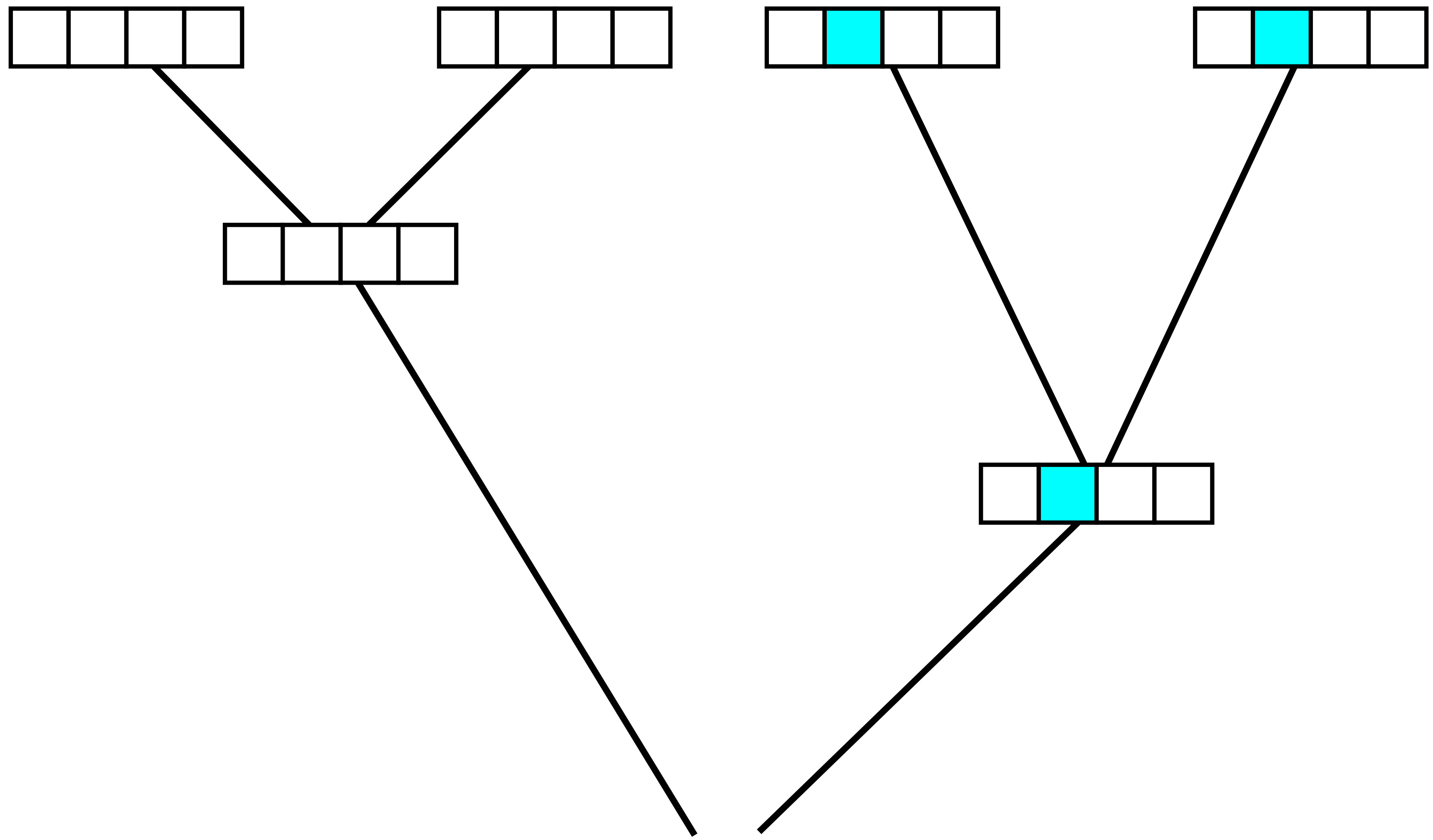


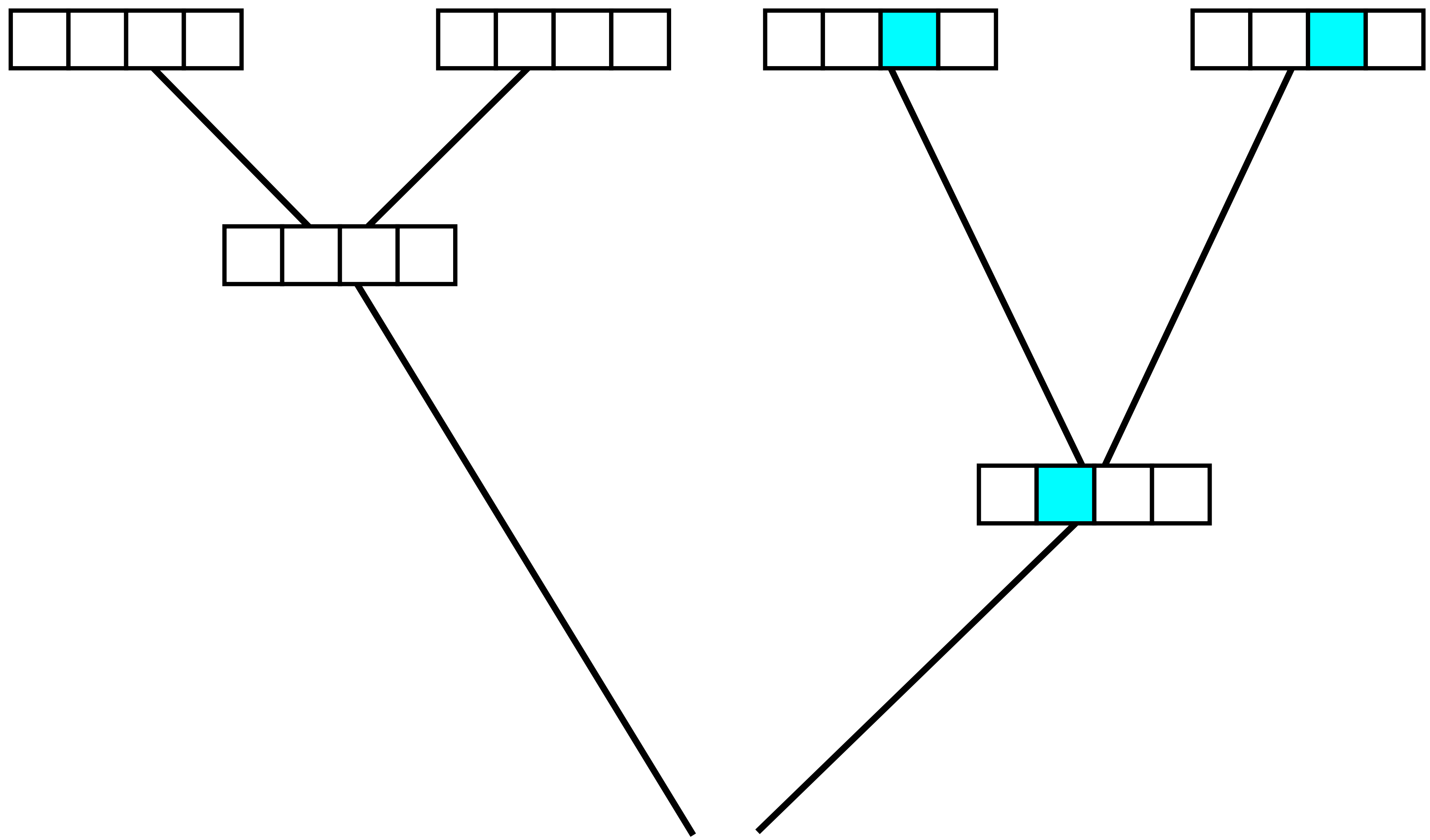


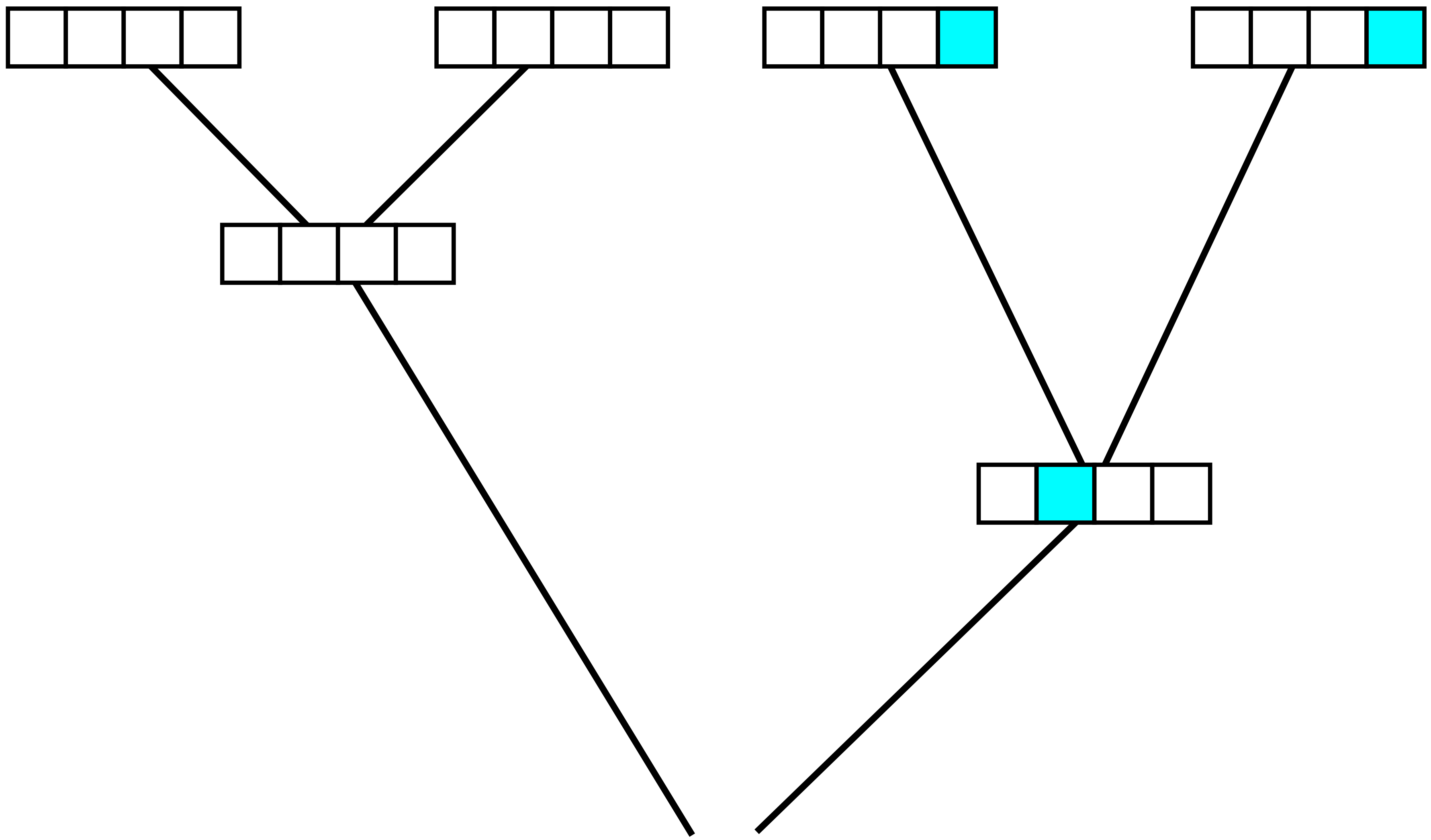


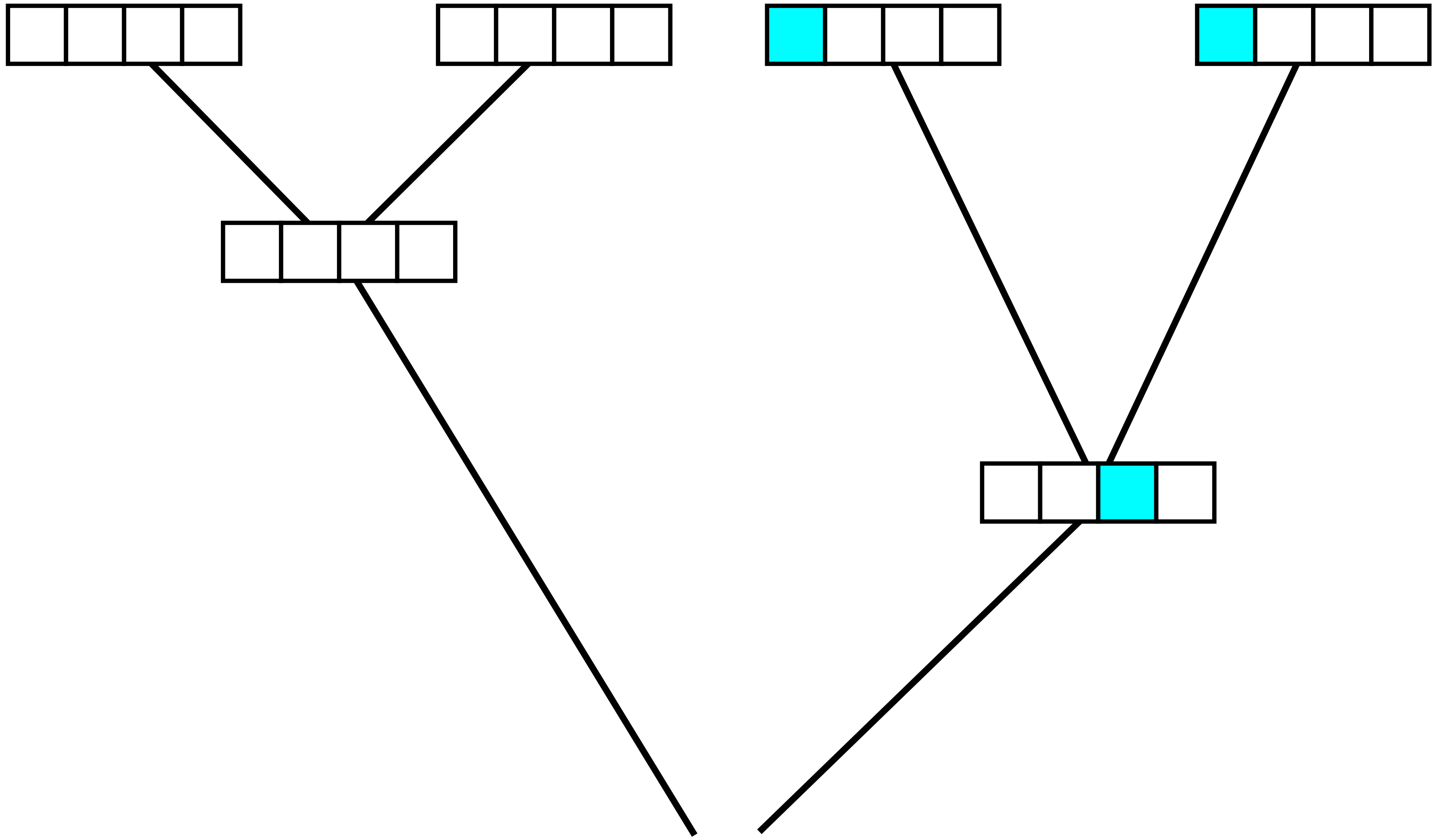


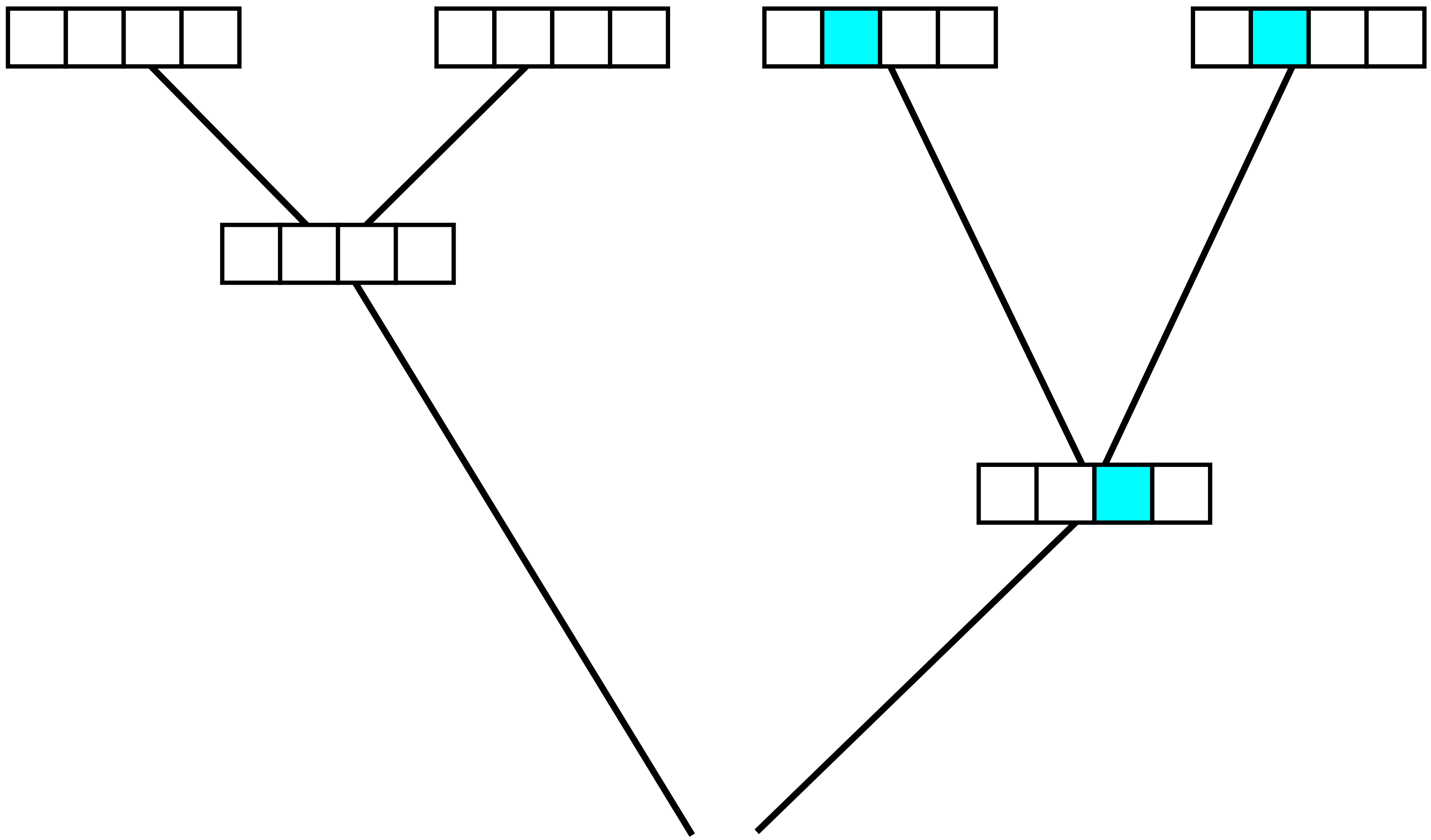


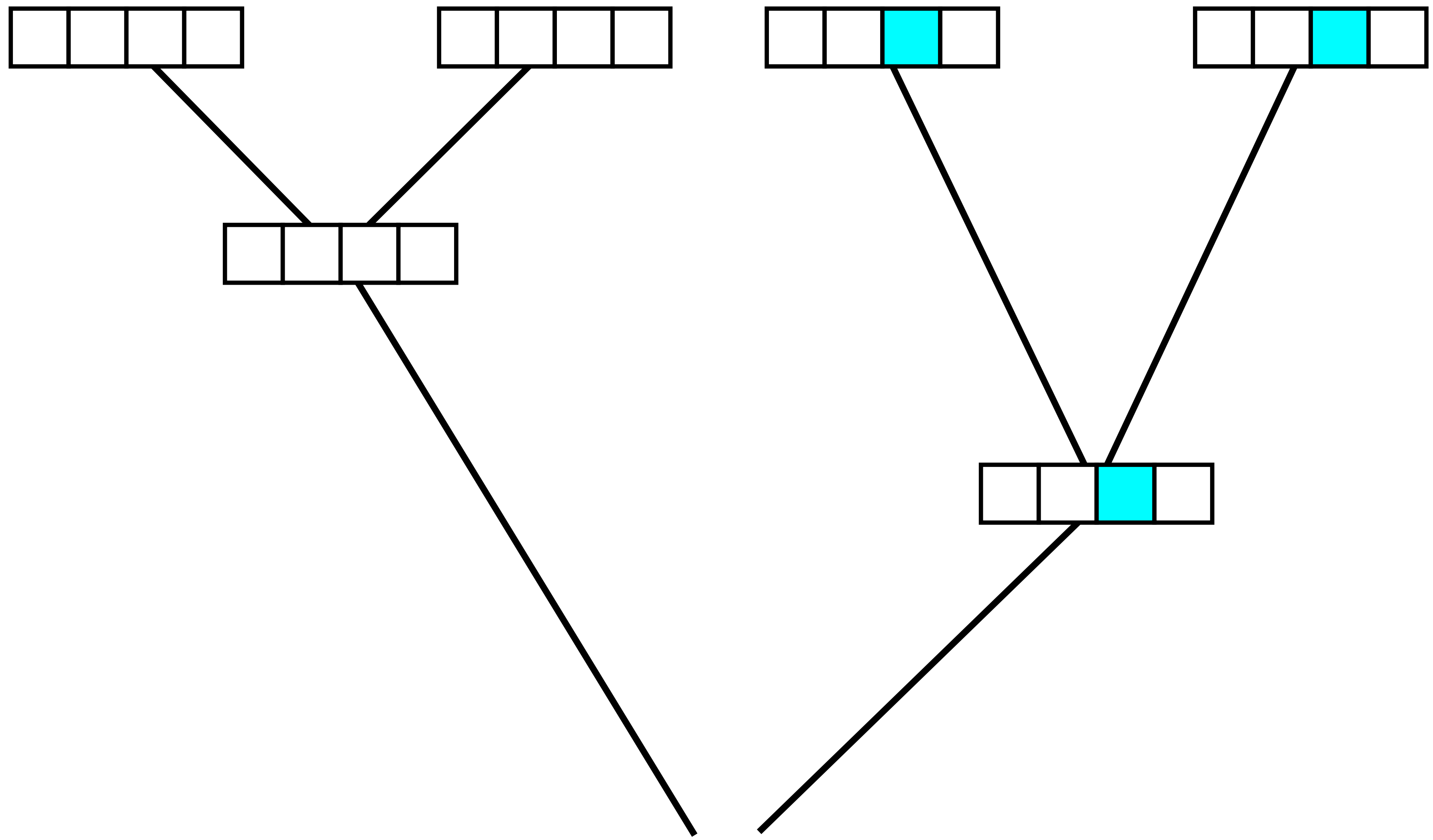


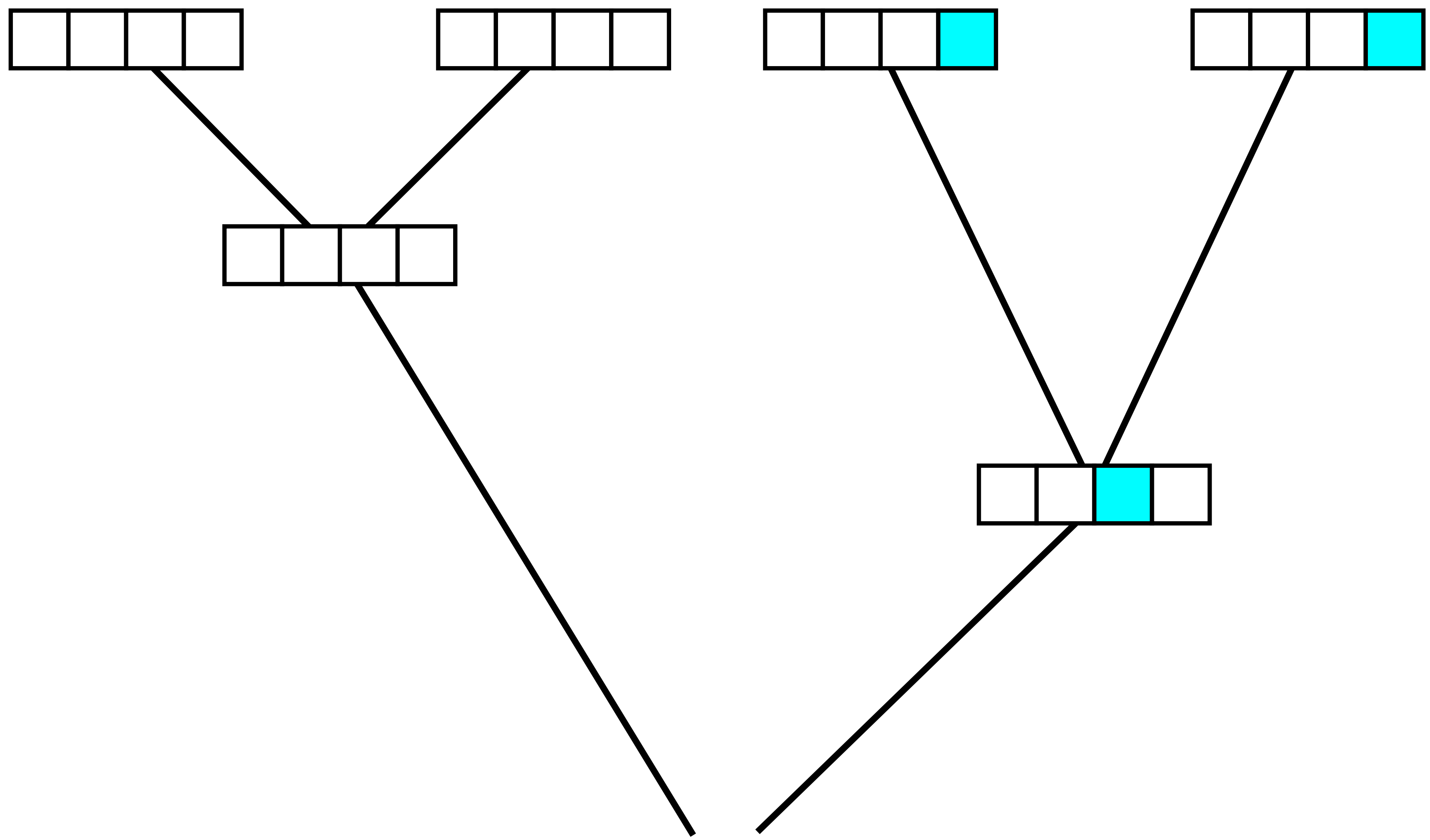


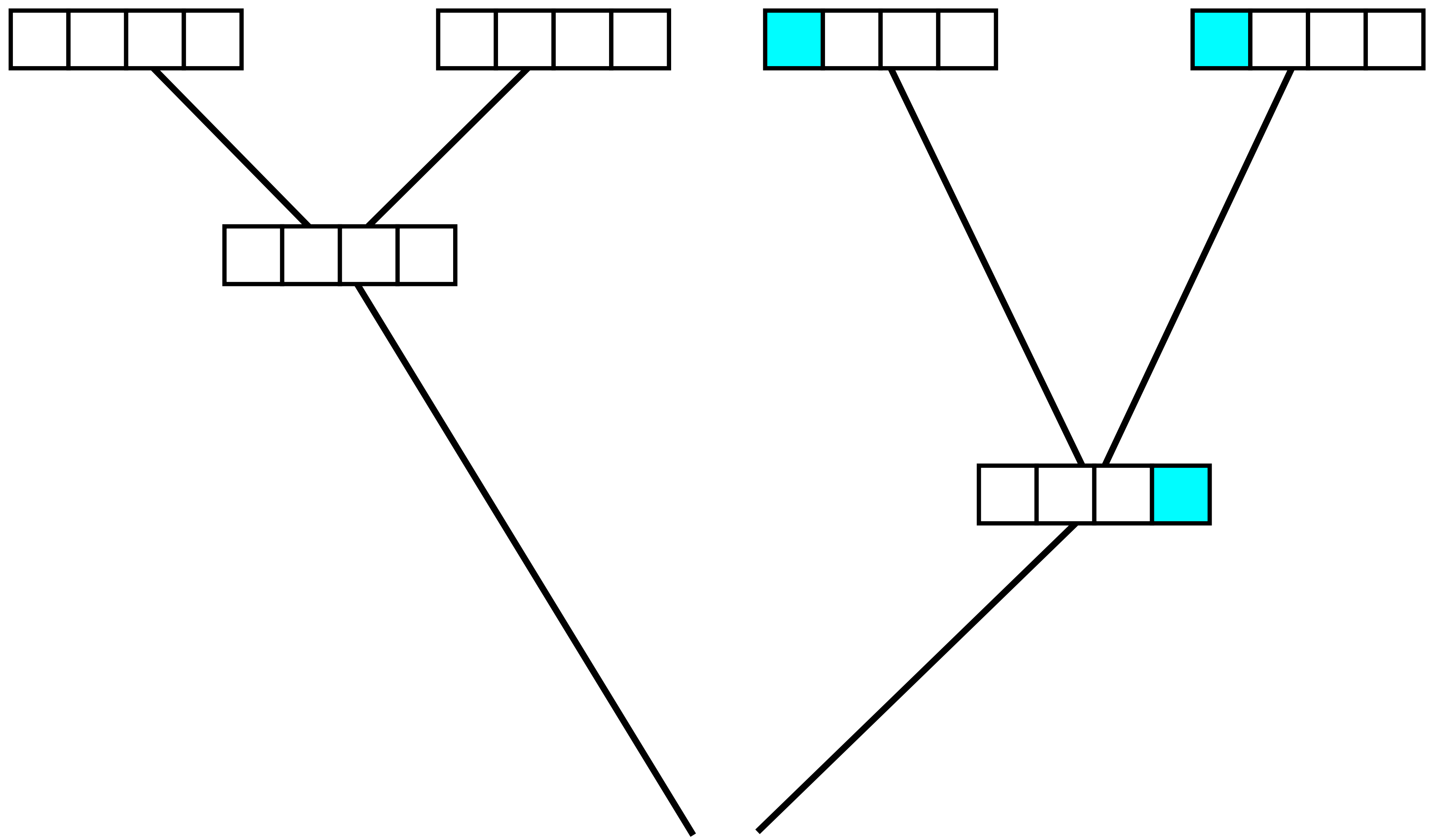


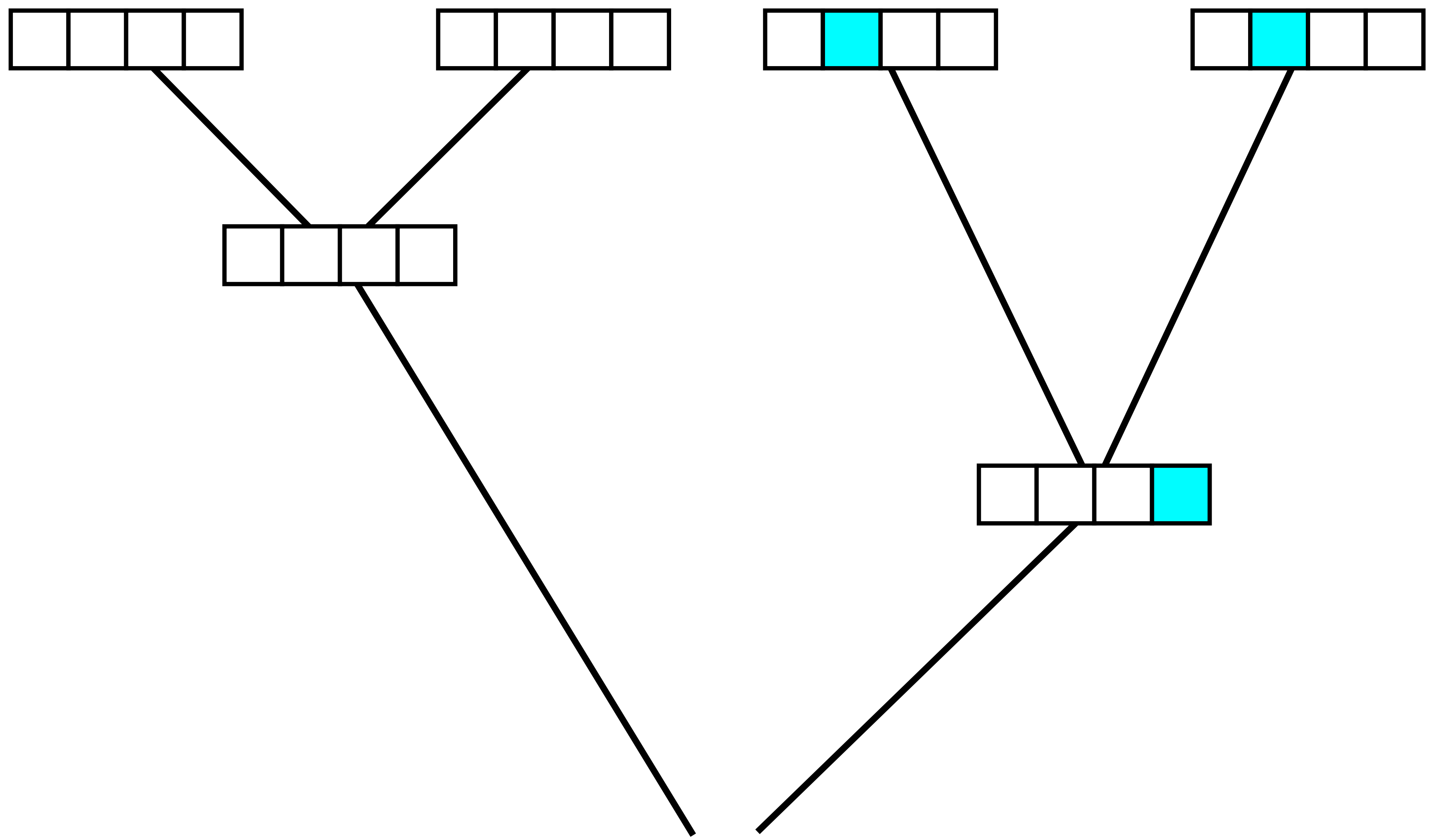


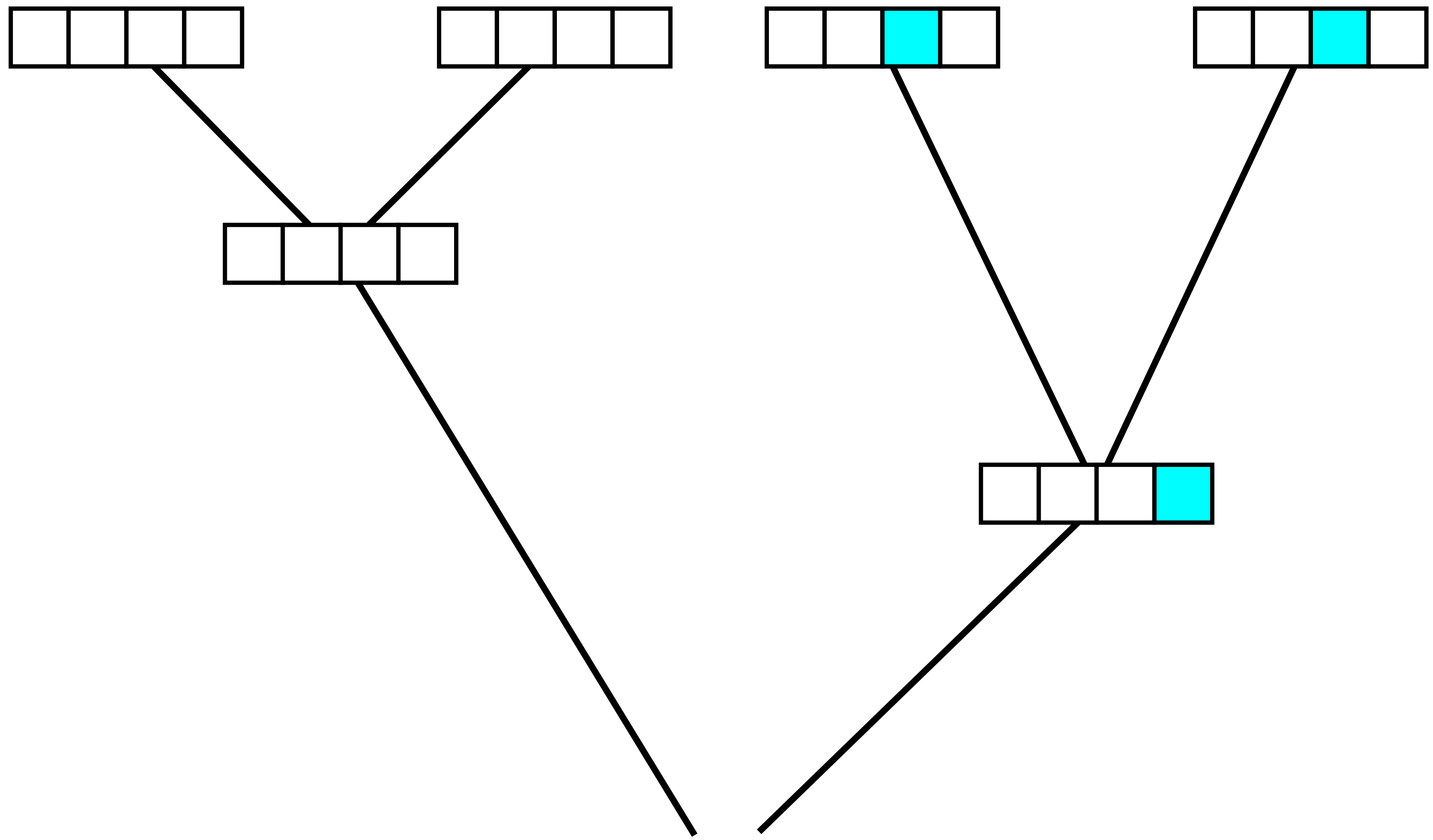


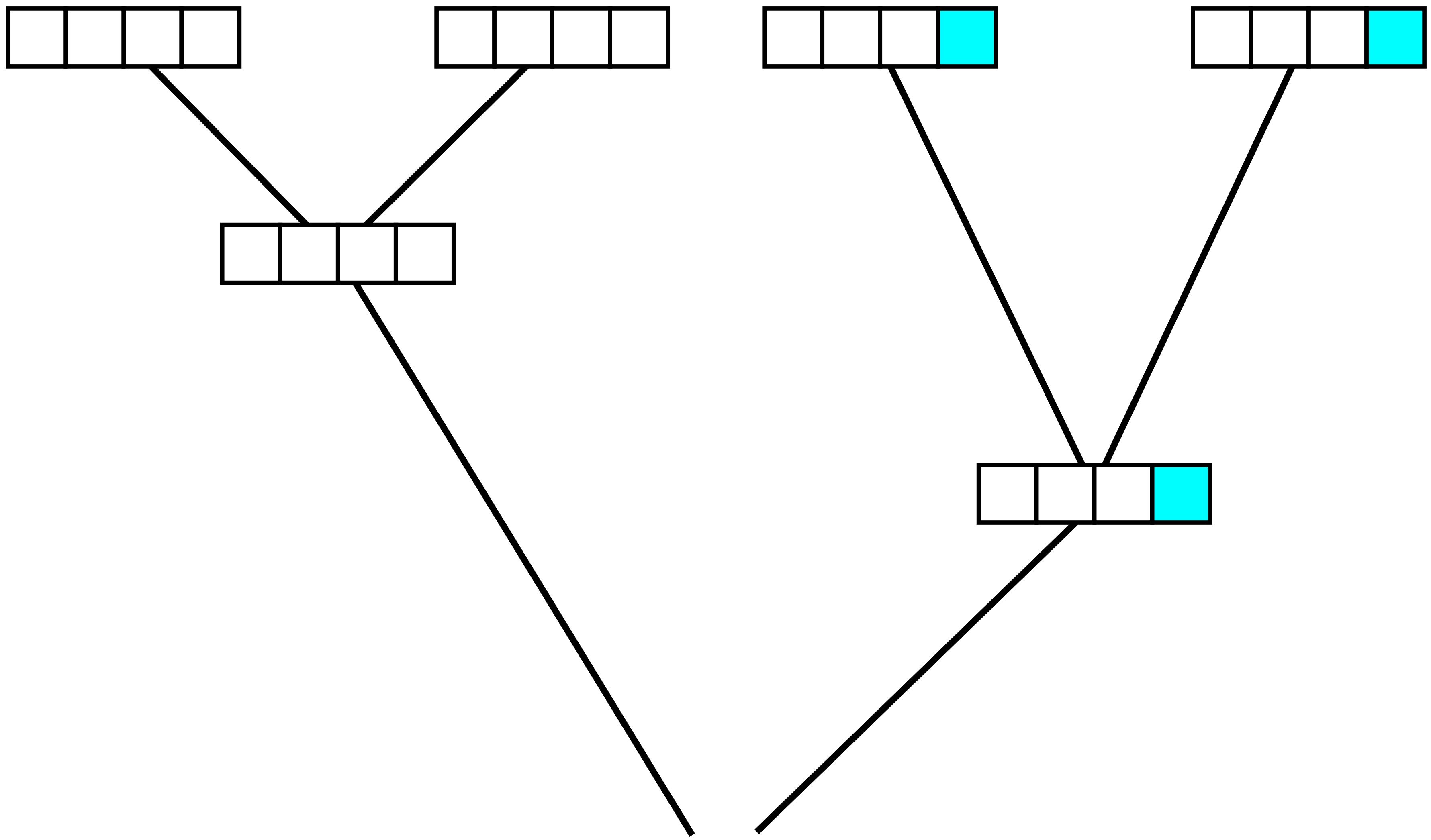


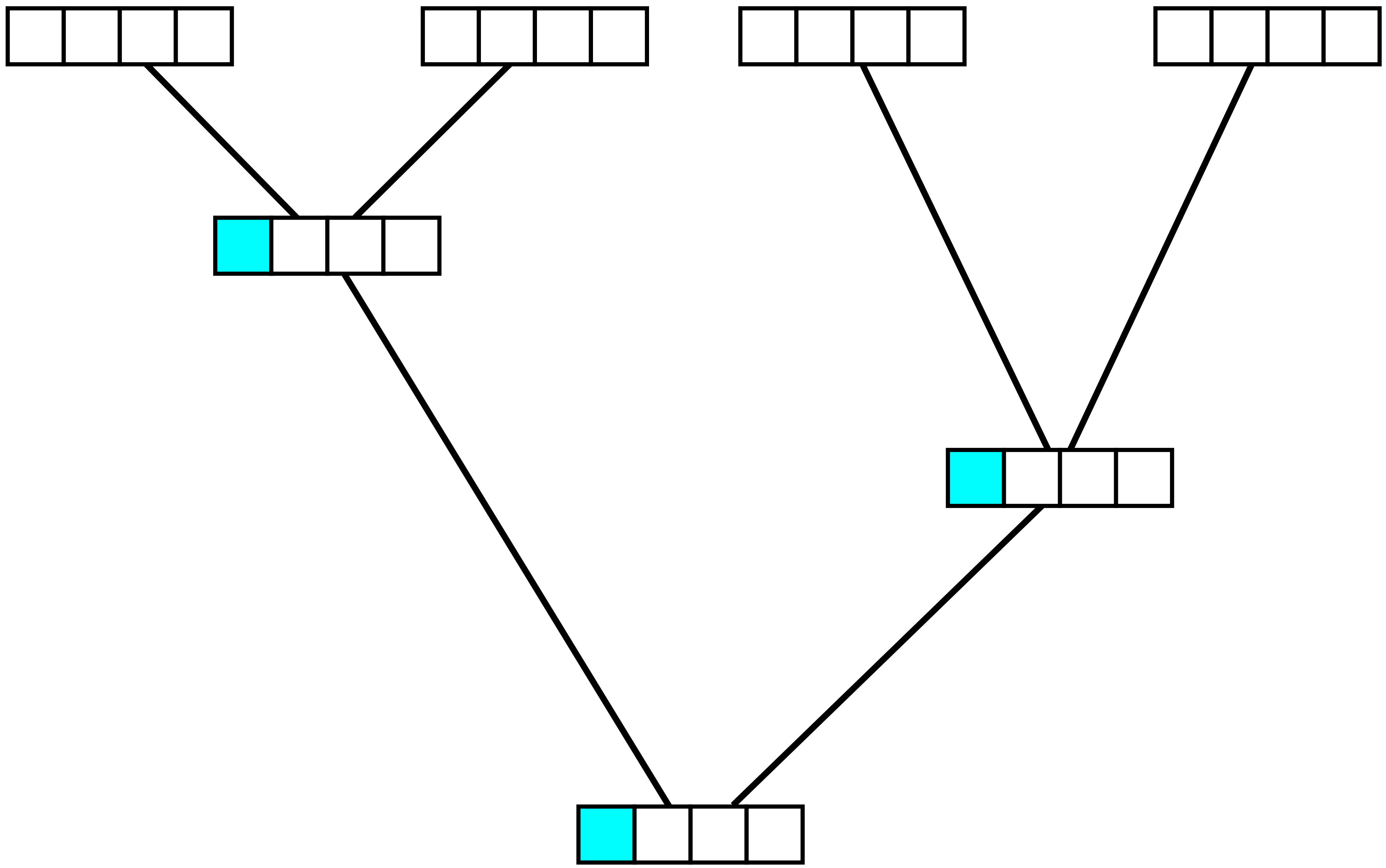


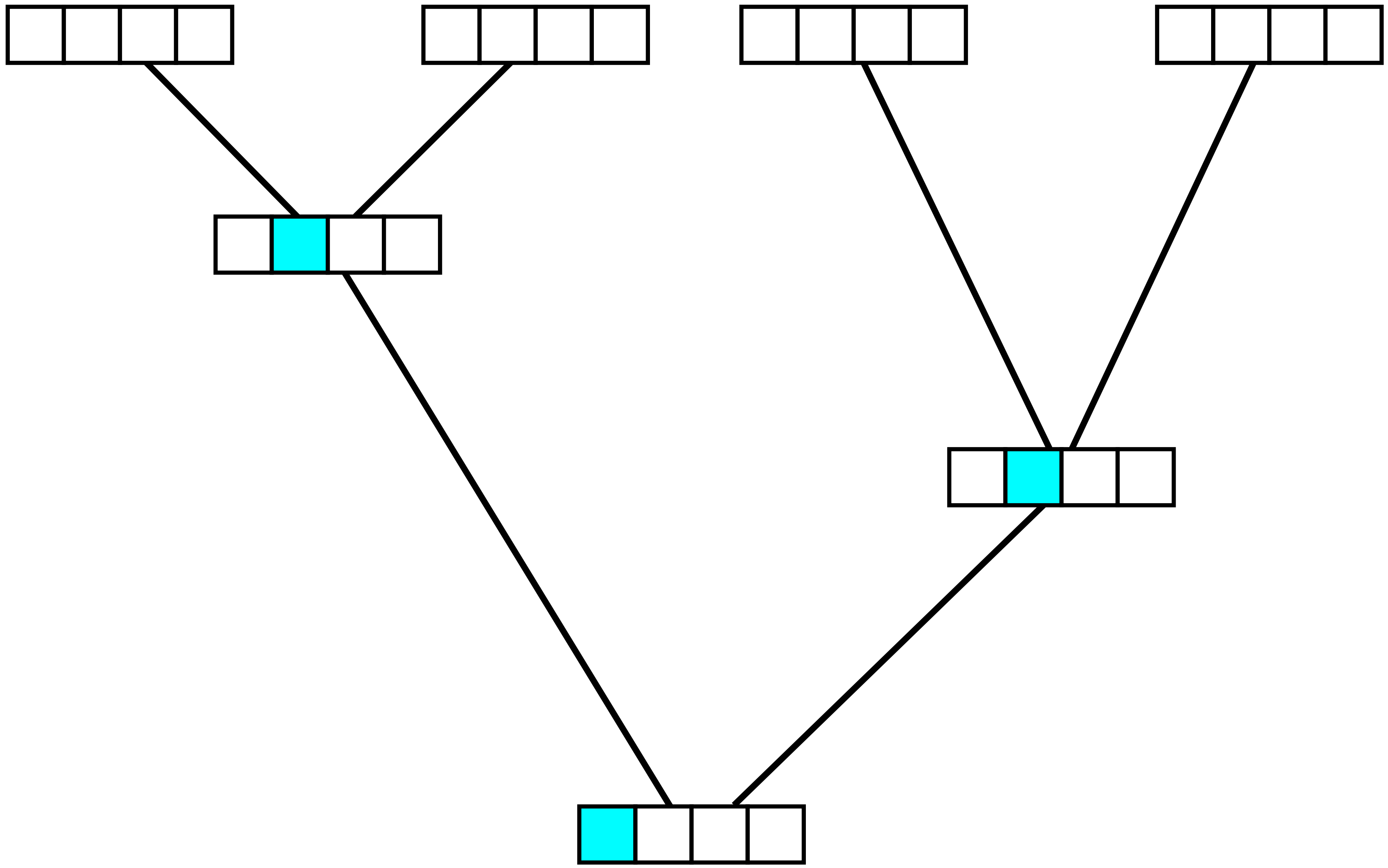


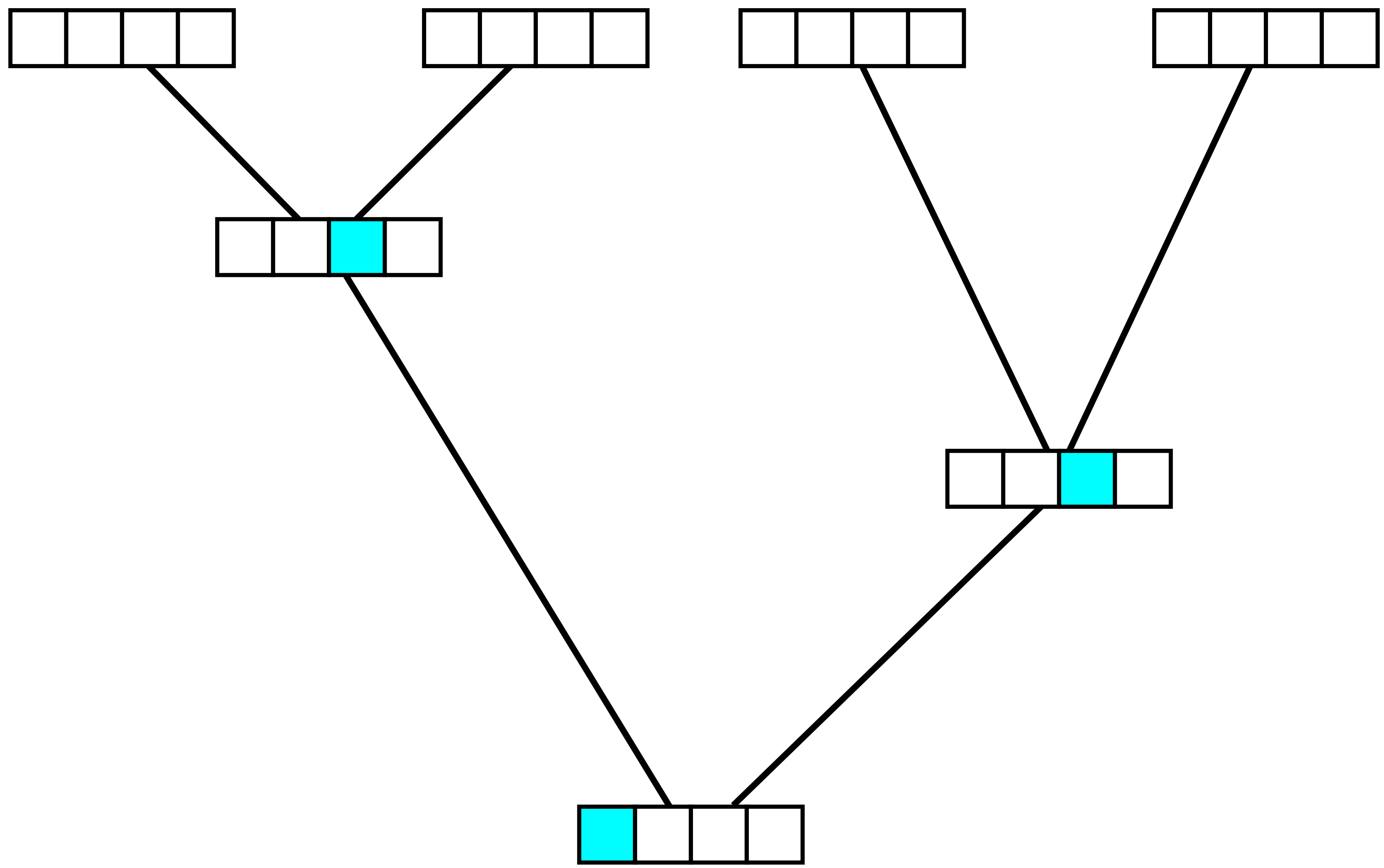


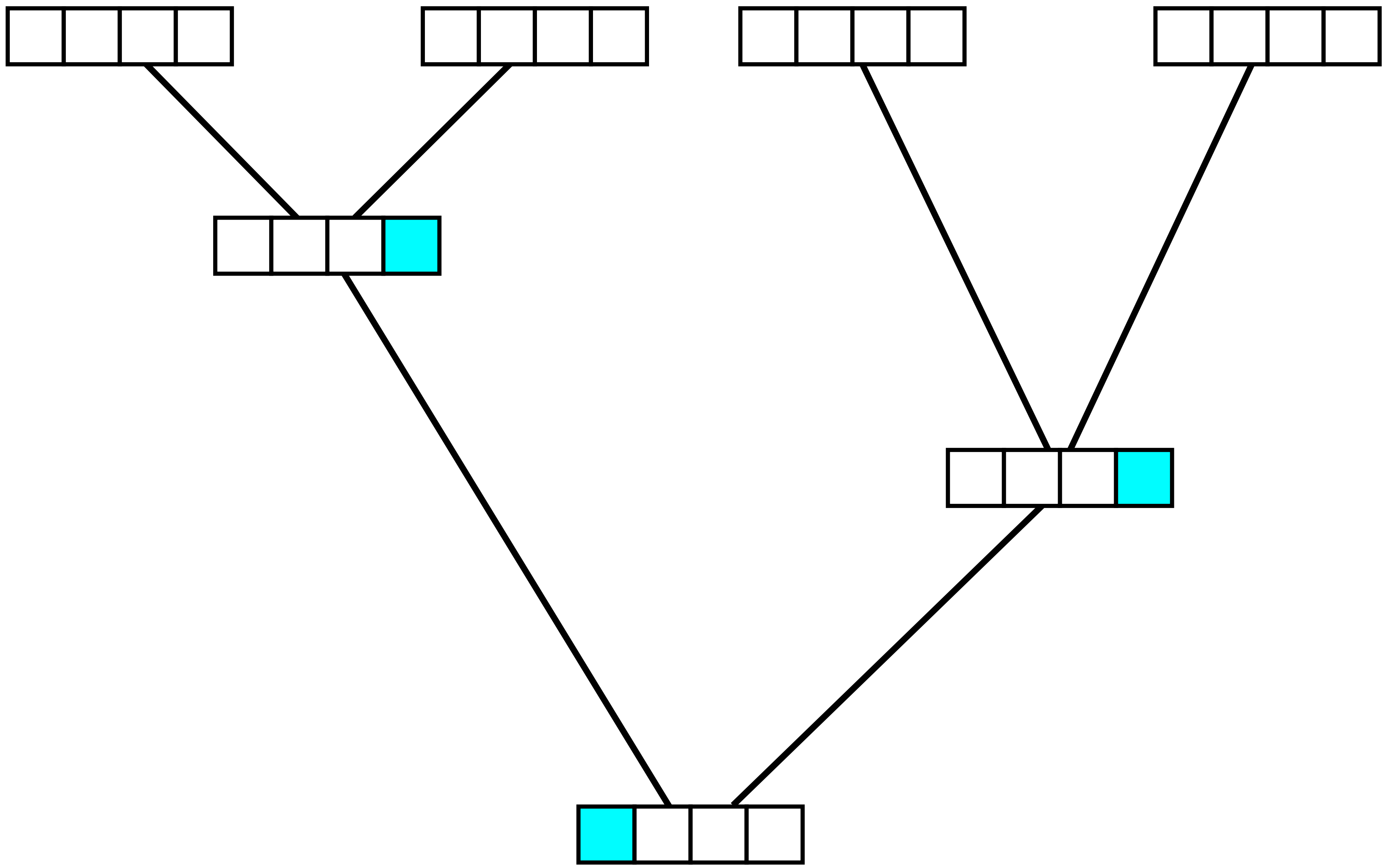


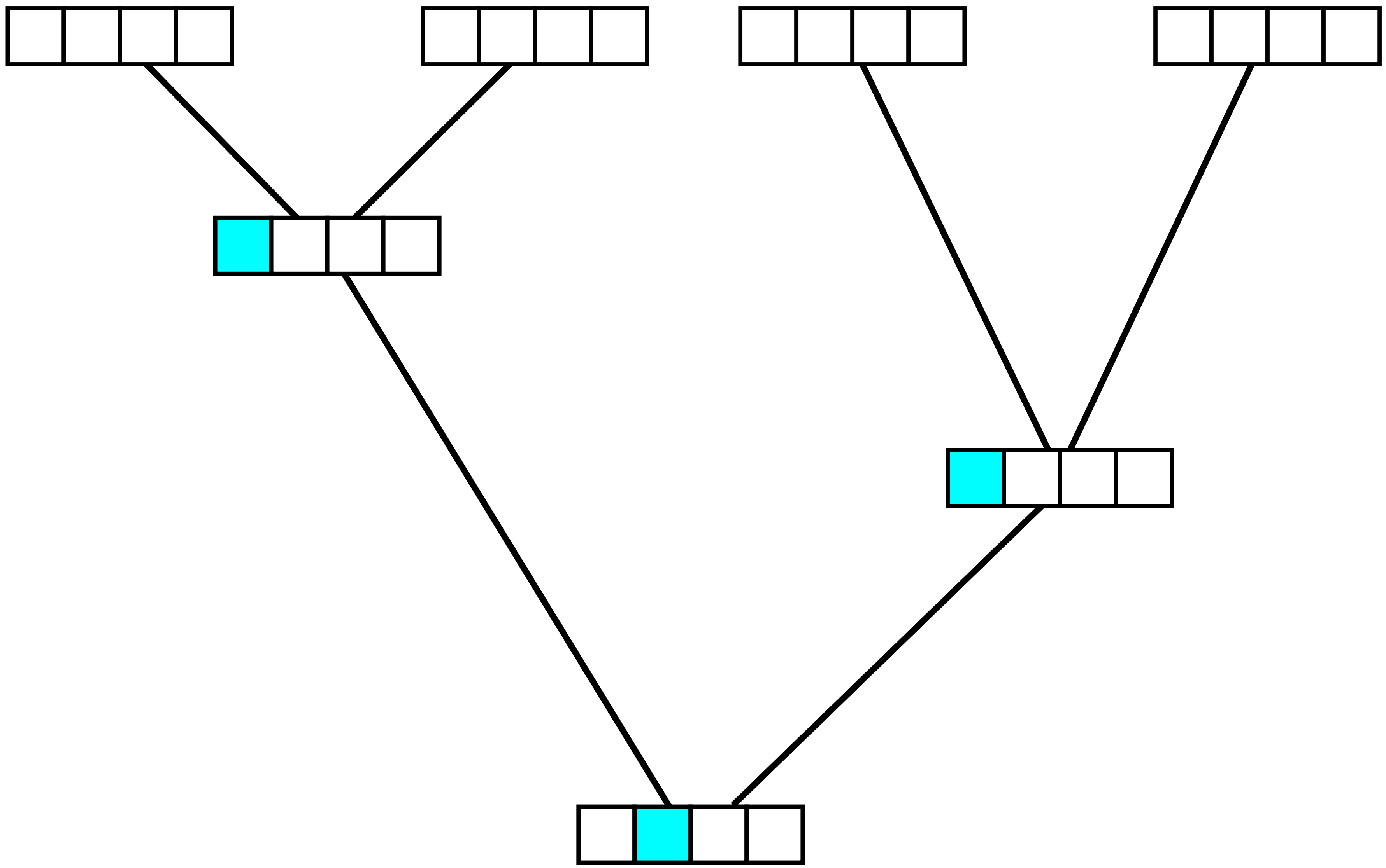


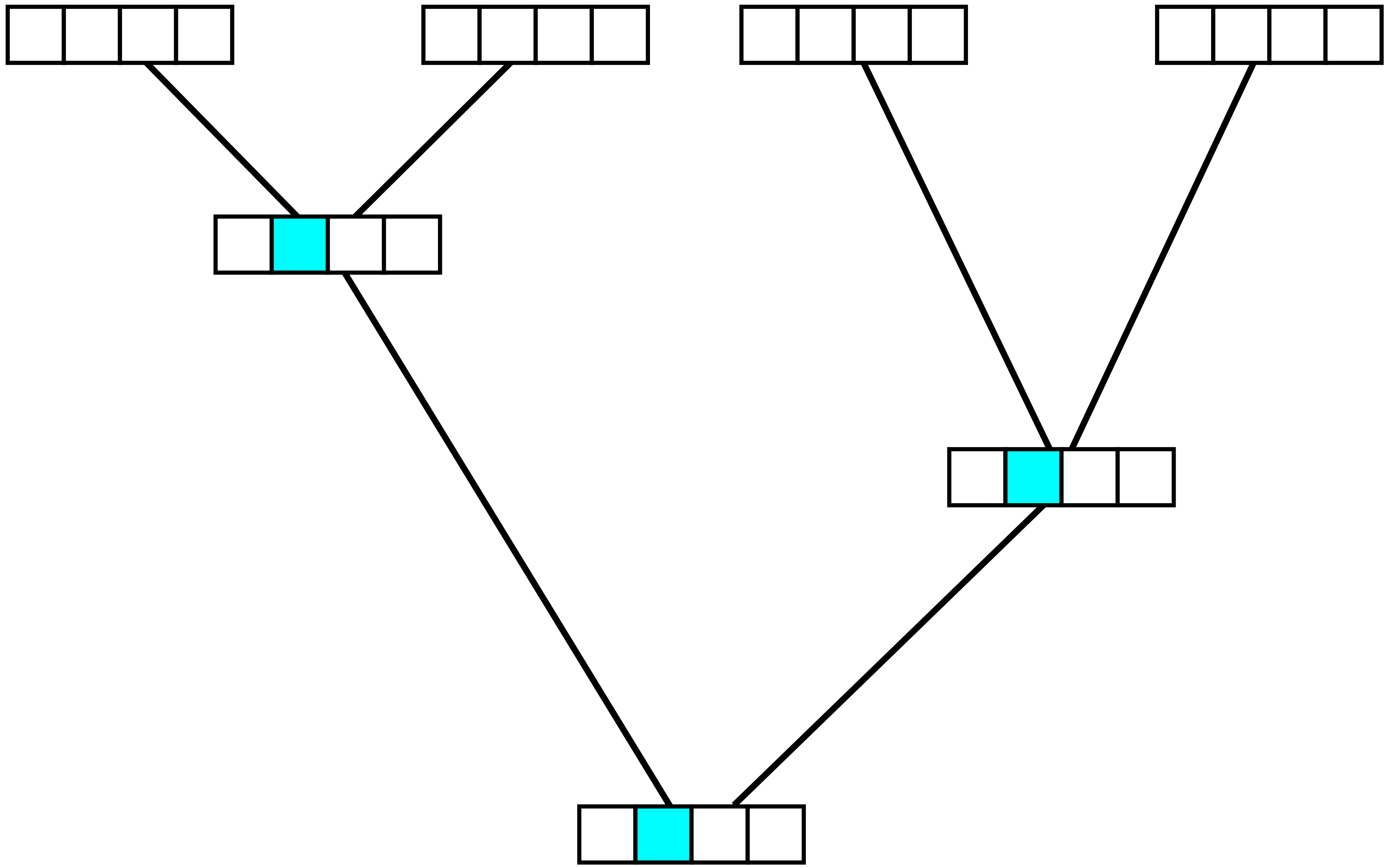


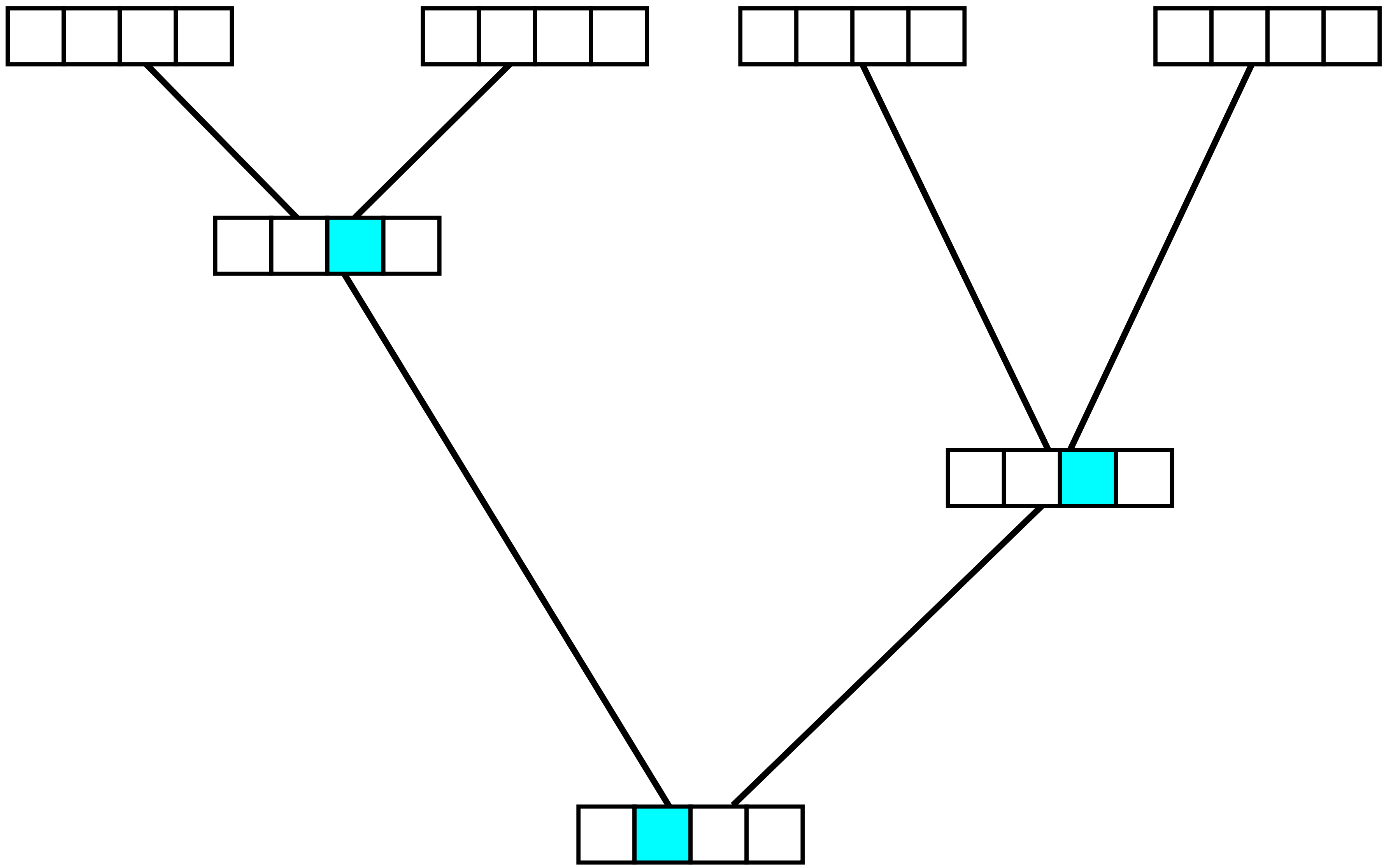


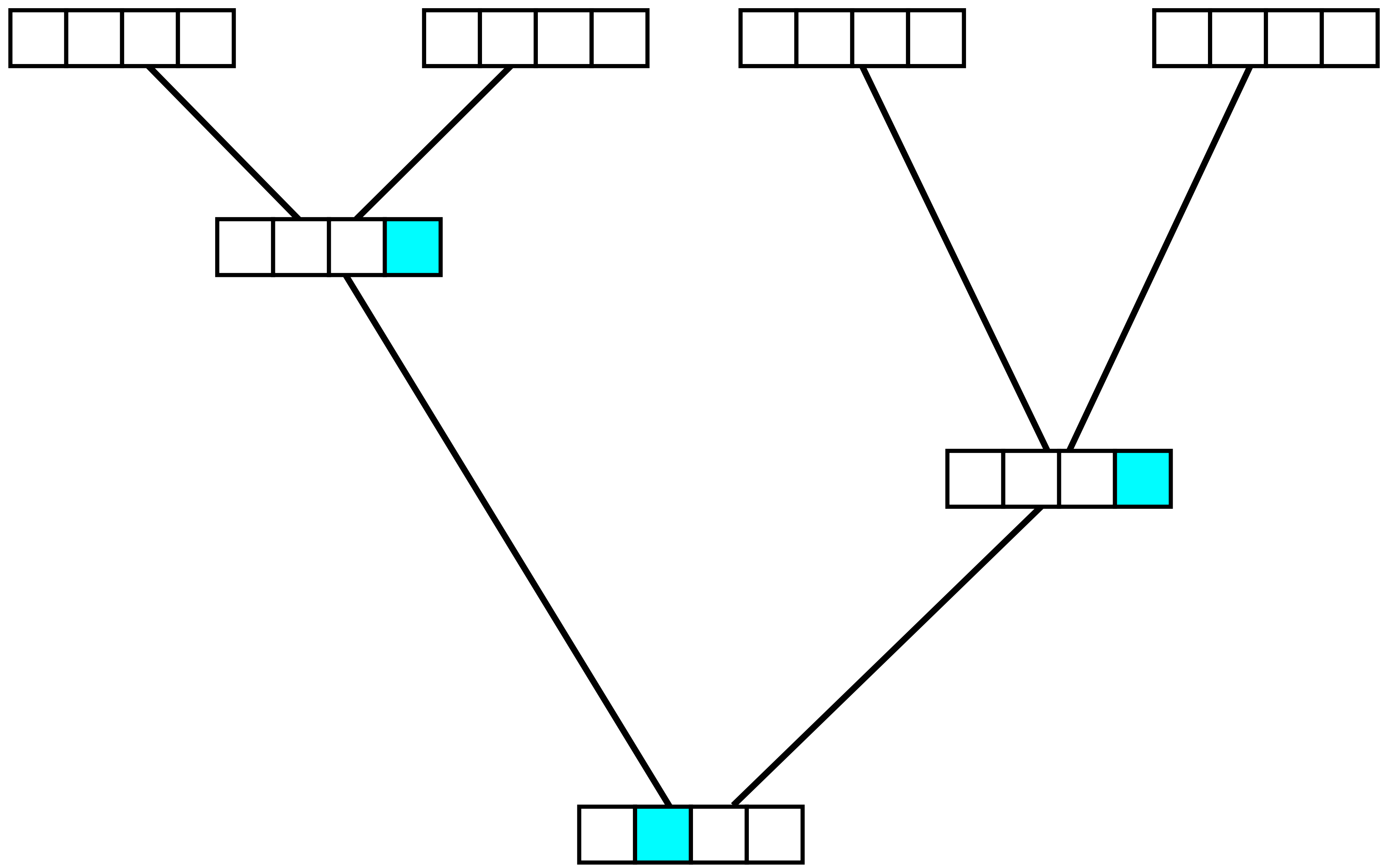


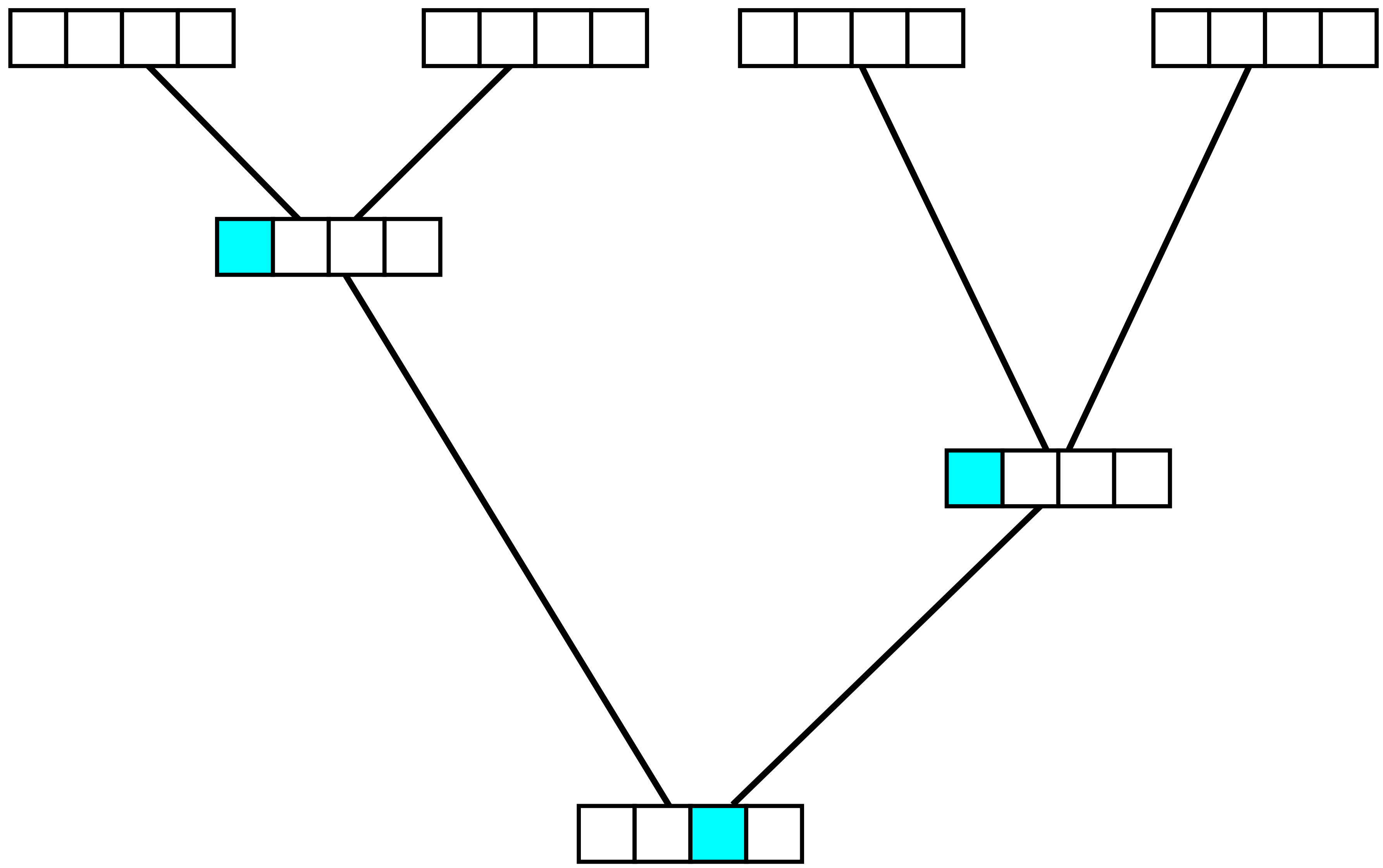


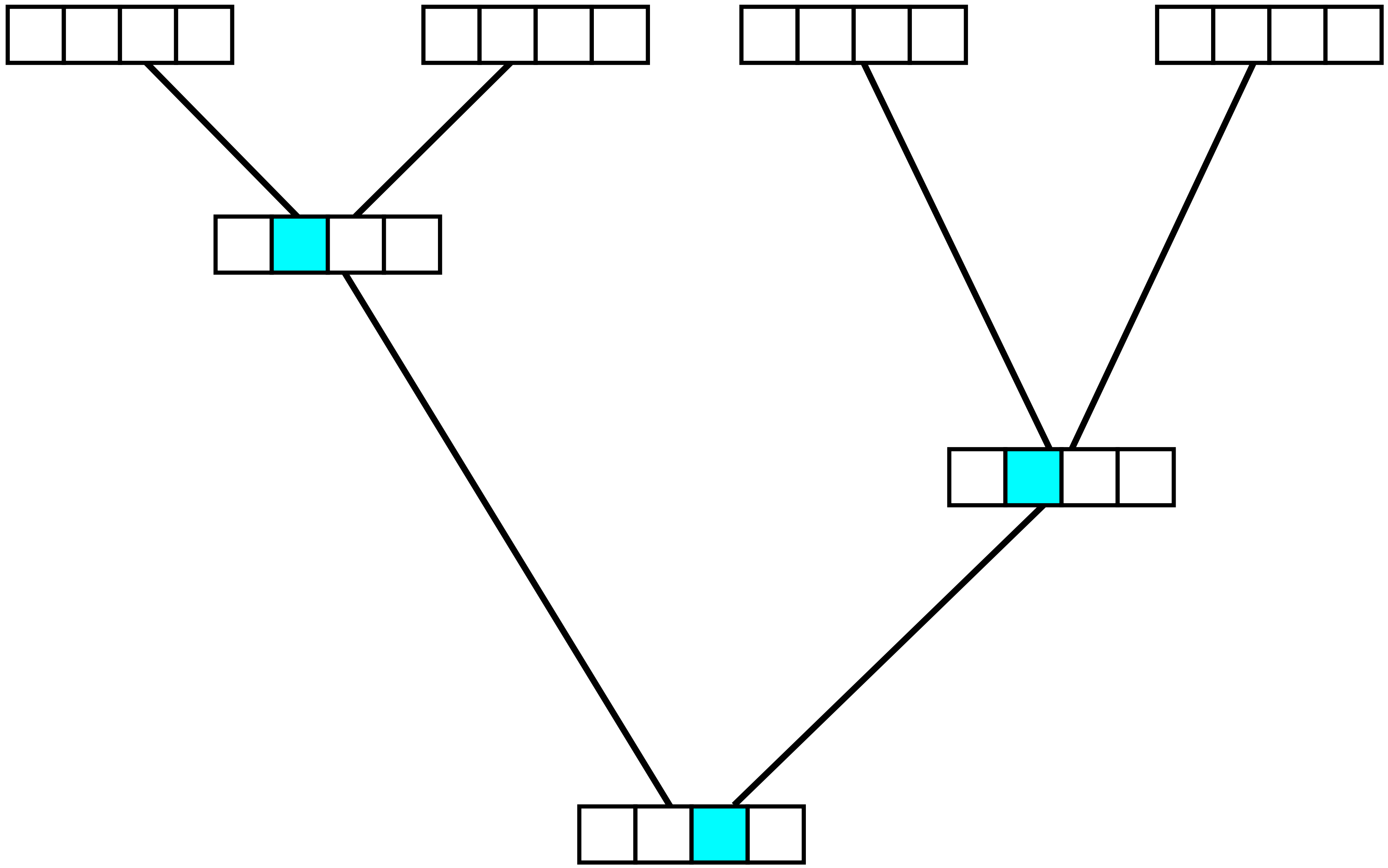


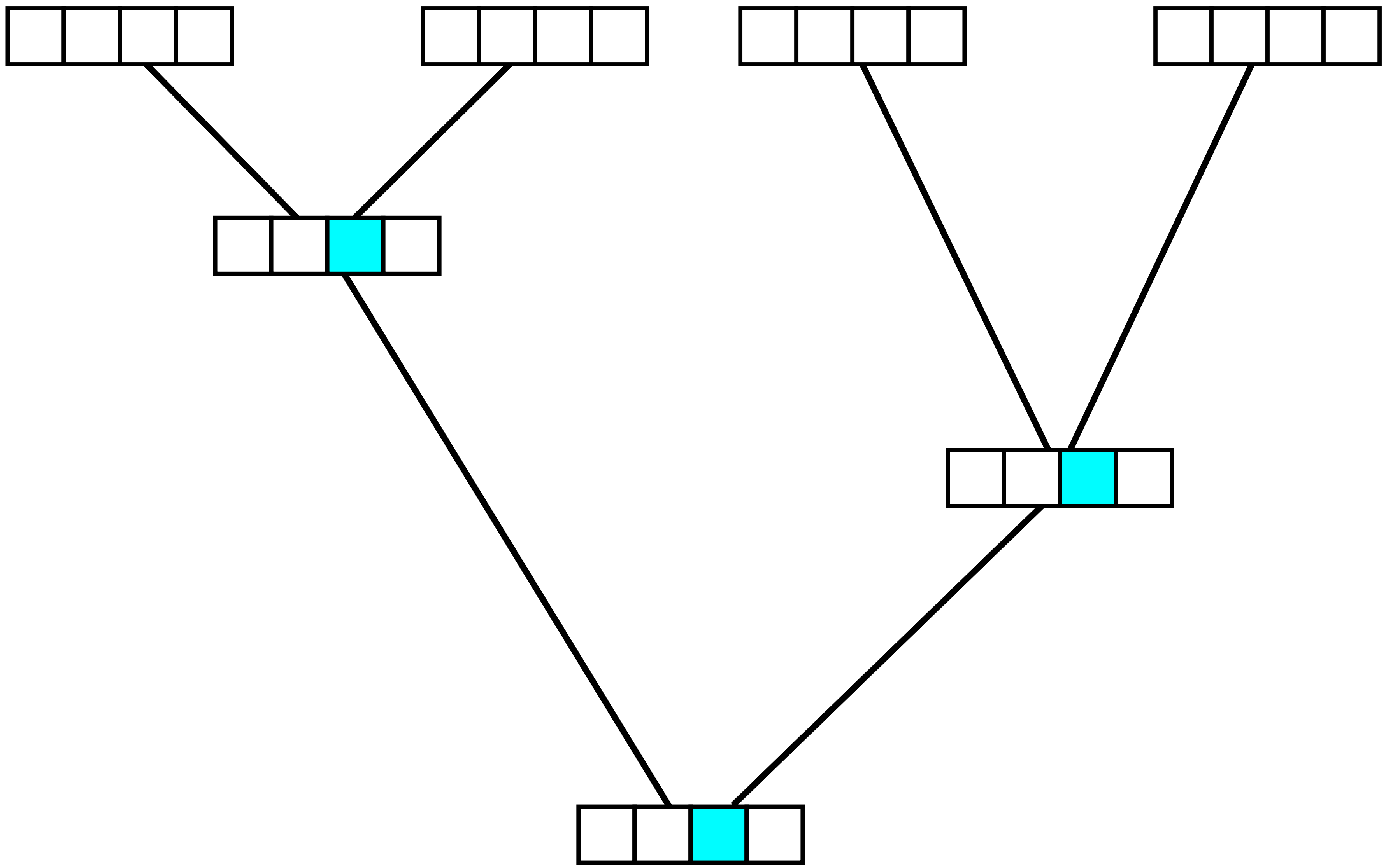


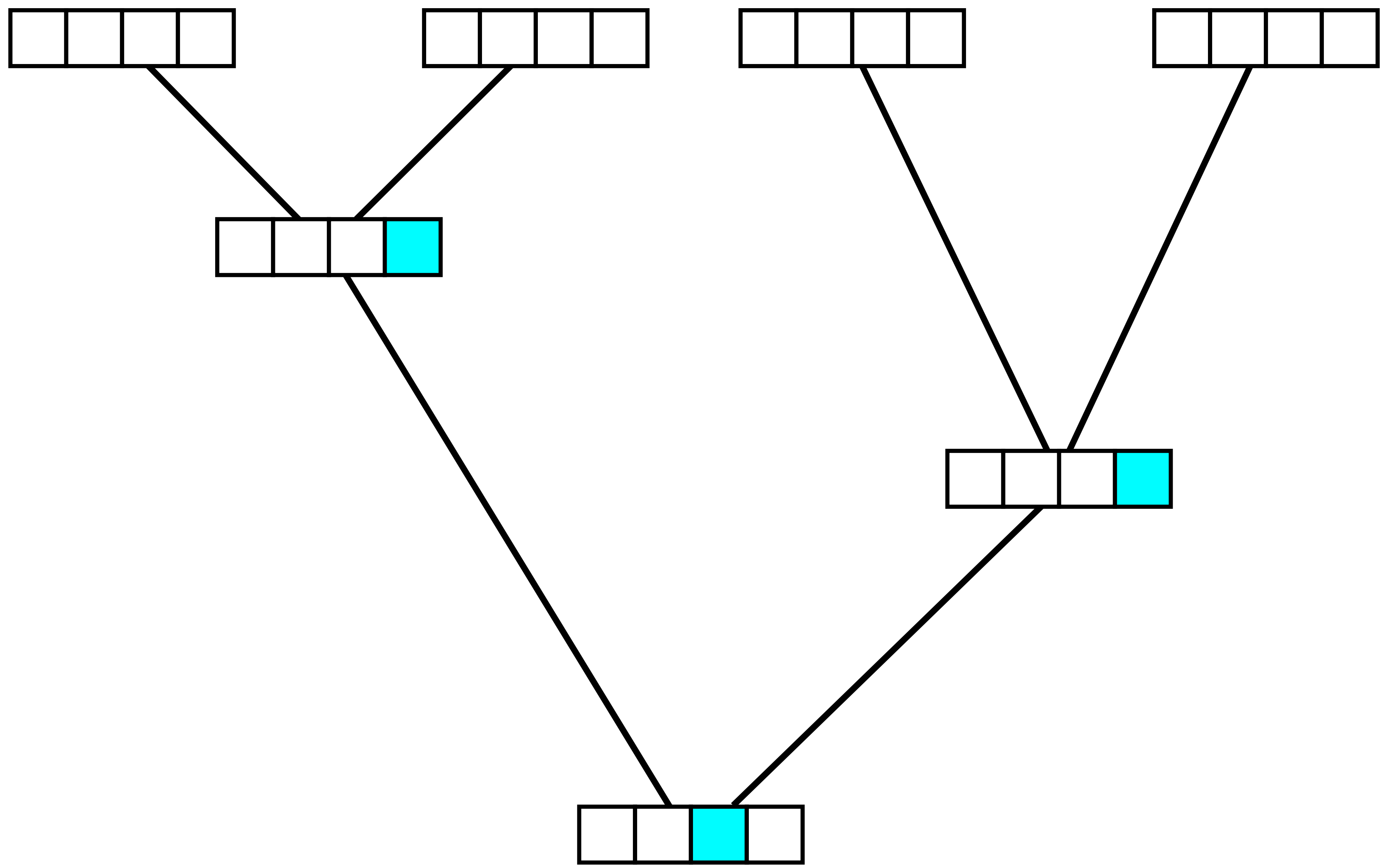


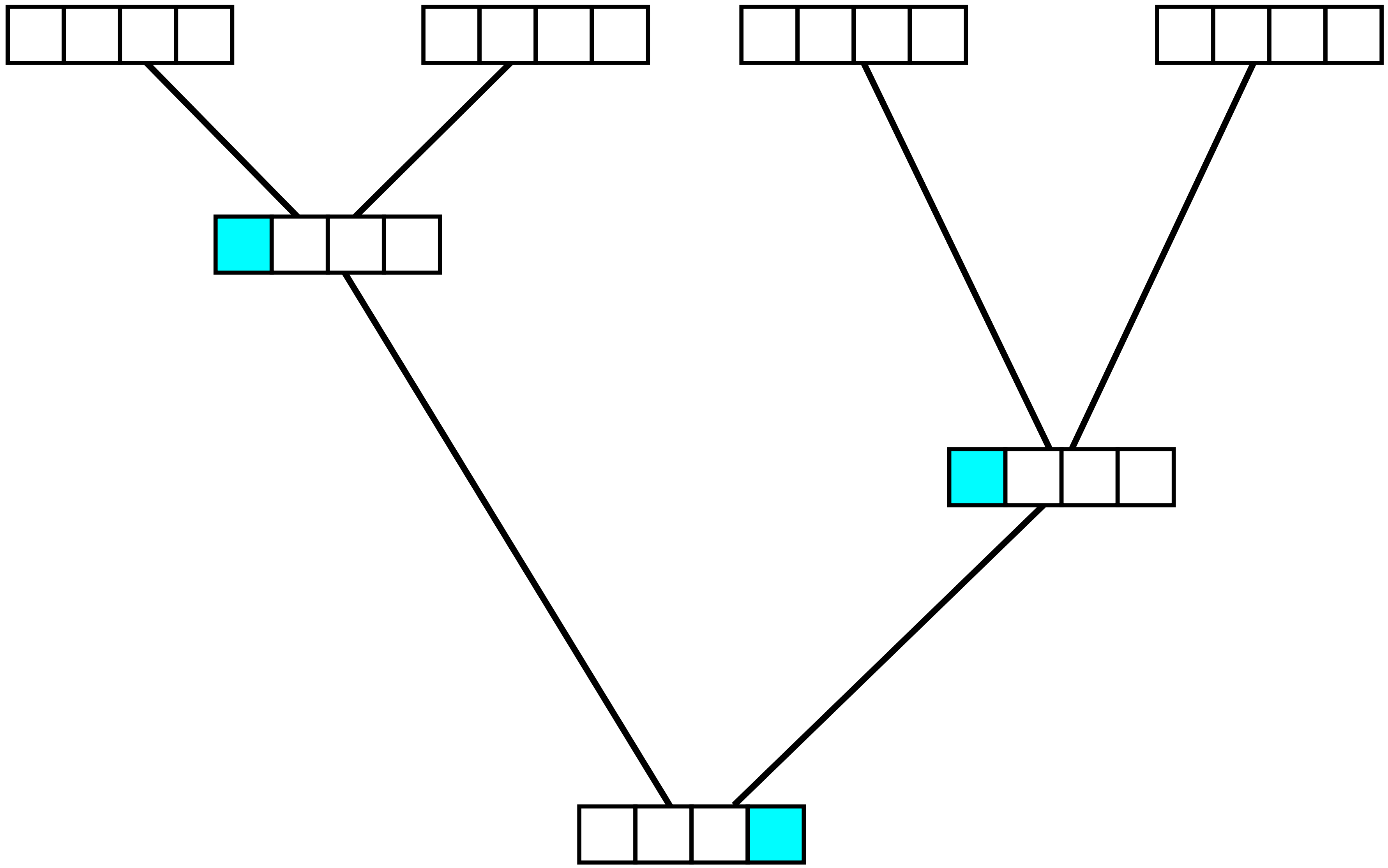


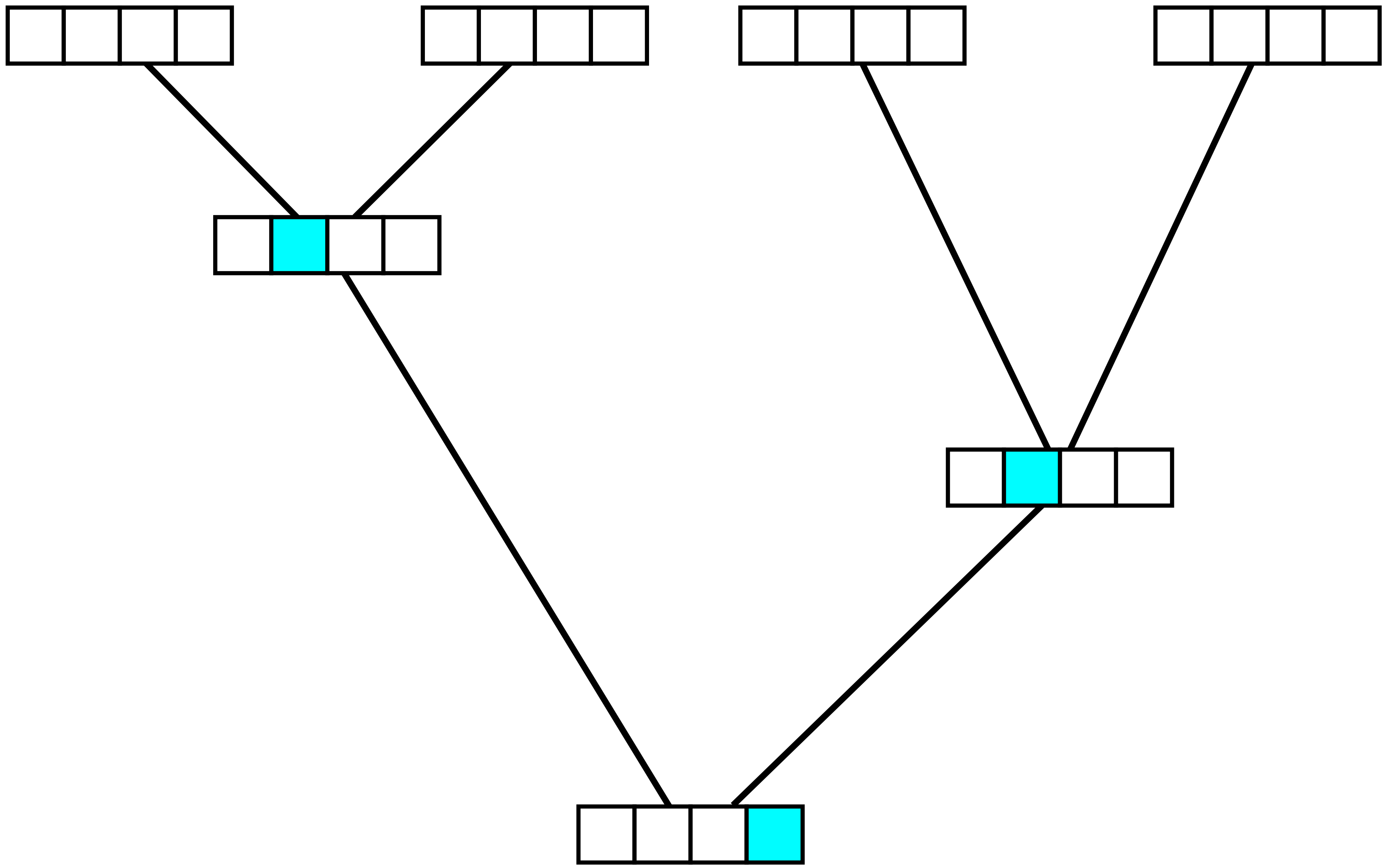


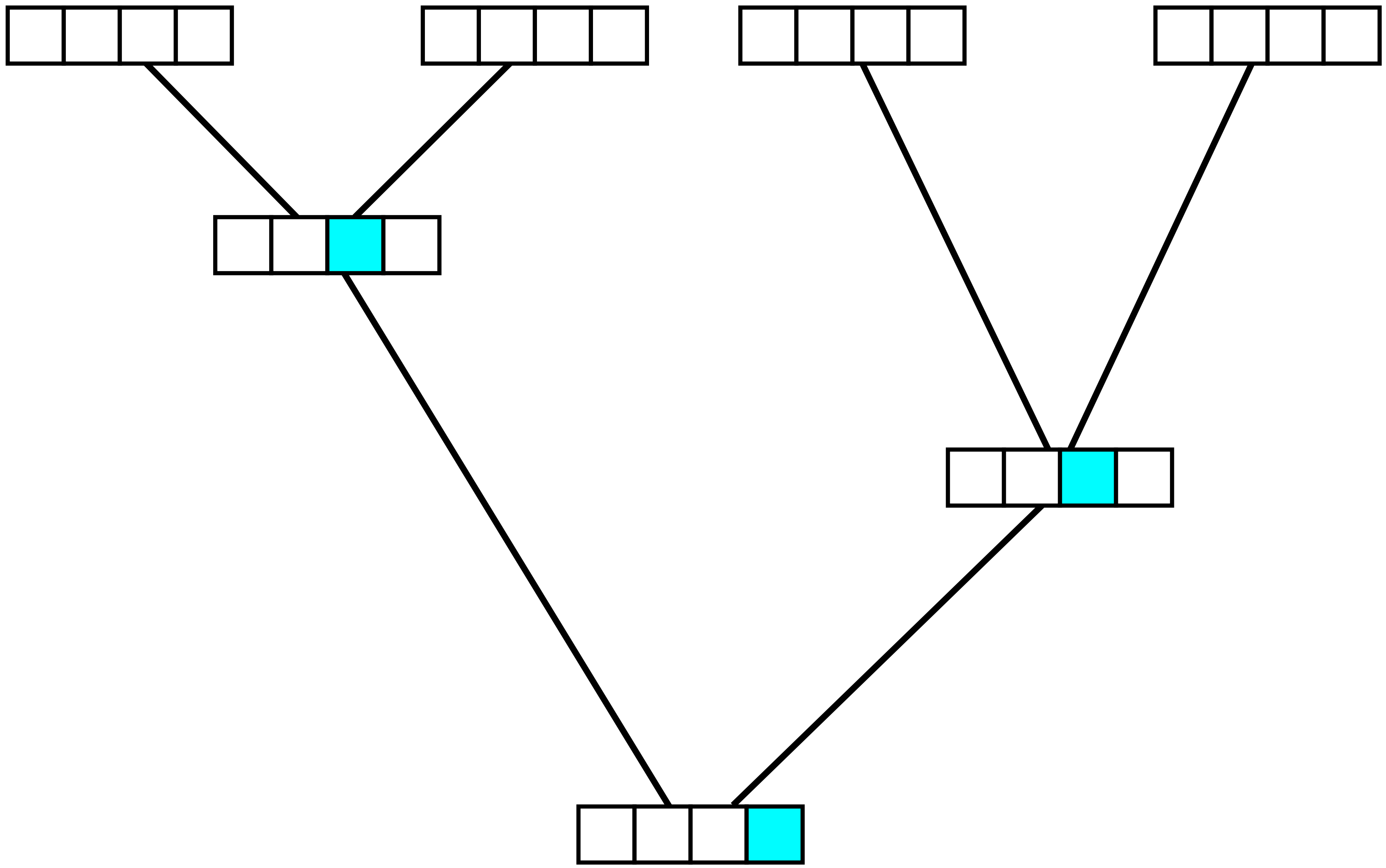


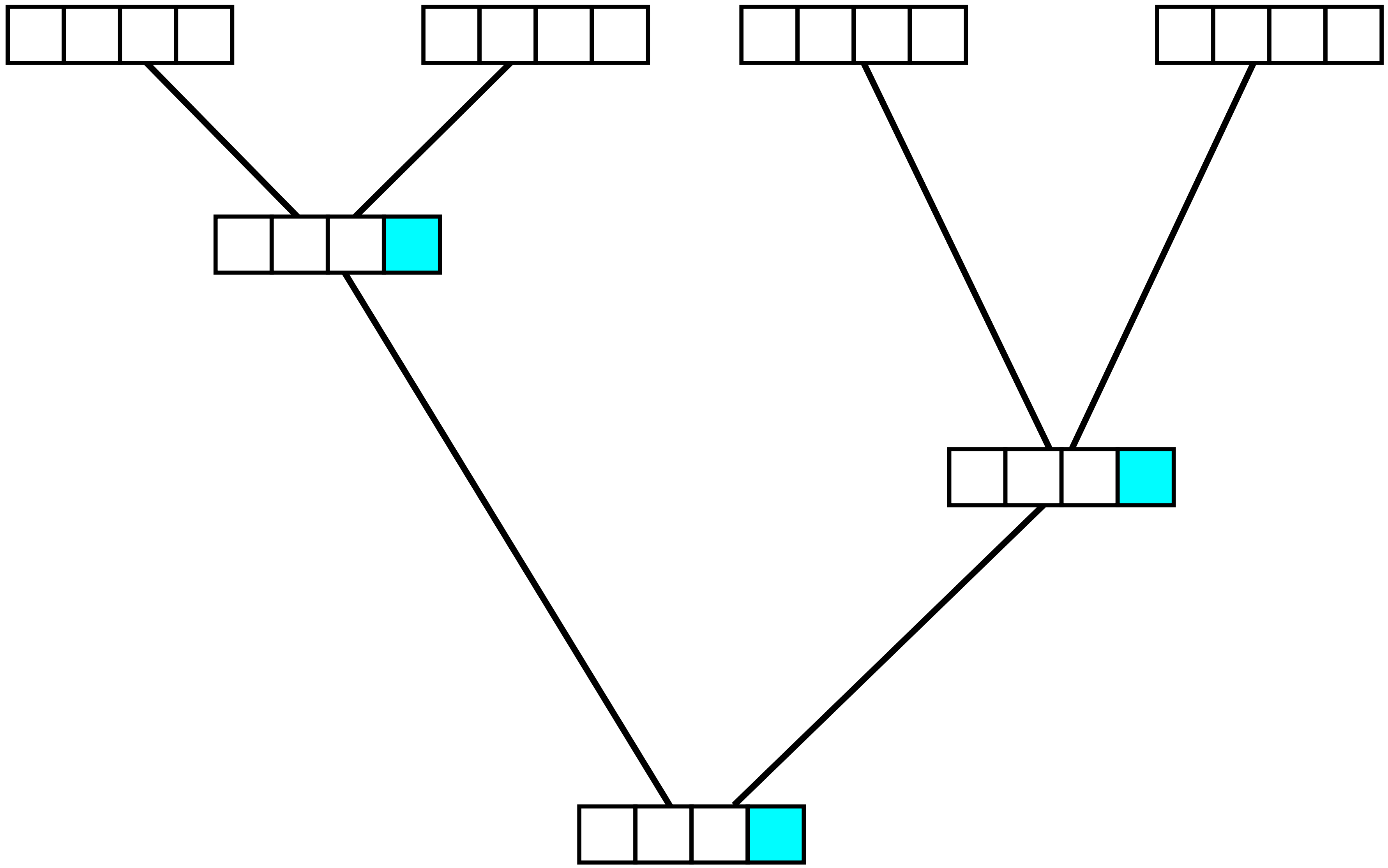


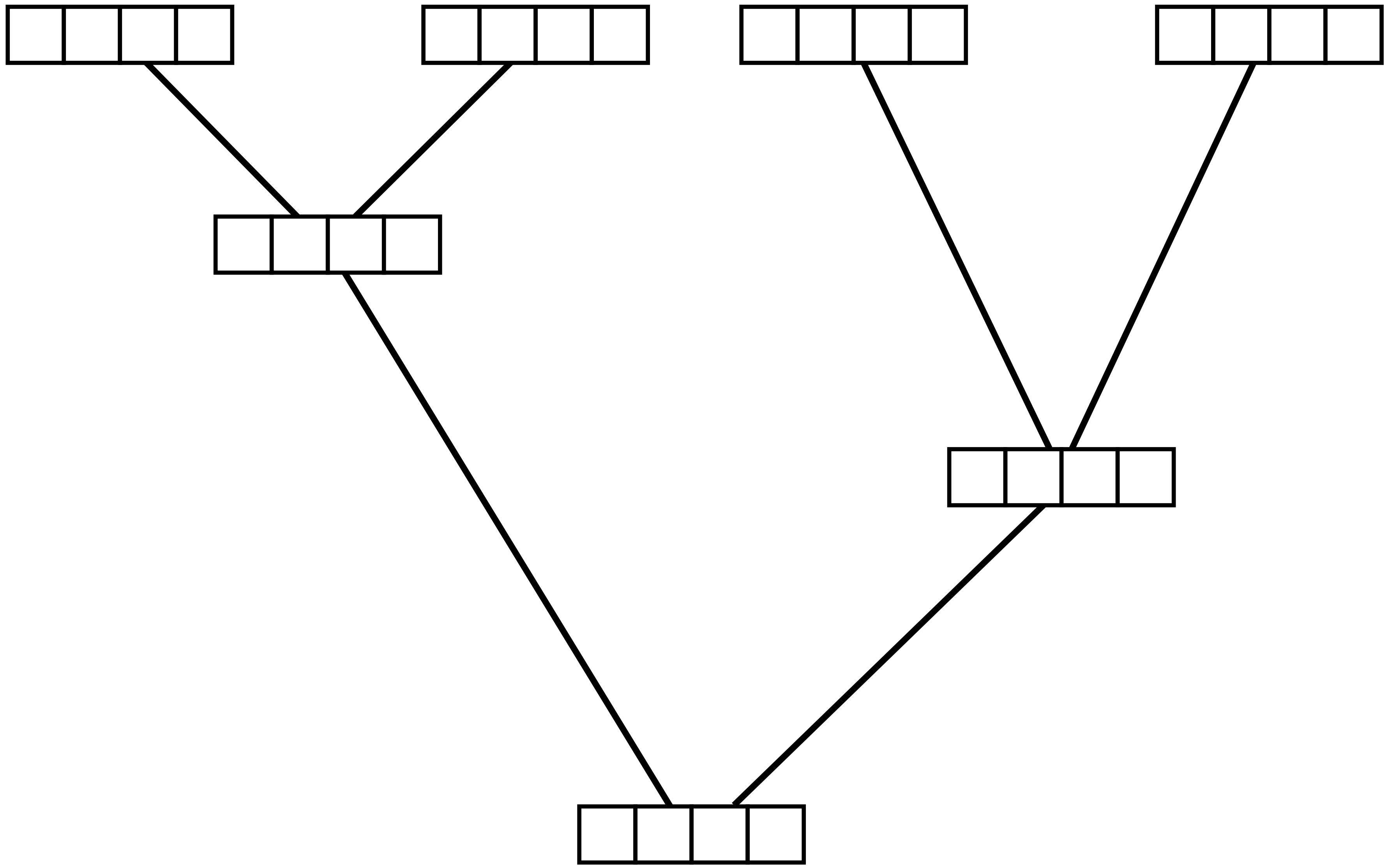




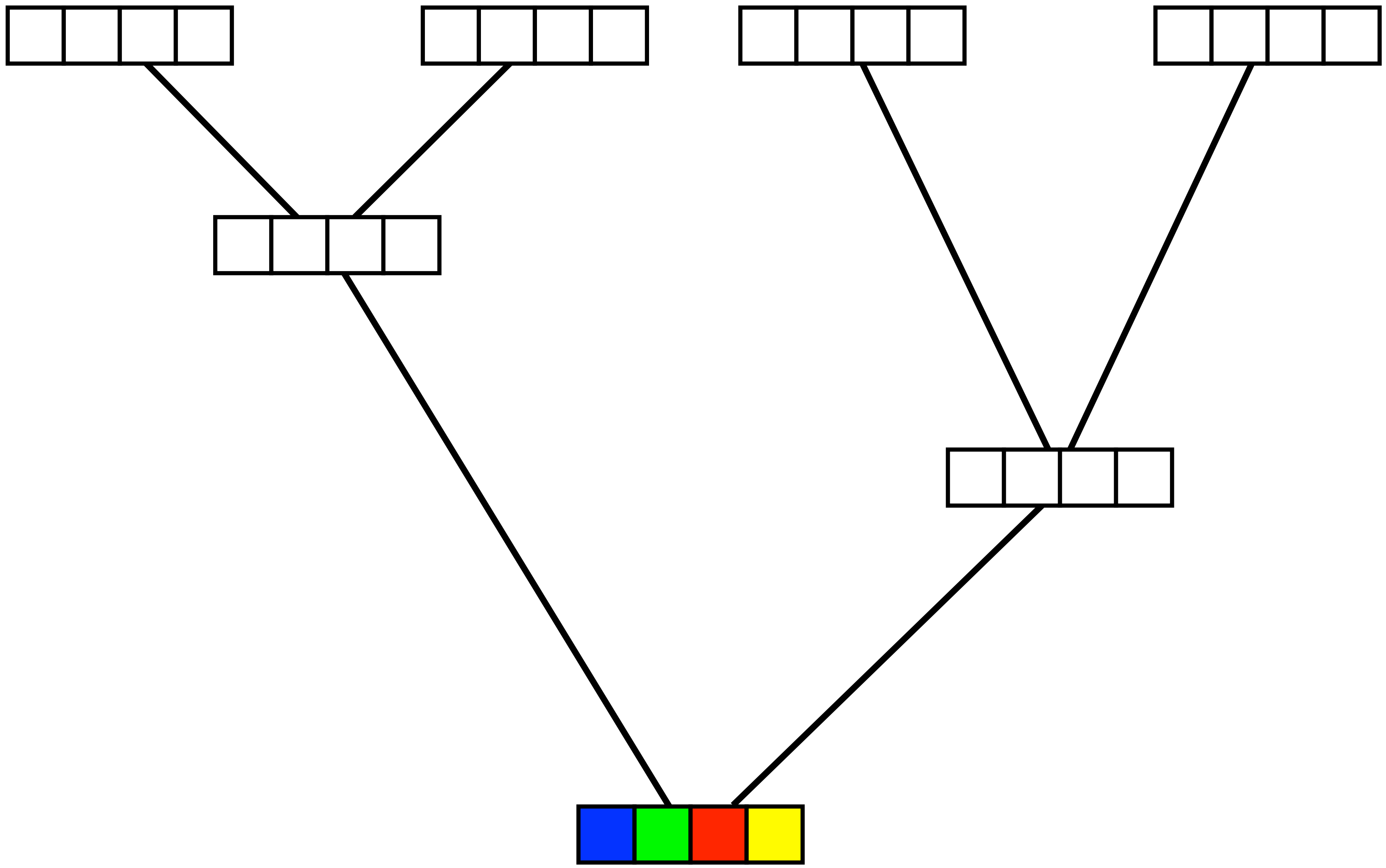




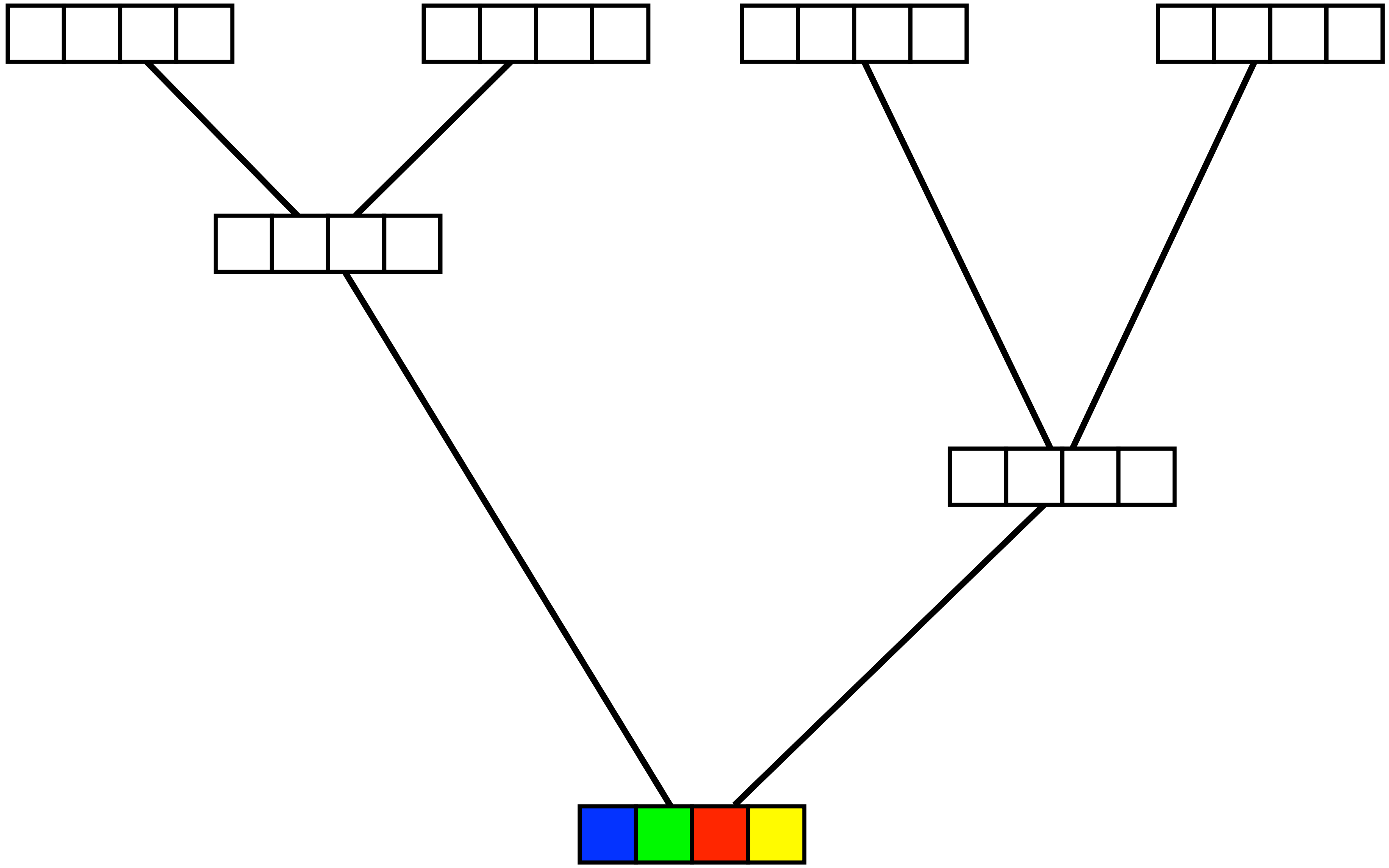




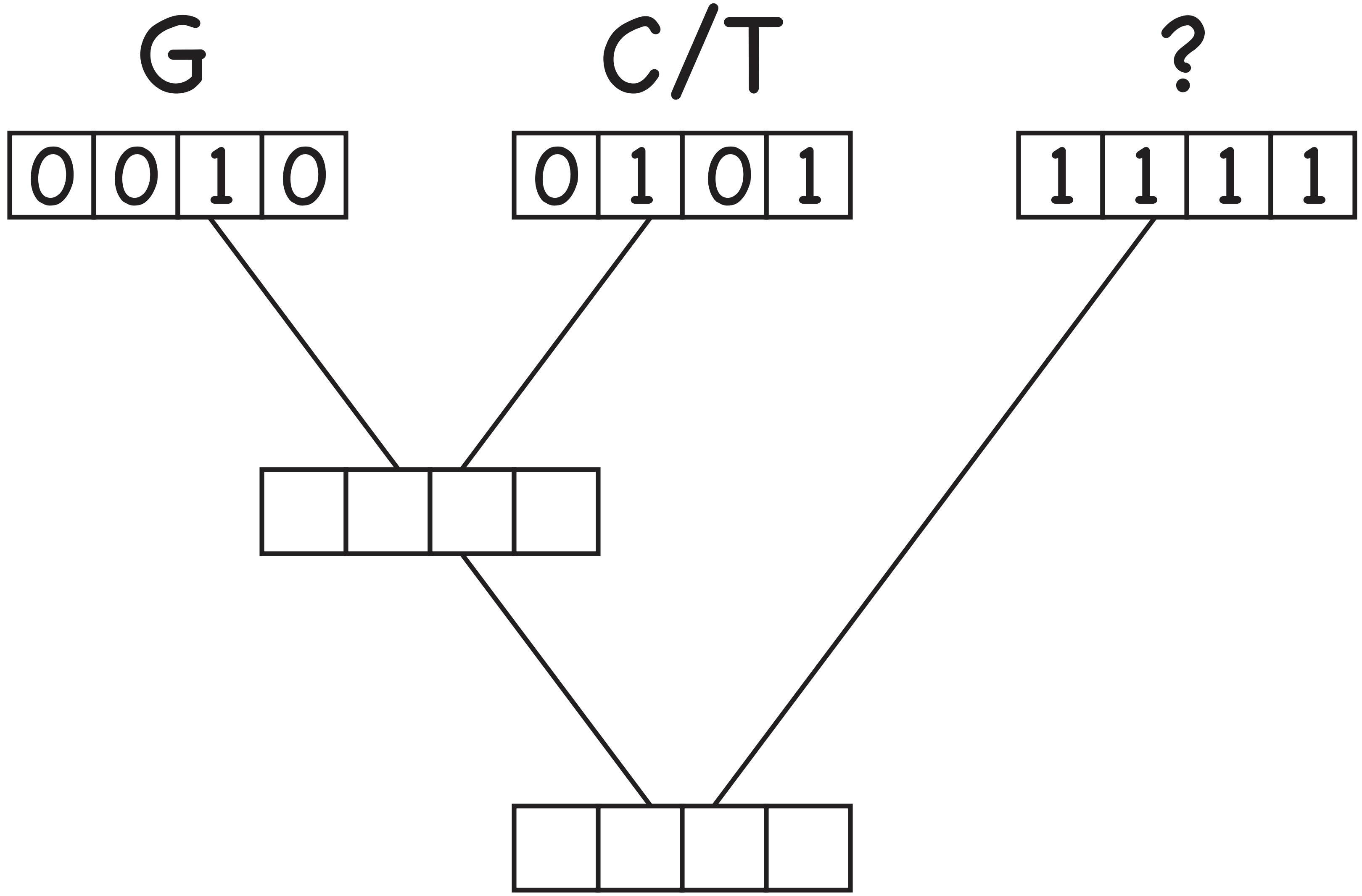
$$\ell_{\text{Site}} = \pi_A \times \ell_A^{\text{Root}} + \pi_C \times \ell_C^{\text{Root}} + \pi_G \times \ell_G^{\text{Root}} + \pi_T \times \ell_T^{\text{Root}}$$

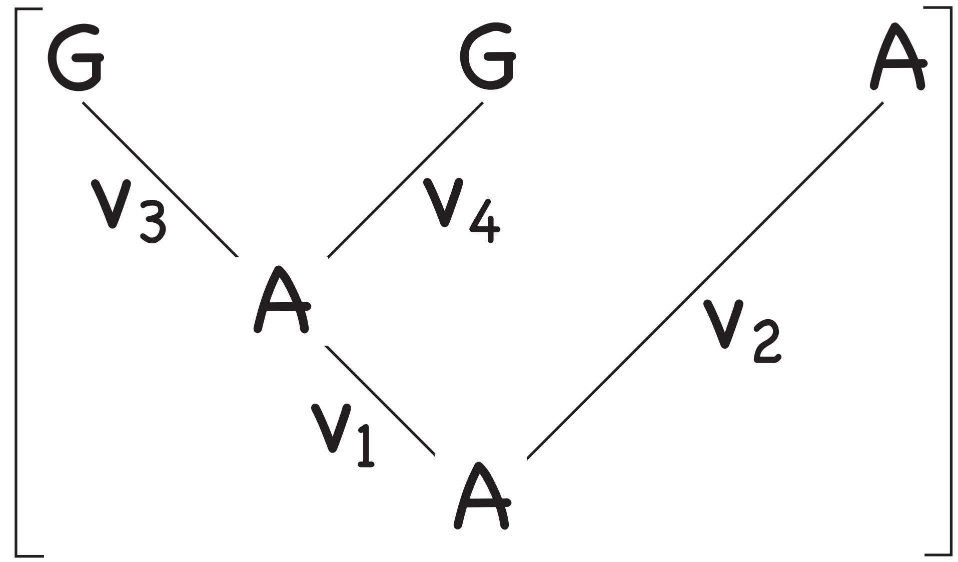


$$\ell_{\text{Site}} = \pi_A \times \ell_A^{\text{Root}} + \pi_C \times \ell_C^{\text{Root}} + \pi_G \times \ell_G^{\text{Root}} + \pi_T \times \ell_T^{\text{Root}}$$



$$\ell_{\text{Site}} = \pi_A \times \ell_A^{\text{Root}} + \pi_C \times \ell_C^{\text{Root}} + \pi_G \times \ell_G^{\text{Root}} + \pi_T \times \ell_T^{\text{Root}}$$


G
C/T
?
0 0 1 0
0 1 0 1
1 1 1 1

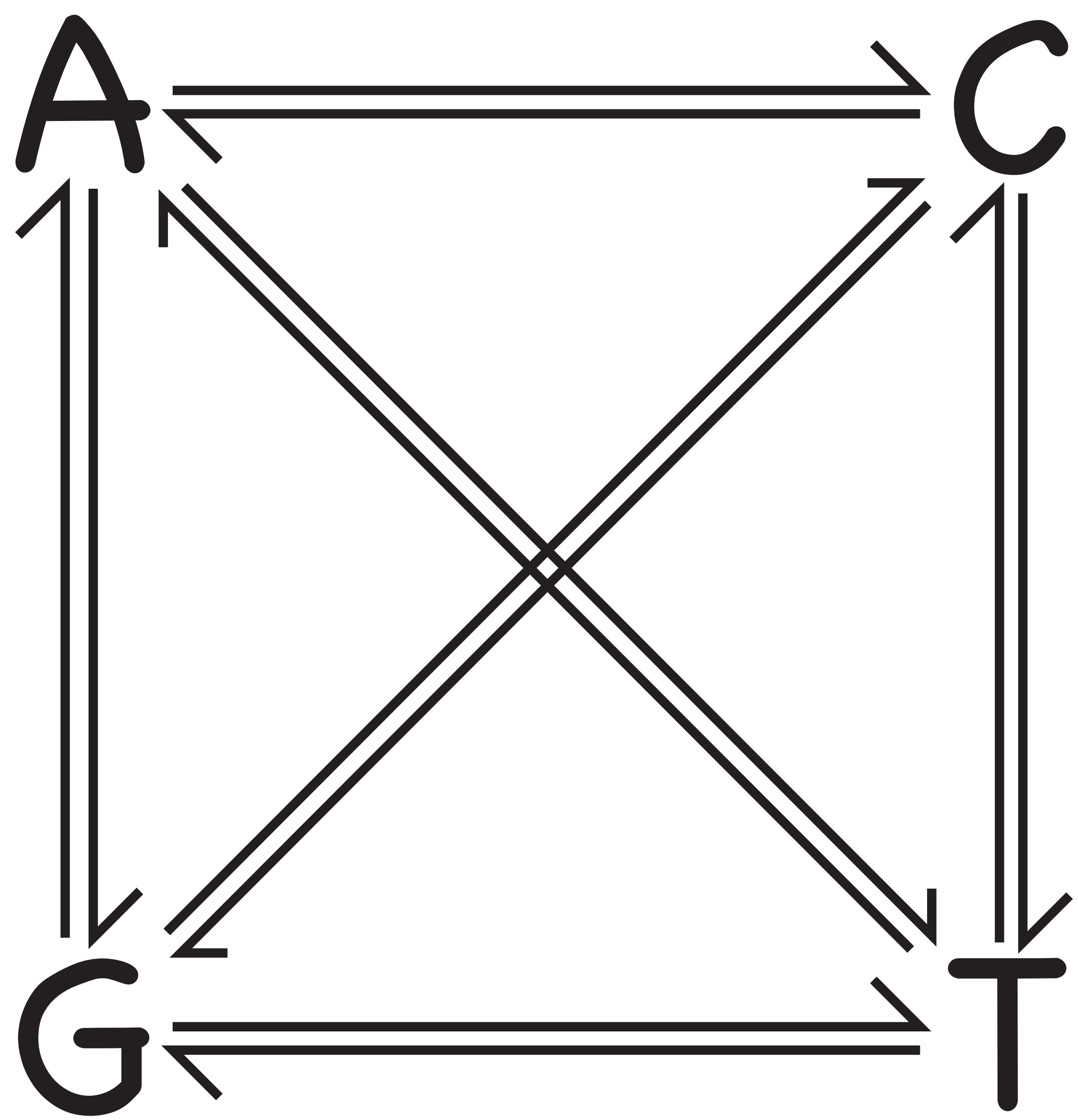
Pr  =

$$\pi_A \times p_{AA}(v_1) \times p_{AA}(v_2) \times p_{AG}(v_3) \times p_{AG}(v_4)$$

π_i – Stationary frequencies

$p_{ij}(v)$ – Transition probabilities

Continuous-Time Markov Chain



To

From		A	C	G	T
	A	−0.886	0.190	0.633	0.063
	C	0.253	−0.696	0.127	0.316
	G	1.266	0.190	−1.519	0.063
	T	0.253	0.949	0.127	−1.329

$$Q = \begin{pmatrix} -0.886 & 0.190 & 0.633 & 0.063 \\ 0.253 & -0.696 & 0.127 & 0.316 \\ 1.266 & 0.190 & -1.519 & 0.063 \\ 0.253 & 0.949 & 0.127 & -1.329 \end{pmatrix}$$

		To		
	A	C	G	T
From A	−0.886	0.190	0.633	0.063
From C	0.253	−0.696	0.127	0.316
From G	1.266	0.190	−1.519	0.063
From T	0.253	0.949	0.127	−1.329

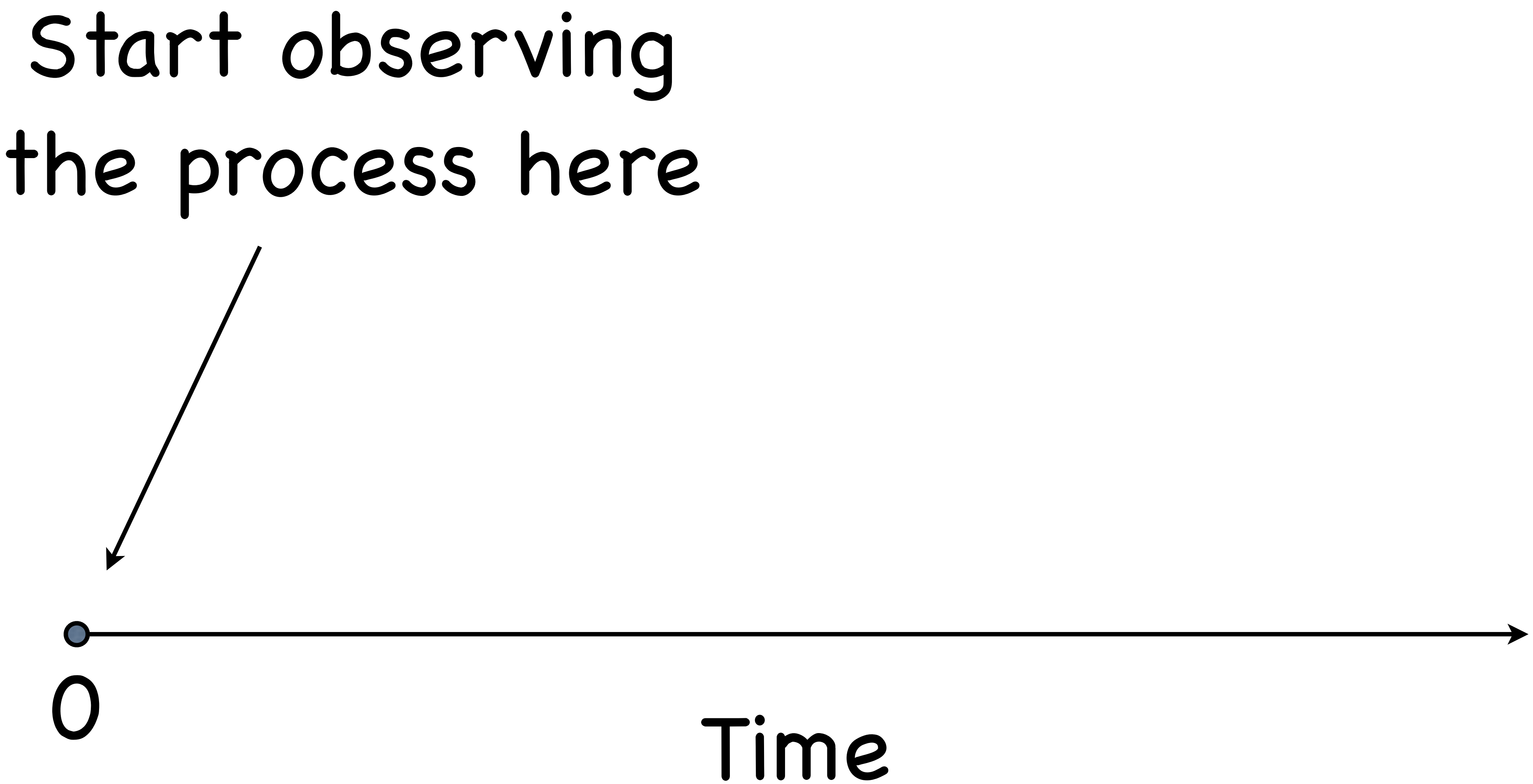
Interpretation: If the process is in state i, we wait an exponentially distributed amount of time with parameter $-q_{ii}$ until the next substitution occurs.

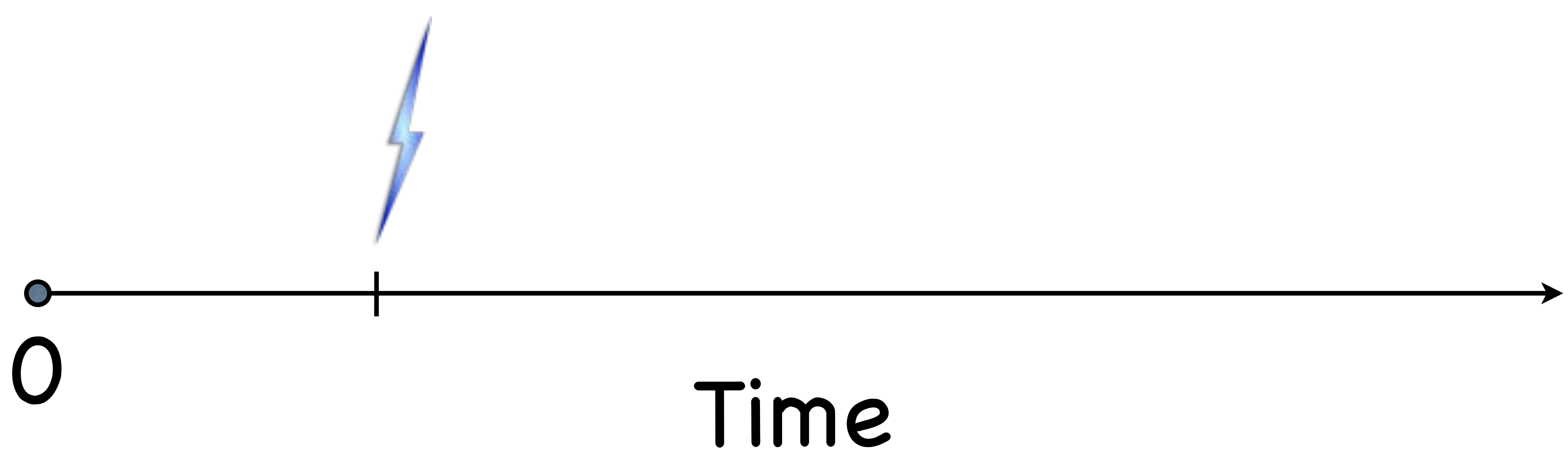
		To			
		A	C	G	T
From	A	−0.886	0.190	0.633	0.063
	C	0.253	−0.696	0.127	0.316
	G	1.266	0.190	−1.519	0.063
	T	0.253	0.949	0.127	−1.329

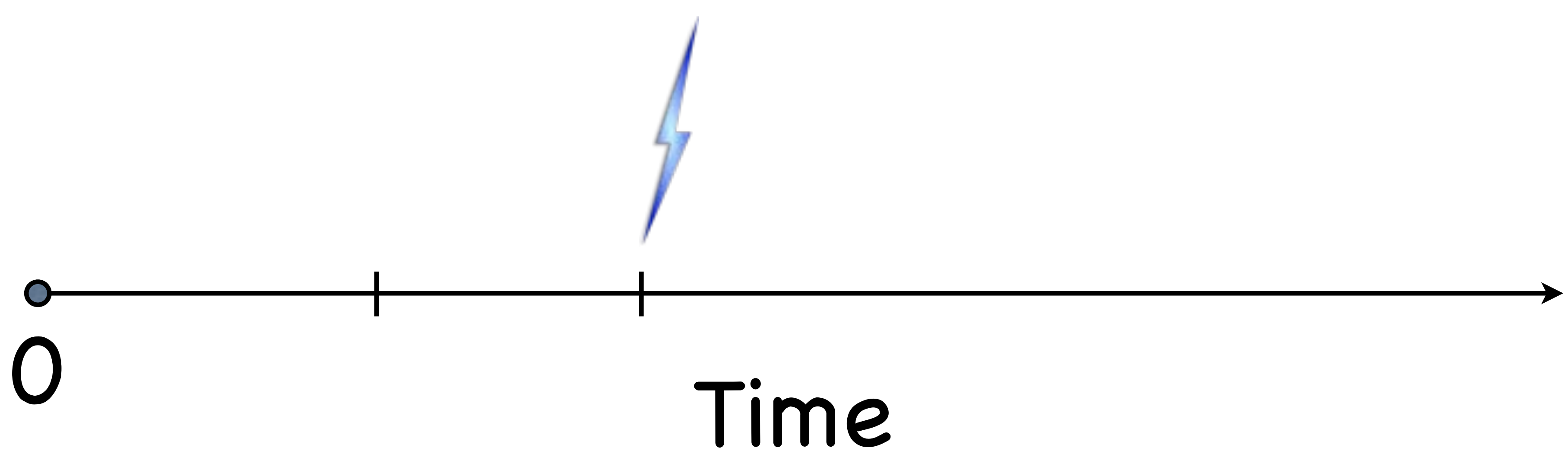
Interpretation: The change is to state j with probability $-q_{ij}/q_{ii}$.

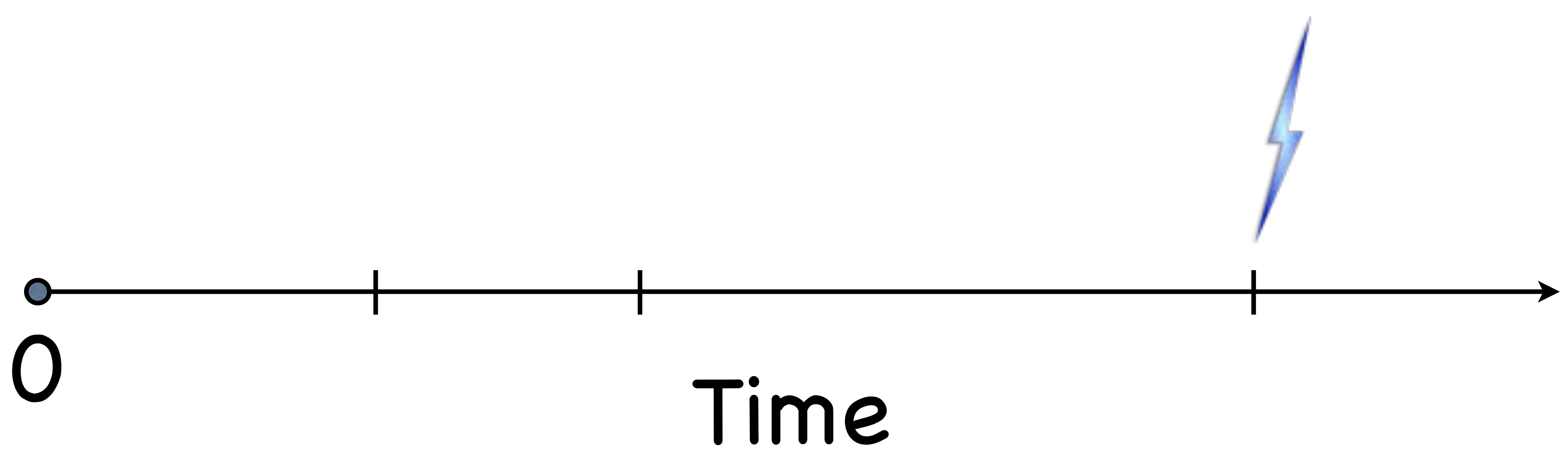
Something — the arrival of a customer, a coal mining disaster, a photon hitting a photodetector, a particle emission from a radioactive substance, a nucleotide substitution — occurs at a constant rate.

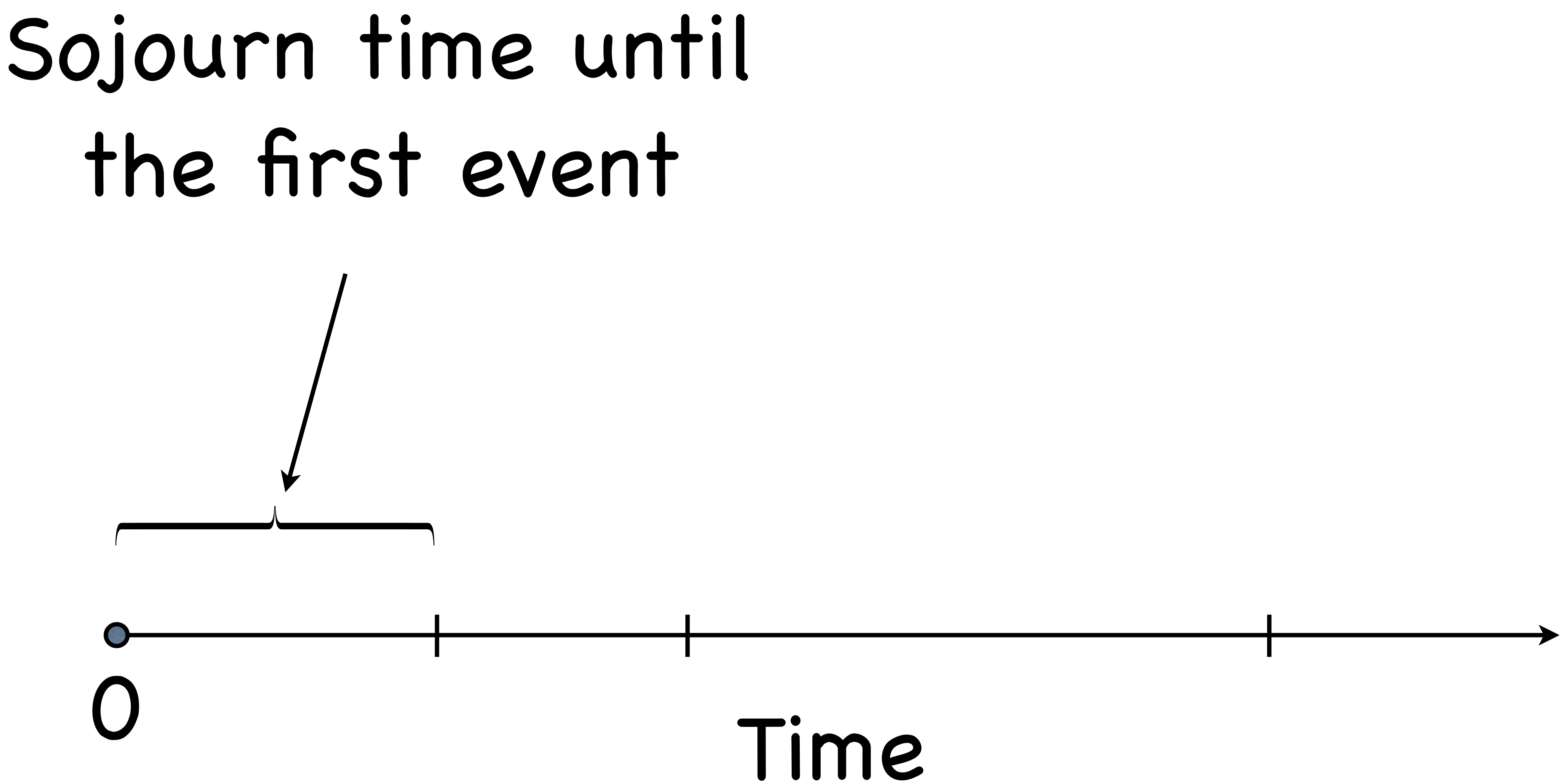
The rate at which the somethings (events) occur is λ.

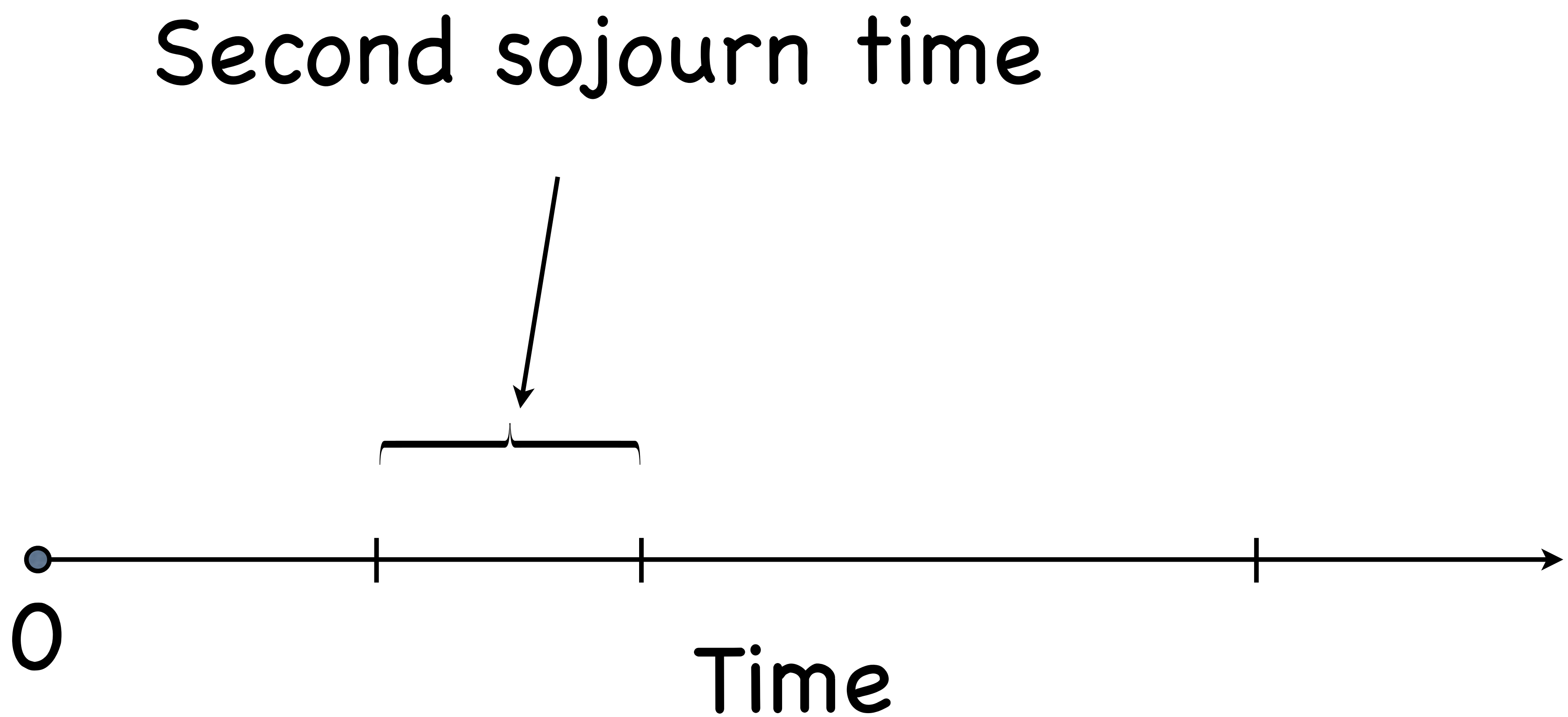

Start observing
the process here
0
Time

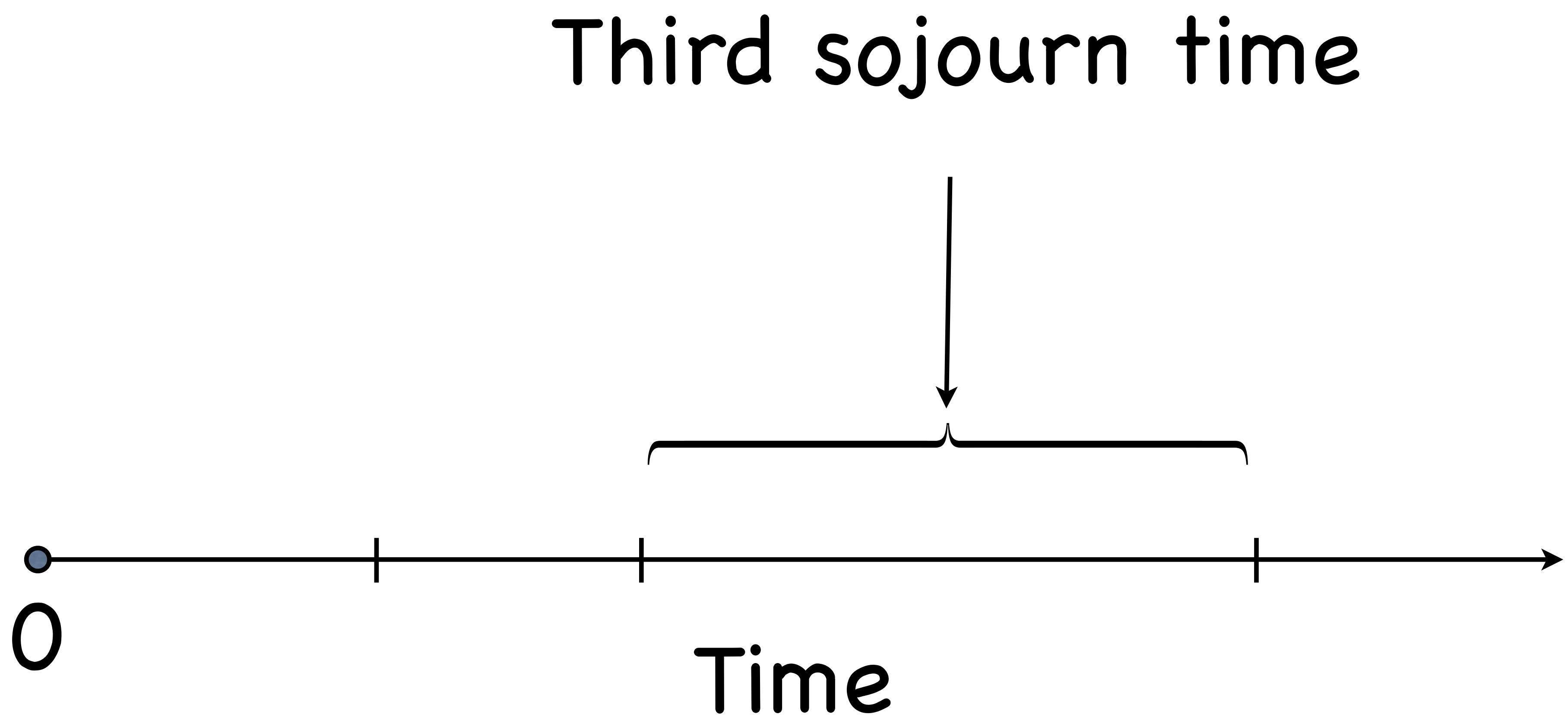

0
Time


0
Time

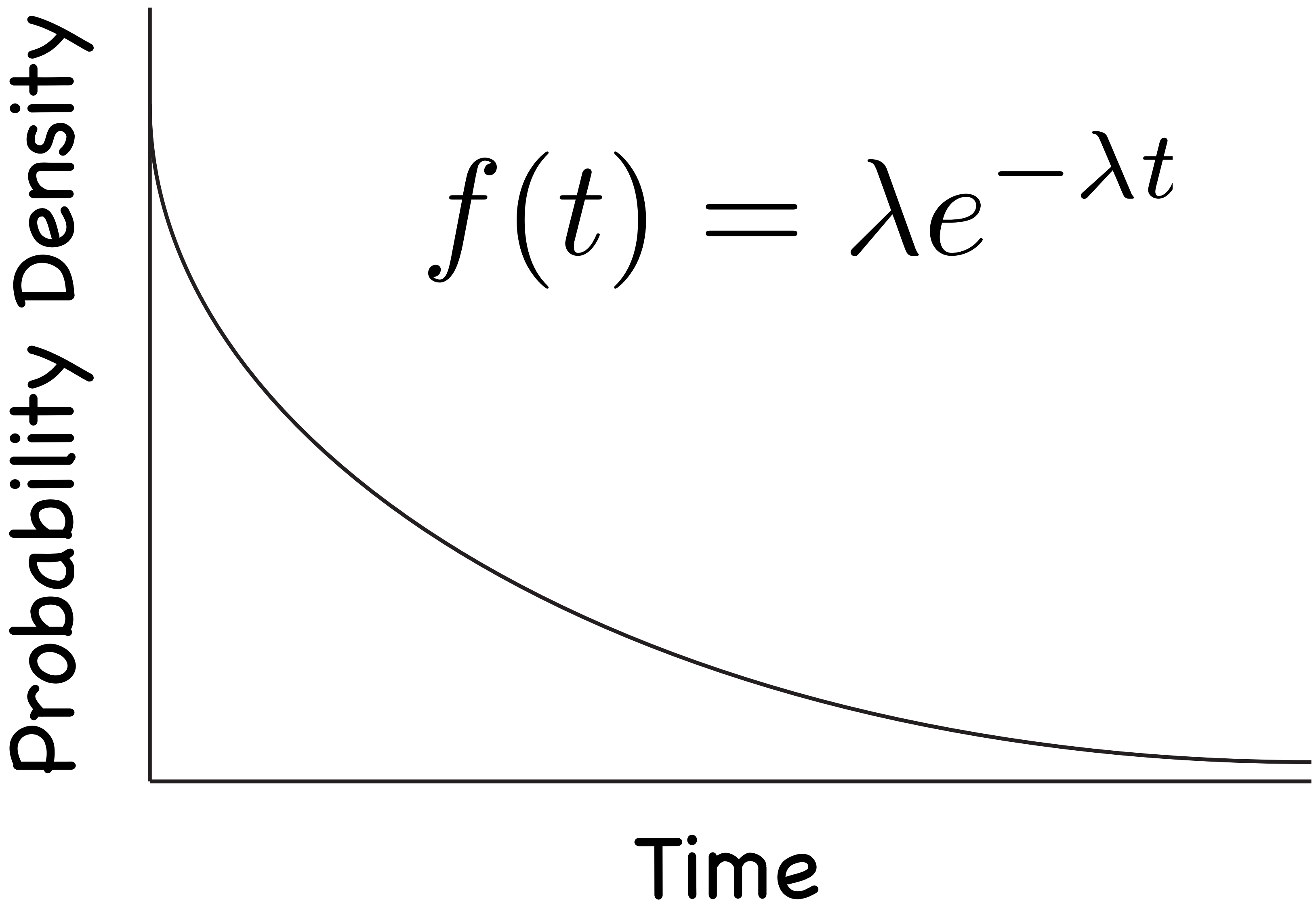

0
Time


Sojourn time until
the first event
0
Time


Second sojourn time
0
Time


Third sojourn time
0
Time

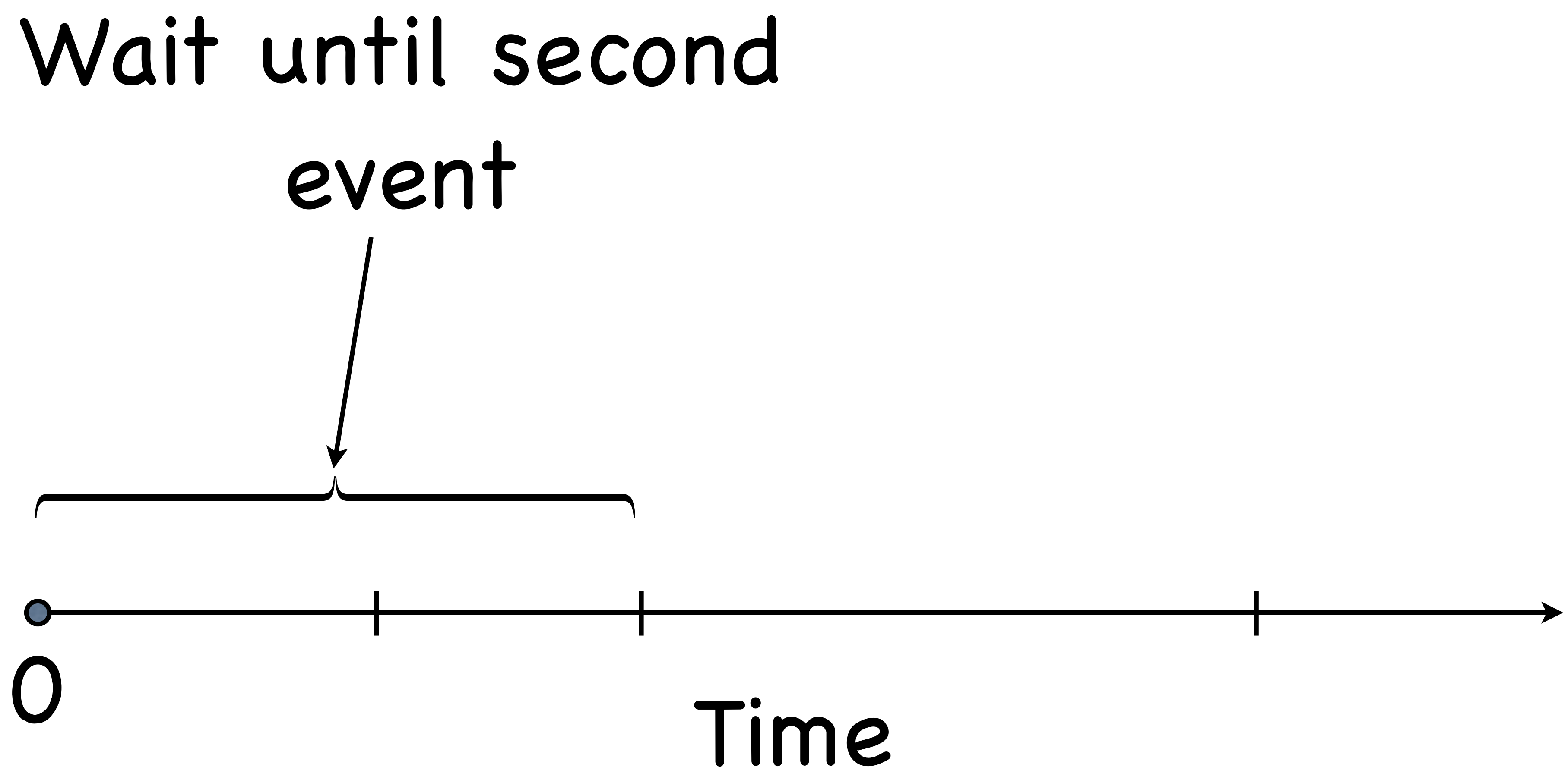
Important fact: The sojourn times are exponentially-distributed random variables

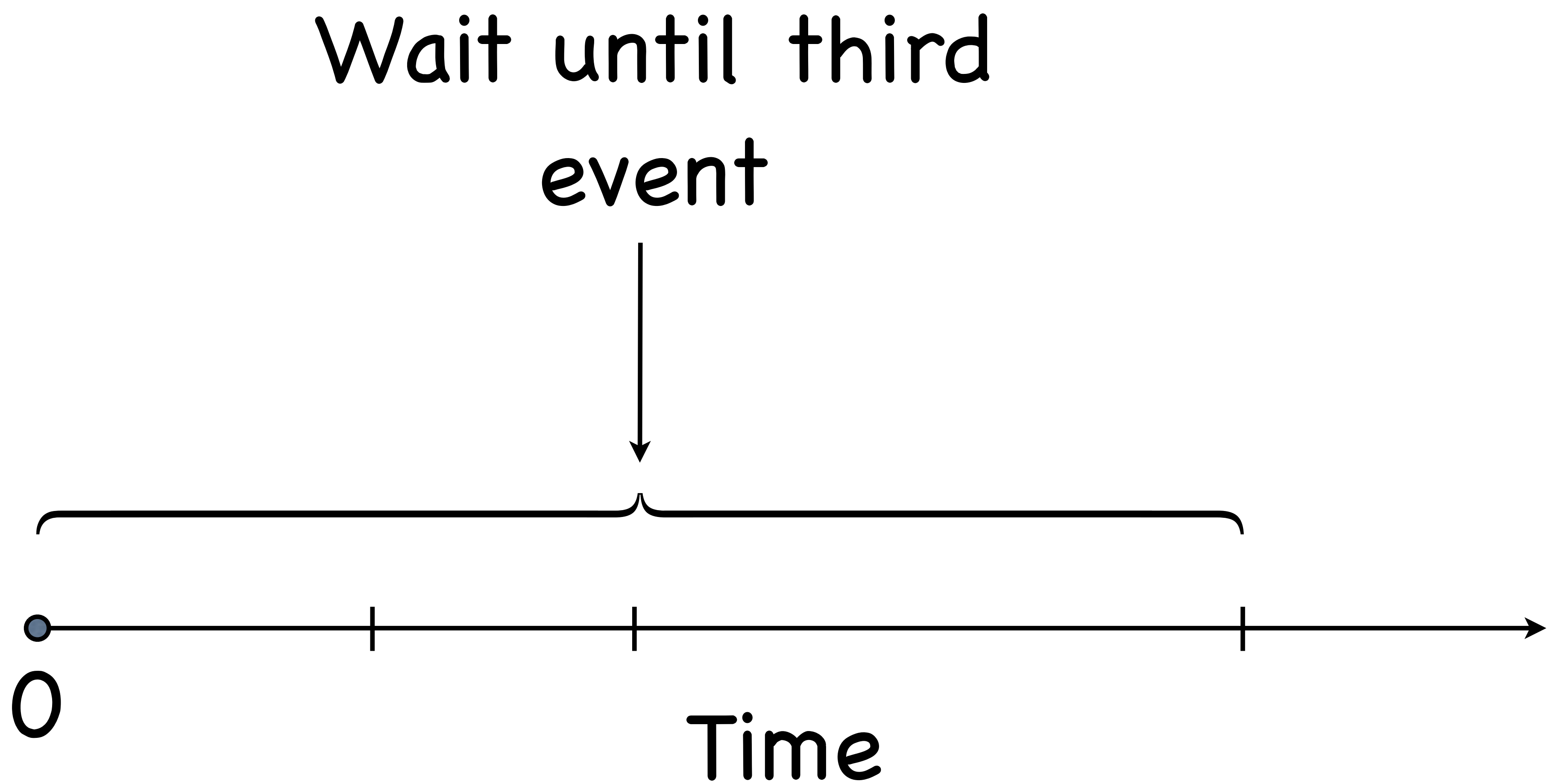


$$f(t) = \lambda e^{-\lambda t}$$

Interesting fact: The sojourn time is the exponentially-distributed time until the next event.

However, one can ask what is the waiting time until the k-th event?

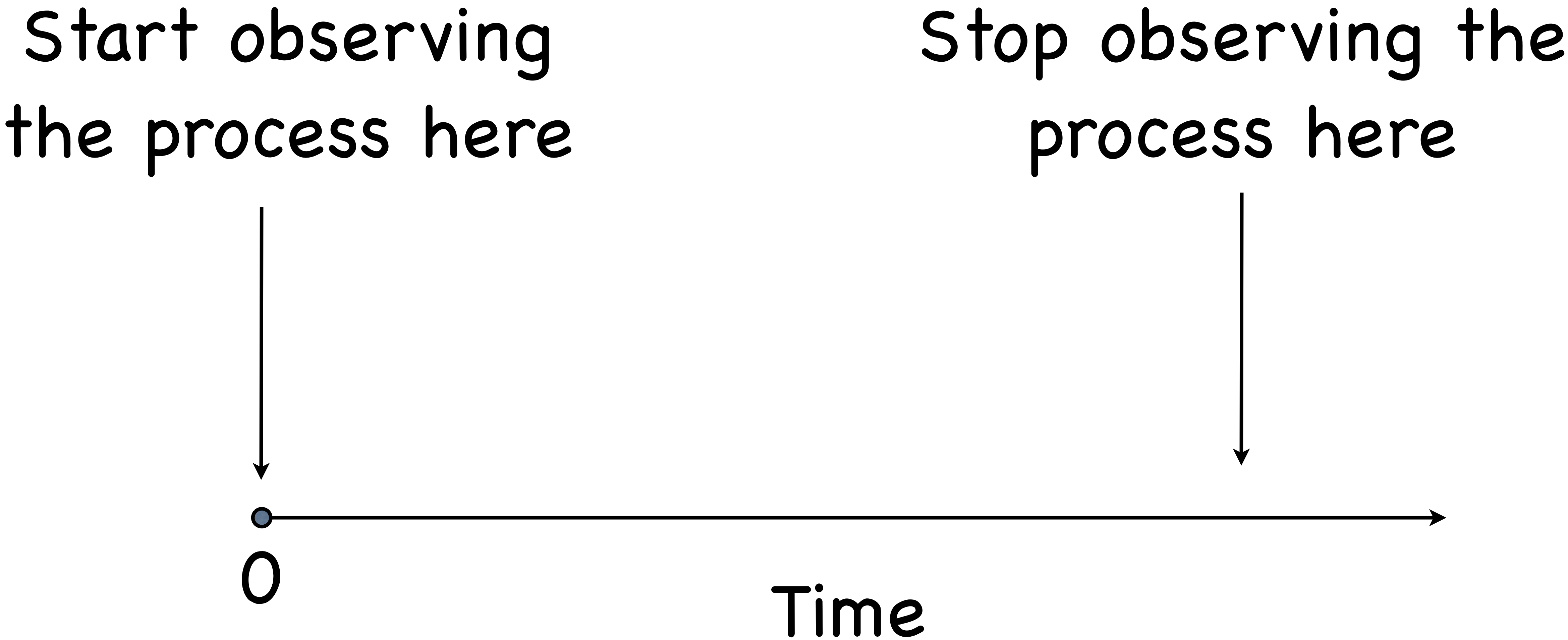

Wait until second event
0
Time


Wait until third event
0
Time

Interesting fact: The waiting time until the k-th event is a gamma-distributed random variable, with parameters k and λ.

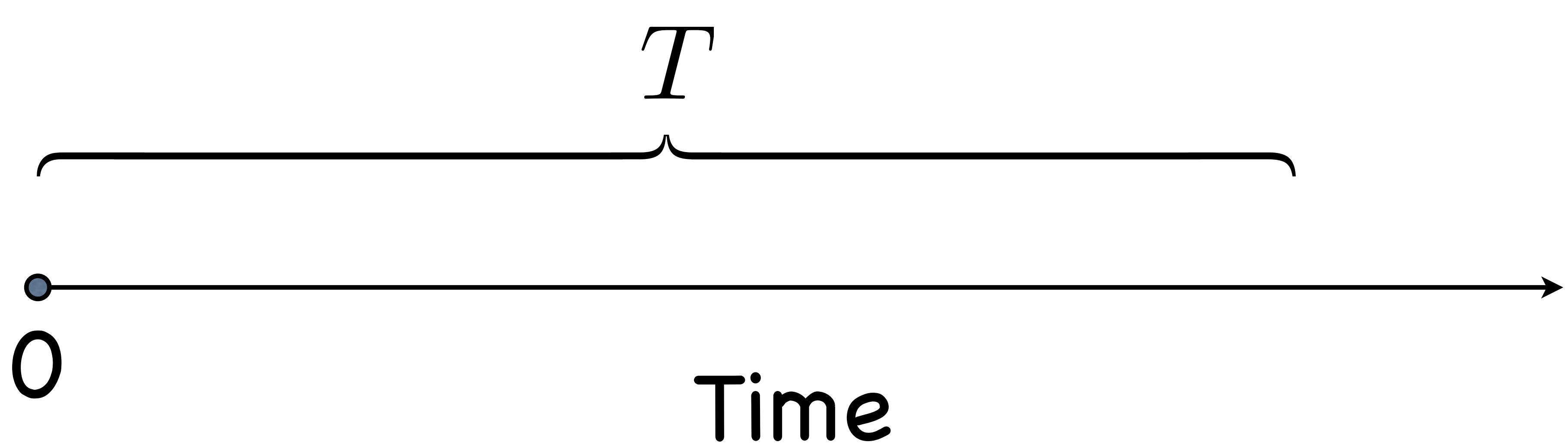
$$f(t) = \frac{\lambda^k}{\Gamma(k)} t^{k-1} e^{-\lambda t}$$

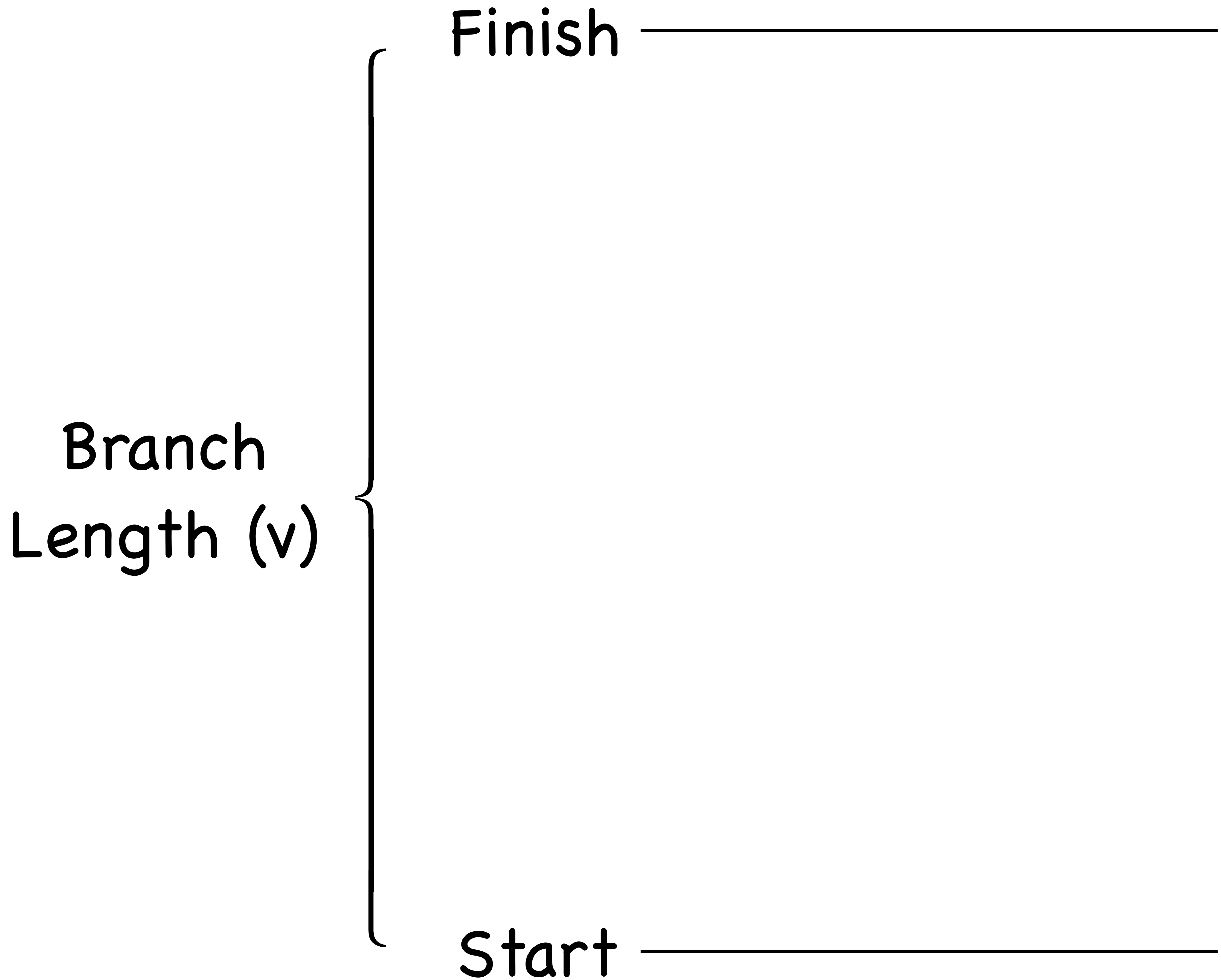
Note: $\Gamma(k) = (k-1)!$ for integer k


Start observing
the process here
Stop observing the
process here
0
Time

Interesting fact: The number of events that occur in the interval T is a Poisson-distributed random variable with parameter λT.

$$\Pr(k \text{ events}) = \frac{e^{-\lambda T}(\lambda T)^k}{k!}$$



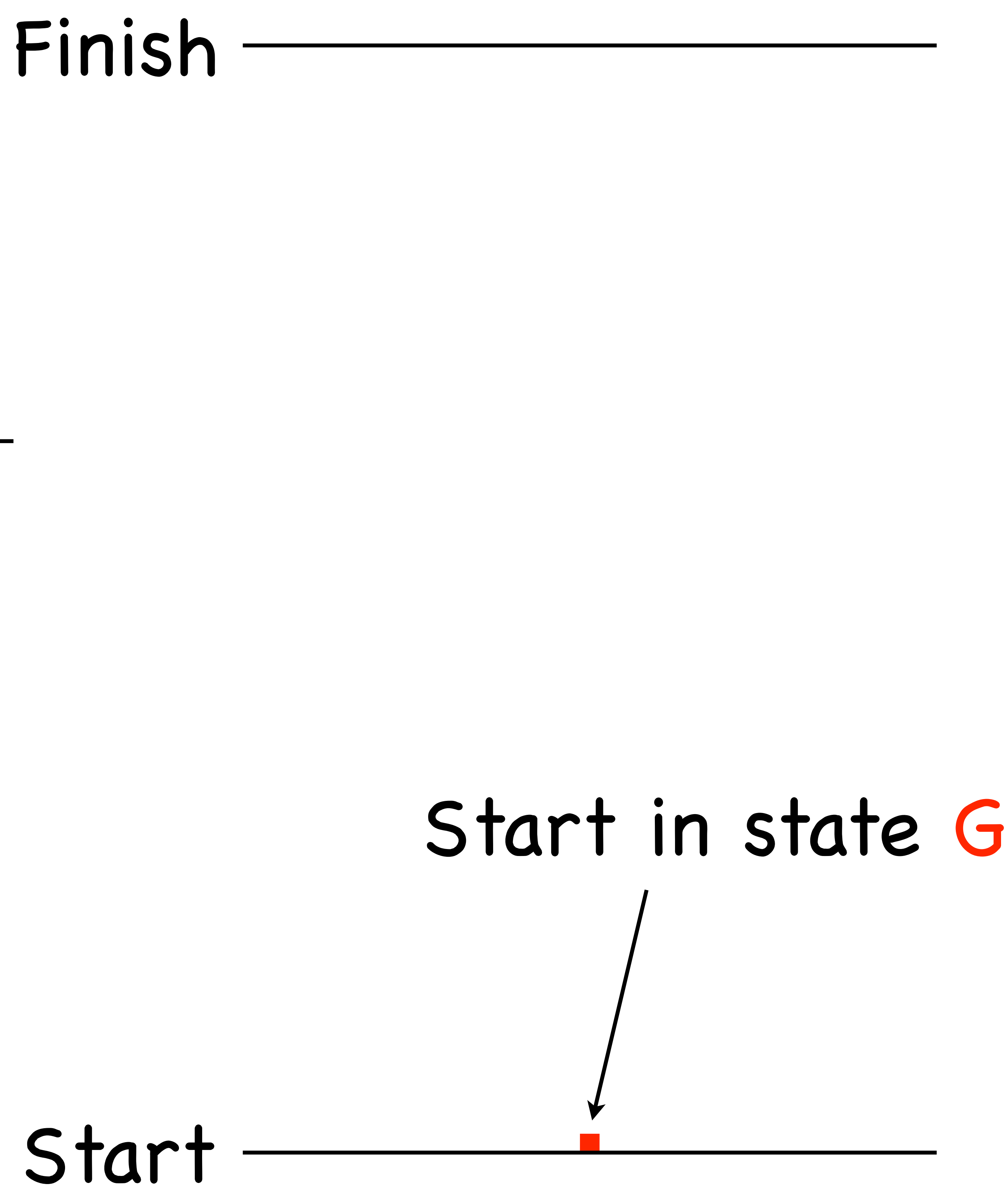

Finish
Branch
Length (v)
Start

Finish

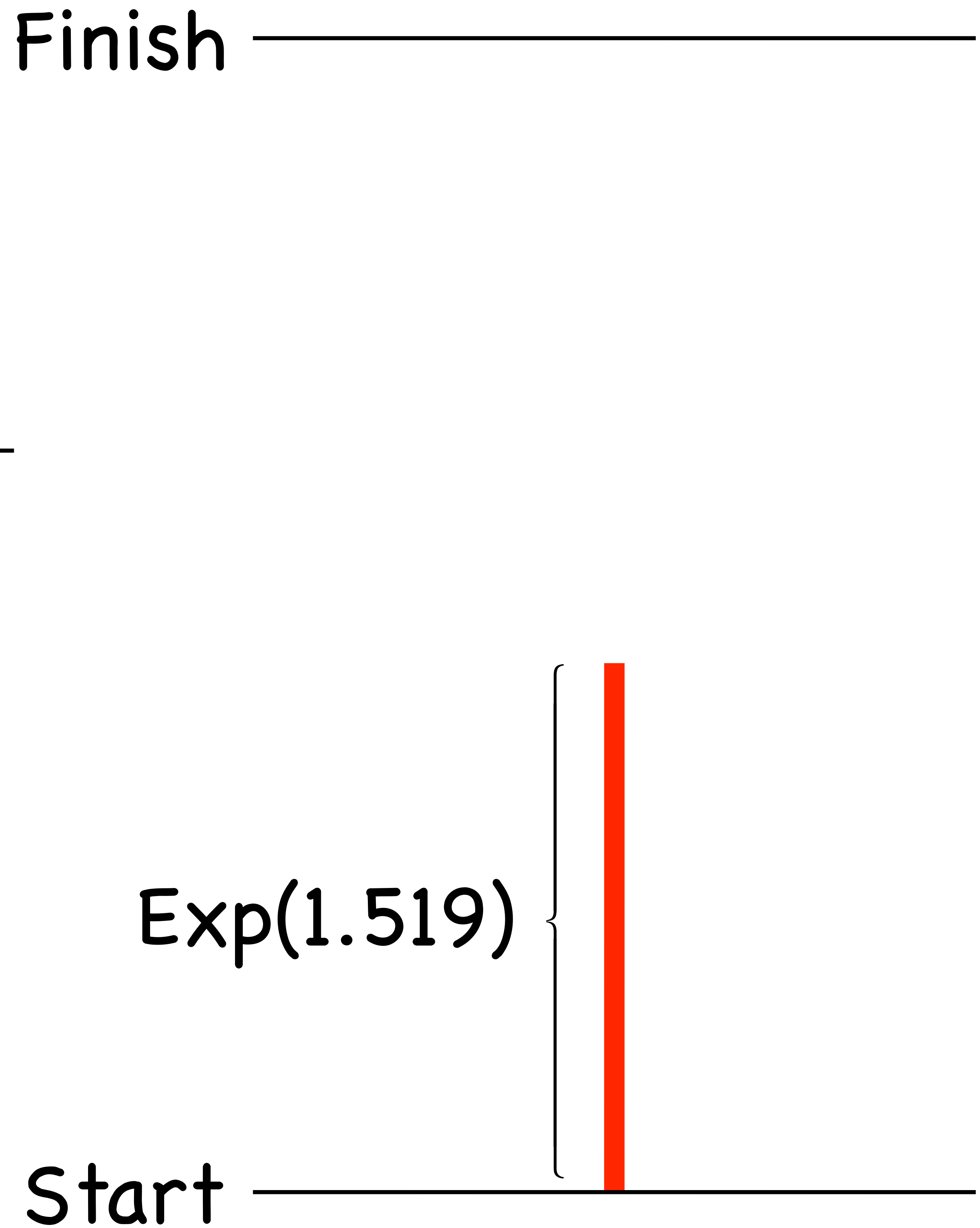
	A	C	G	T
A	−0.886	0.190	0.633	0.063
C	0.253	−0.696	0.127	0.316
G	1.266	0.190	−1.519	0.063
T	0.253	0.949	0.127	−1.329

Start

	A	C	G	T
A	−0.886	0.190	0.633	0.063
C	0.253	−0.696	0.127	0.316
G	1.266	0.190	−1.519	0.063
T	0.253	0.949	0.127	−1.329



	A	C	G	T
A	−0.886	0.190	0.633	0.063
C	0.253	−0.696	0.127	0.316
G	1.266	0.190	−1.519	0.063
T	0.253	0.949	0.127	−1.329



Finish

	A	C	G	T
A	−0.886	0.190	0.633	0.063
C	0.253	−0.696	0.127	0.316
G	1.266	0.190	−1.519	0.063
T	0.253	0.949	0.127	−1.329

$$p_A = \frac{1.266}{1.519} = 0.833$$

$$p_C = \frac{0.190}{1.519} = 0.125$$

$$p_T = \frac{0.063}{1.519} = 0.042$$

Start

Finish

	A	C	G	T
A	−0.886	0.190	0.633	0.063
C	0.253	−0.696	0.127	0.316
G	1.266	0.190	−1.519	0.063
T	0.253	0.949	0.127	−1.329

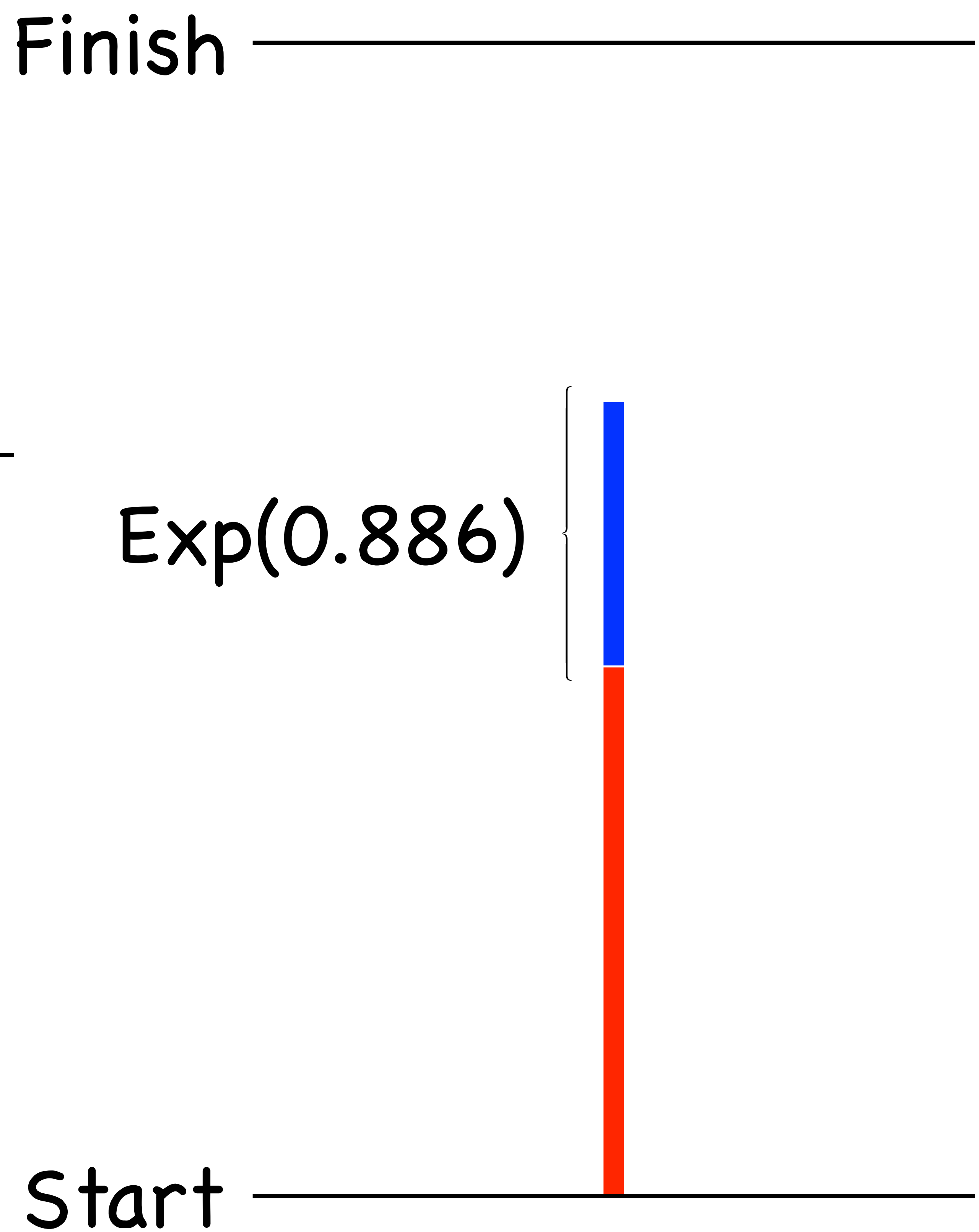
$$p_A = \frac{1.266}{1.519} = 0.833$$

$$p_C = \frac{0.190}{1.519} = 0.125$$

$$p_T = \frac{0.063}{1.519} = 0.042$$

Start

	A	C	G	T
A	−0.886	0.190	0.633	0.063
C	0.253	−0.696	0.127	0.316
G	1.266	0.190	−1.519	0.063
T	0.253	0.949	0.127	−1.329

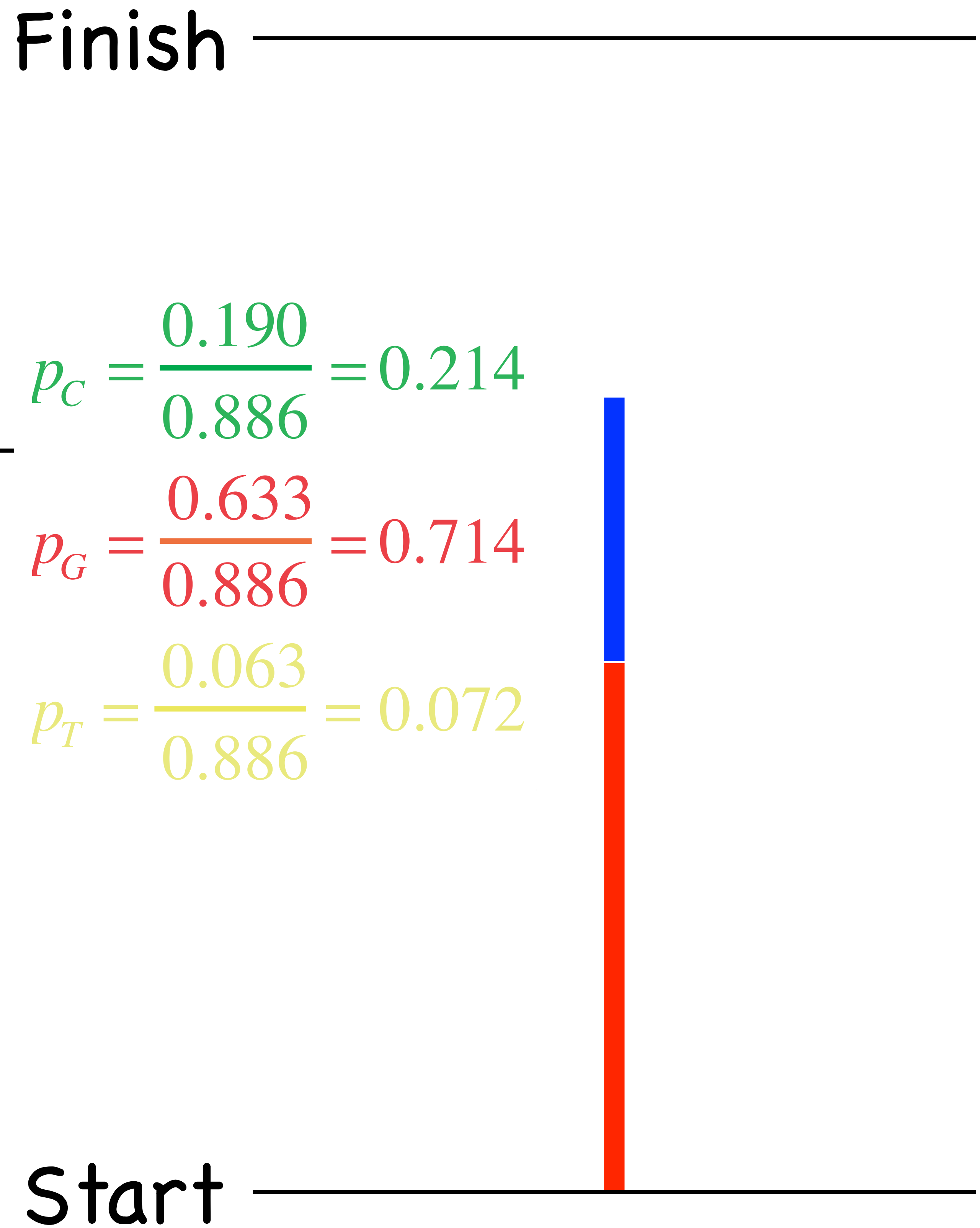


	A	C	G	T
A	−0.886	0.190	0.633	0.063
C	0.253	−0.696	0.127	0.316
G	1.266	0.190	−1.519	0.063
T	0.253	0.949	0.127	−1.329

$$p_C = \frac{0.190}{0.886} = 0.214$$

$$p_G = \frac{0.633}{0.886} = 0.714$$

$$p_T = \frac{0.063}{0.886} = 0.072$$

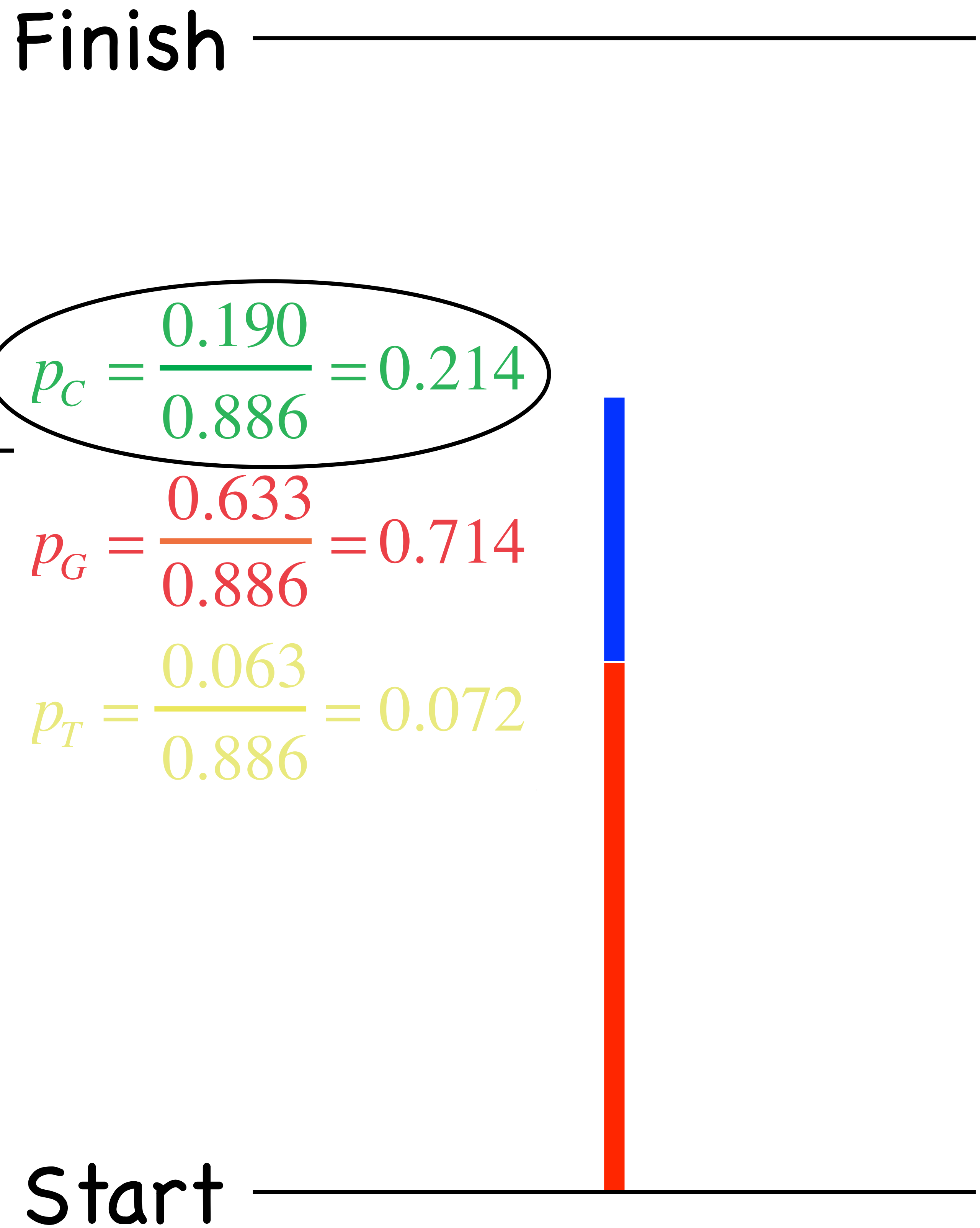


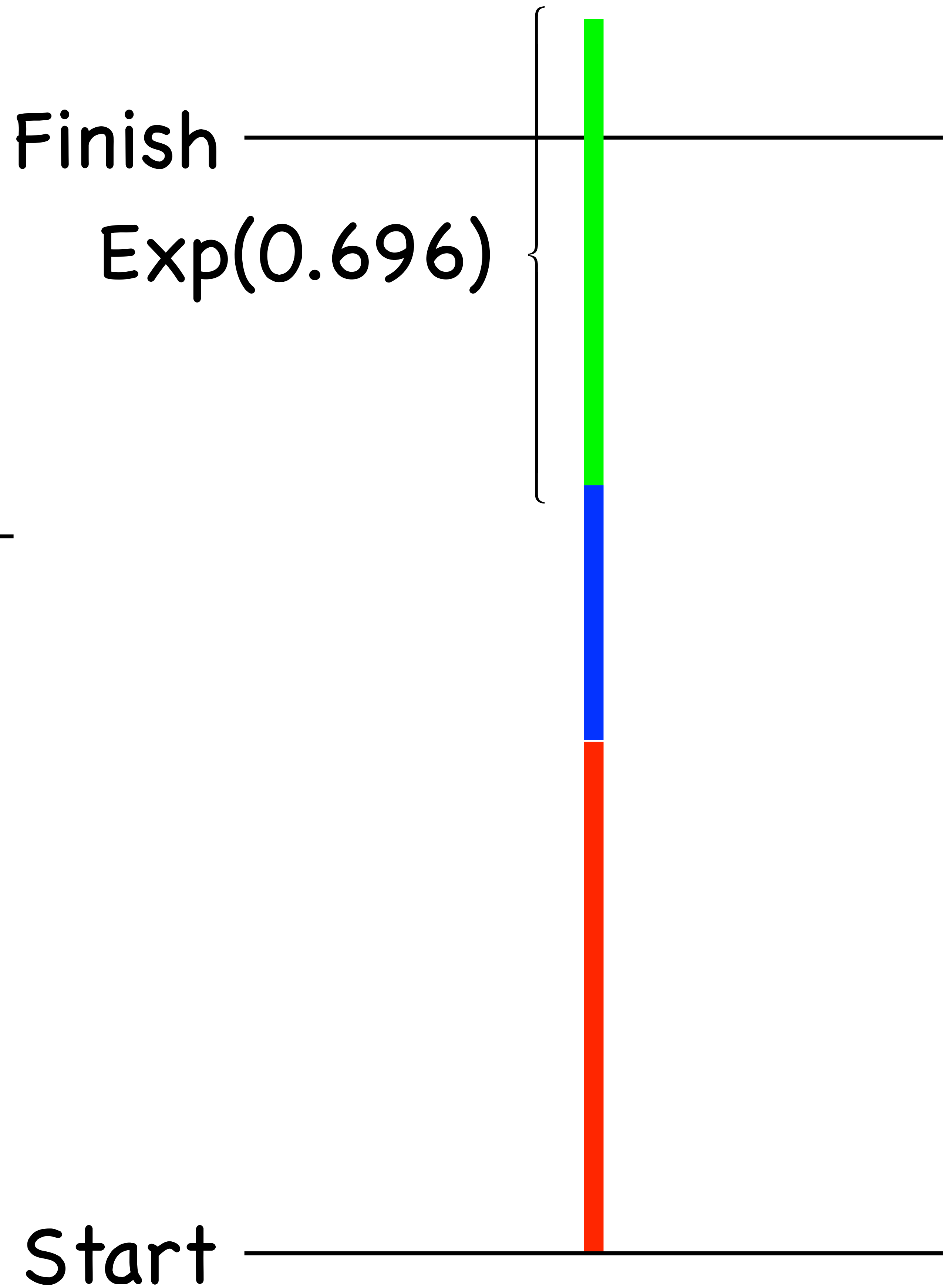
	A	C	G	T
A	−0.886	0.190	0.633	0.063
C	0.253	−0.696	0.127	0.316
G	1.266	0.190	−1.519	0.063
T	0.253	0.949	0.127	−1.329

$$p_C = \frac{0.190}{0.886} = 0.214$$

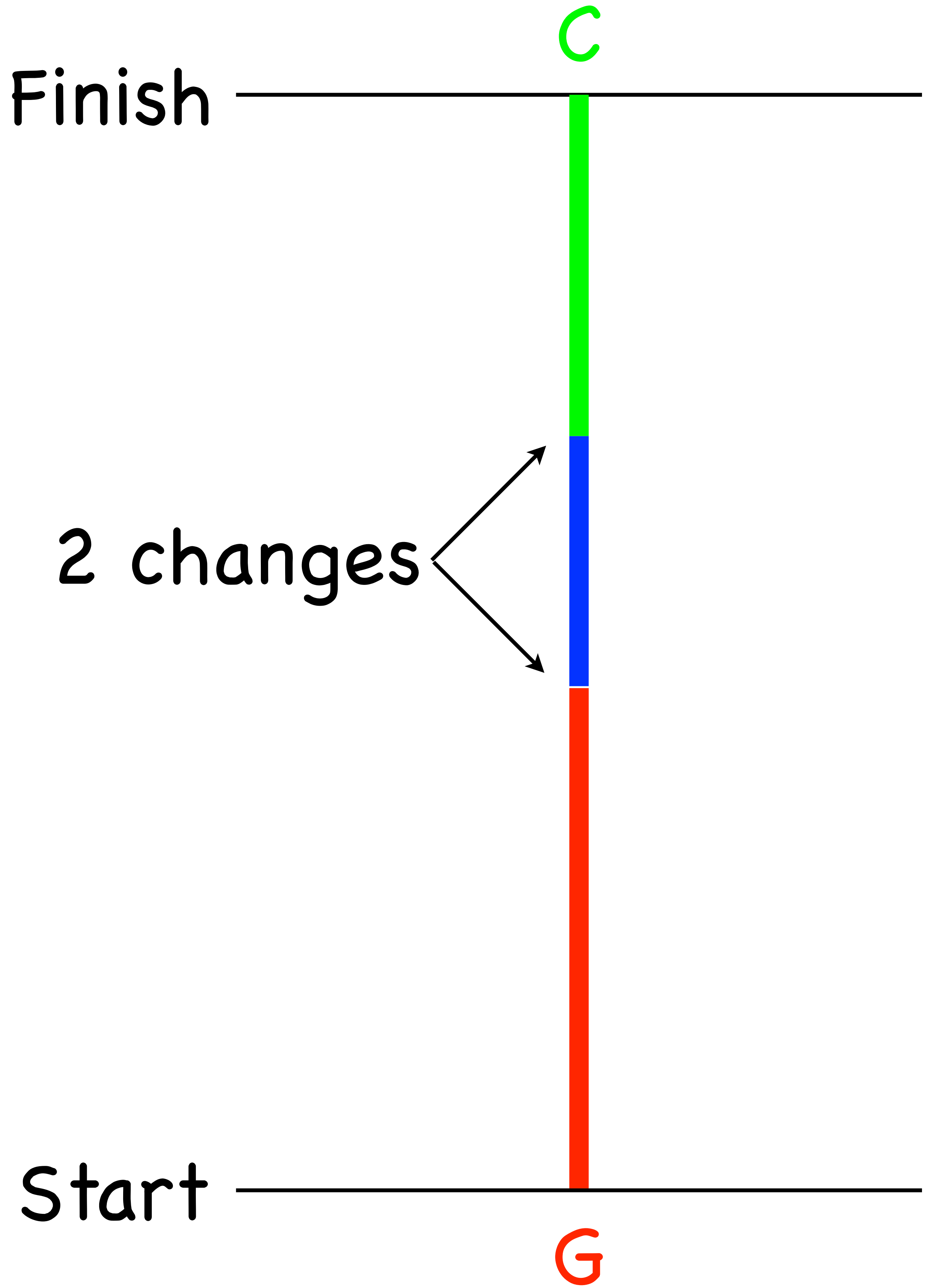
$$p_G = \frac{0.633}{0.886} = 0.714$$

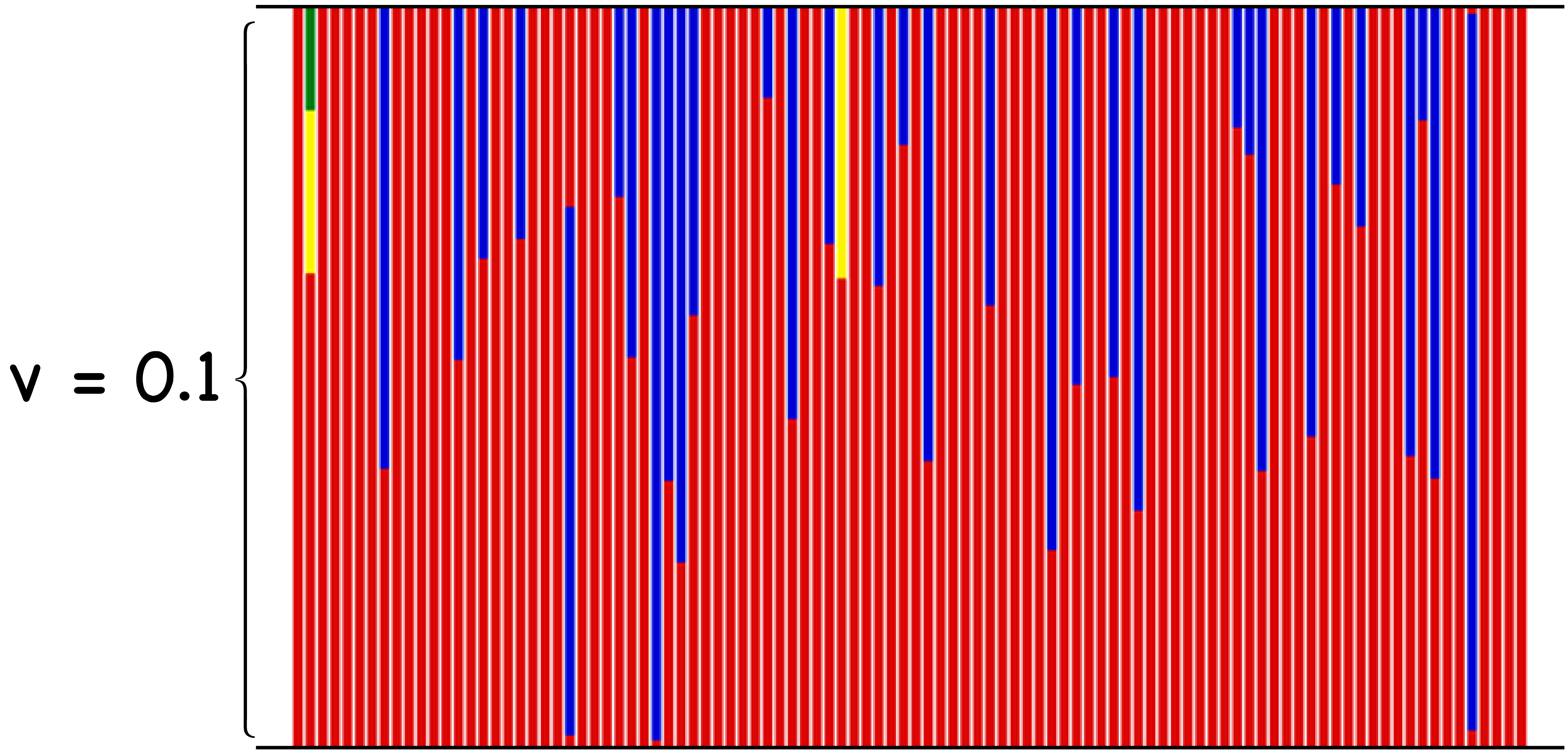
$$p_T = \frac{0.063}{0.886} = 0.072$$





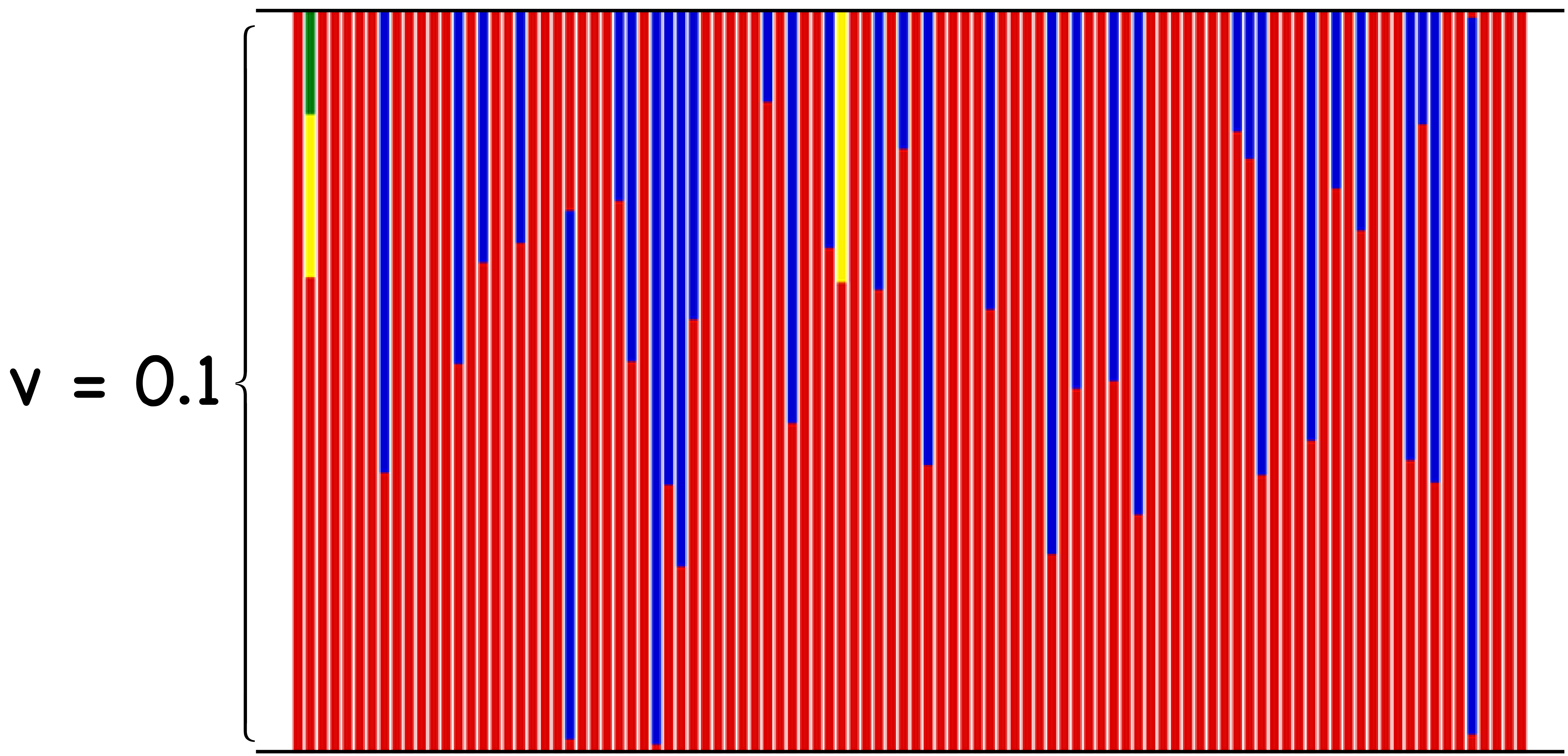
	A	C	G	T
A	−0.886	0.190	0.633	0.063
C	0.253	−0.696	0.127	0.316
G	1.266	0.190	−1.519	0.063
T	0.253	0.949	0.127	−1.329


C
Finish
2 changes
Start
G



Start in G 100 times

End in A 31 times; end in C 1 time; end in G 67 times; end in T 1 time



Start in G 100 times

Transition probabilities for $v = 0.1$

		Ended In			
		A	C	G	T
Started In	A				
	C				
	G	0.31	0.01	0.67	0.01
	T				

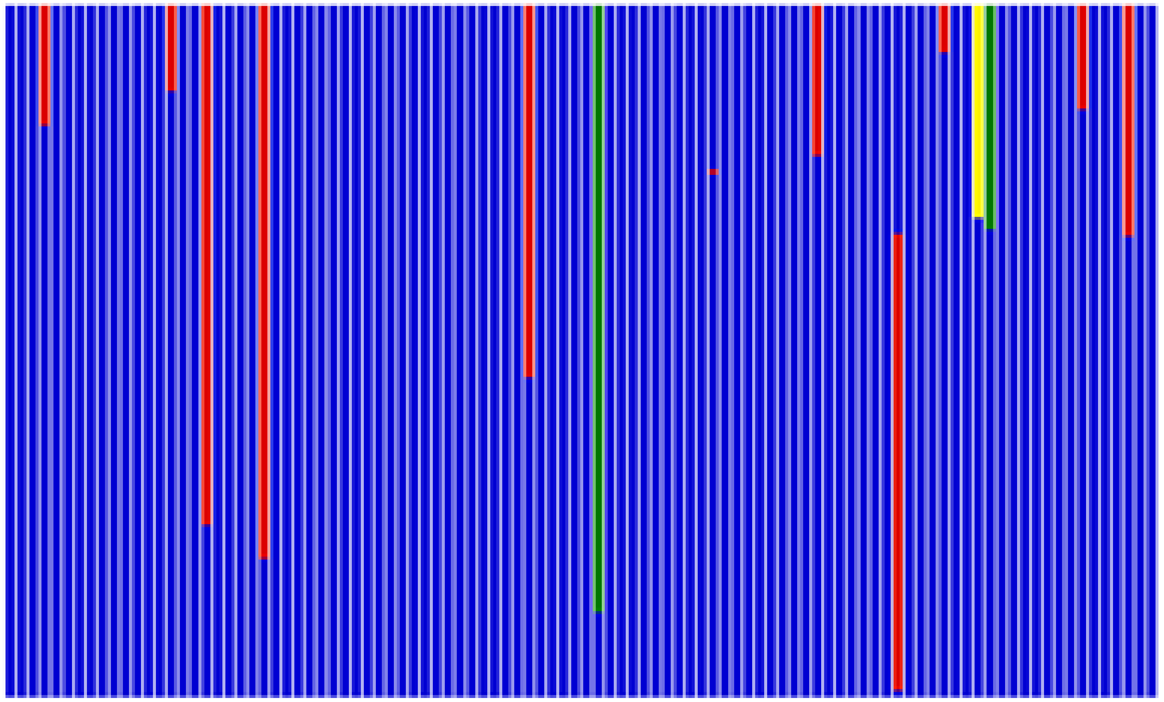
(Monte Carlo estimates of transition probabilities based on a total of 100 simulations)

Transition probabilities for v = 0.1

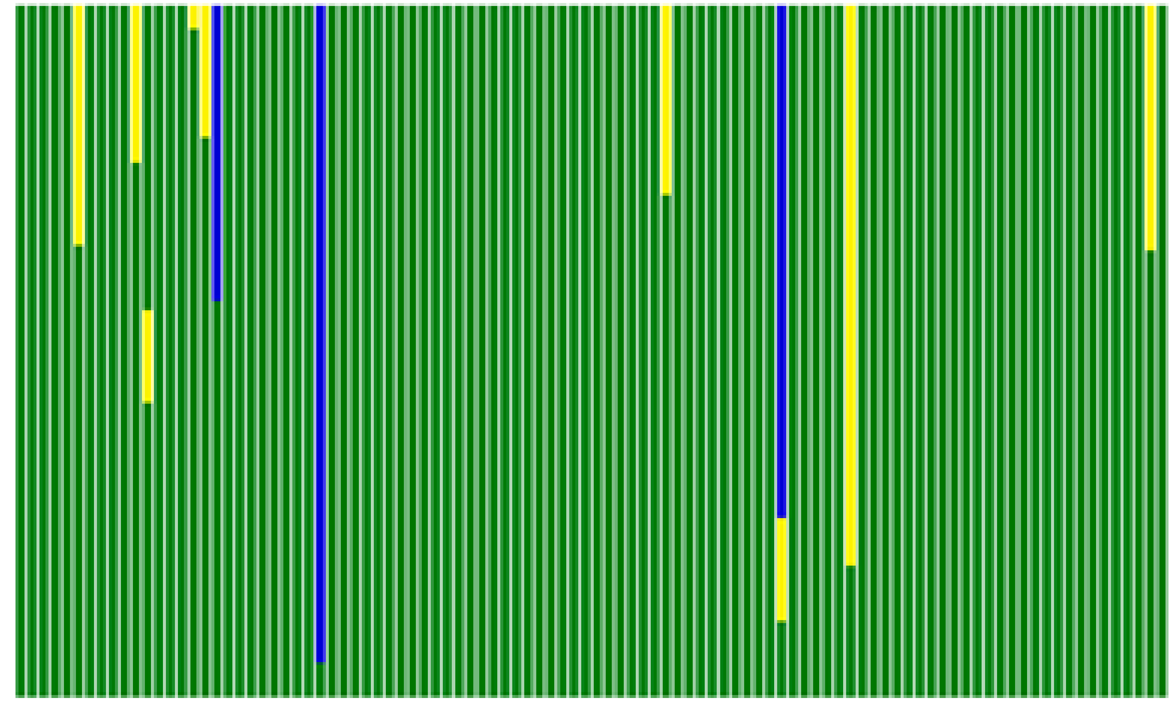
		Ended In			
		A	C	G	T
Started In	A				
	C				
	G	0.1125	0.0182	0.8634	0.0058
	T				

(Monte Carlo estimates of transition probabilities based on a total of 50,000 simulations)

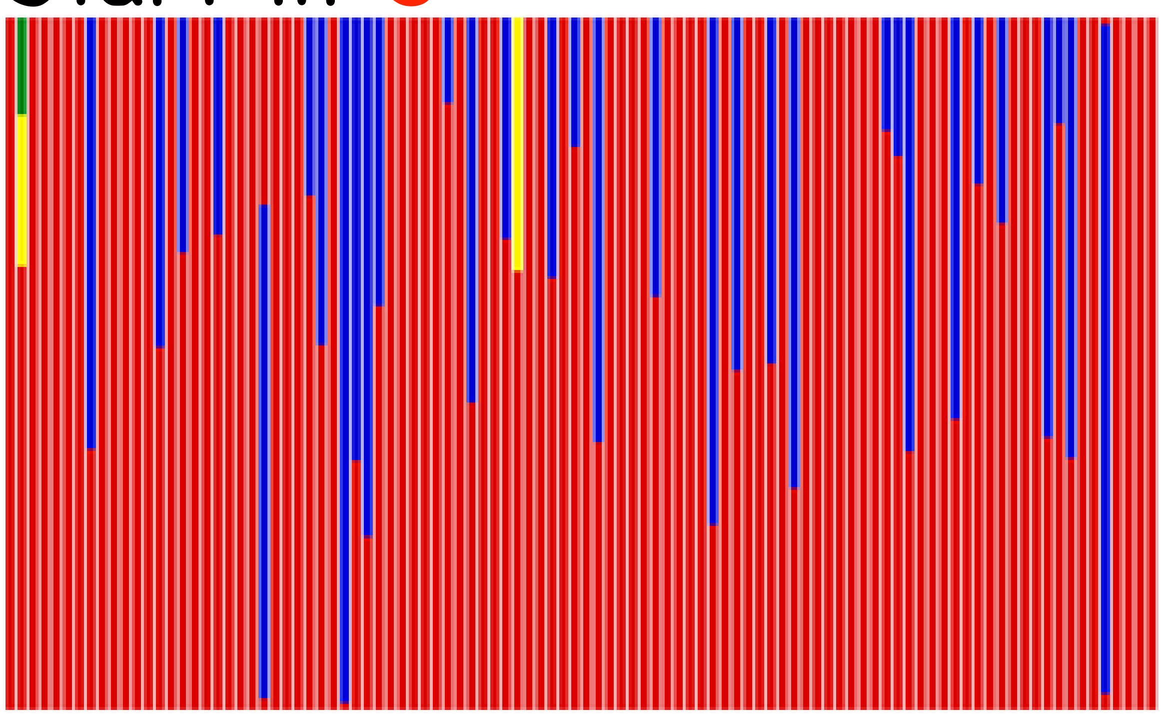
Start in A



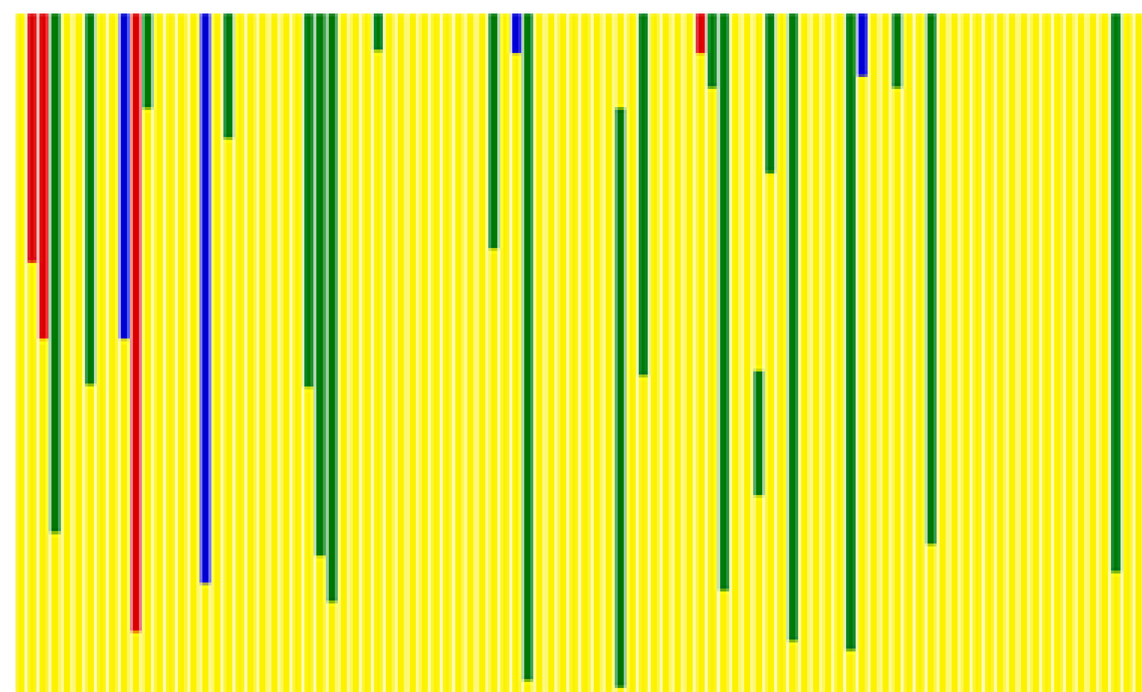
Start in C



Start in G



Start in T



Transition probabilities for v = 0.1

		Ended In			
		A	C	G	T
Started In	A	0.88	0.02	0.09	0.01
	C	0.03	0.90	0.00	0.07
	G	0.31	0.01	0.67	0.01
	T	0.04	0.20	0.04	0.72

(Monte Carlo estimates of transition probabilities based on a total of 100 simulations)

Transition probabilities for v = 0.1

Started In \ Ended In	A	C	G	T
A	0.9180	0.0182	0.0577	0.0060
C	0.0249	0.9346	0.0125	0.0279
G	0.1125	0.0182	0.8634	0.0058
T	0.0241	0.0877	0.0113	0.8767

(Monte Carlo estimates of transition probabilities based on a total of 50,000 simulations)

Monte Carlo
(50,000 reps)

Ended In

Started In	A	C	G	T
A	0.9180	0.0182	0.0577	0.0060
C	0.0249	0.9346	0.0125	0.0279
G	0.1125	0.0182	0.8634	0.0058
T	0.0241	0.0877	0.0113	0.8767

Exact: $\mathbf{P}(t) = e^{\mathbf{Q}t}$

Ended In

Started In	A	C	G	T
A	0.9191	0.0184	0.0563	0.0061
C	0.0245	0.9344	0.0123	0.0287
G	0.1127	0.0183	0.8627	0.0061
T	0.0245	0.0862	0.0123	0.8770

Monte Carlo
(50,000 reps)

Started In / Ended In

	A	C	G	T
A	0.9180	0.0182	0.0577	0.0060
C	0.0249	0.9346	0.0125	0.0279
G	0.1125	0.0182	0.8634	0.0058
T	0.0241	0.0877	0.0113	0.8767

Exact: $\mathbf{P}(t) = e^{\mathbf{Q}t}$

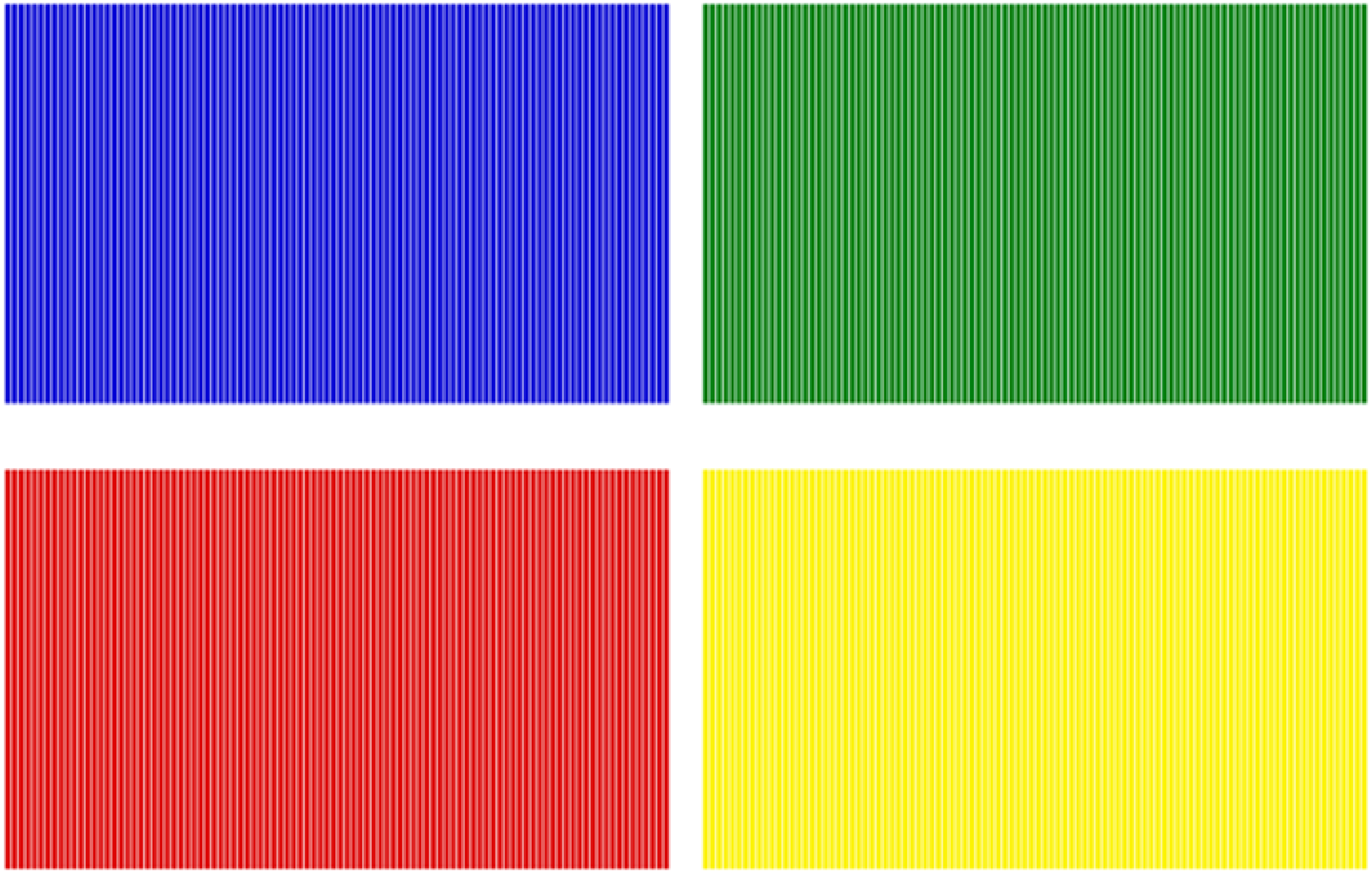
Accounts for all the ways that the process, starting in state i, can end in state j.

Started In / Ended In

	A	C	G	T
A				
C				
G				
T	0.0245	0.0862	0.0123	0.8770

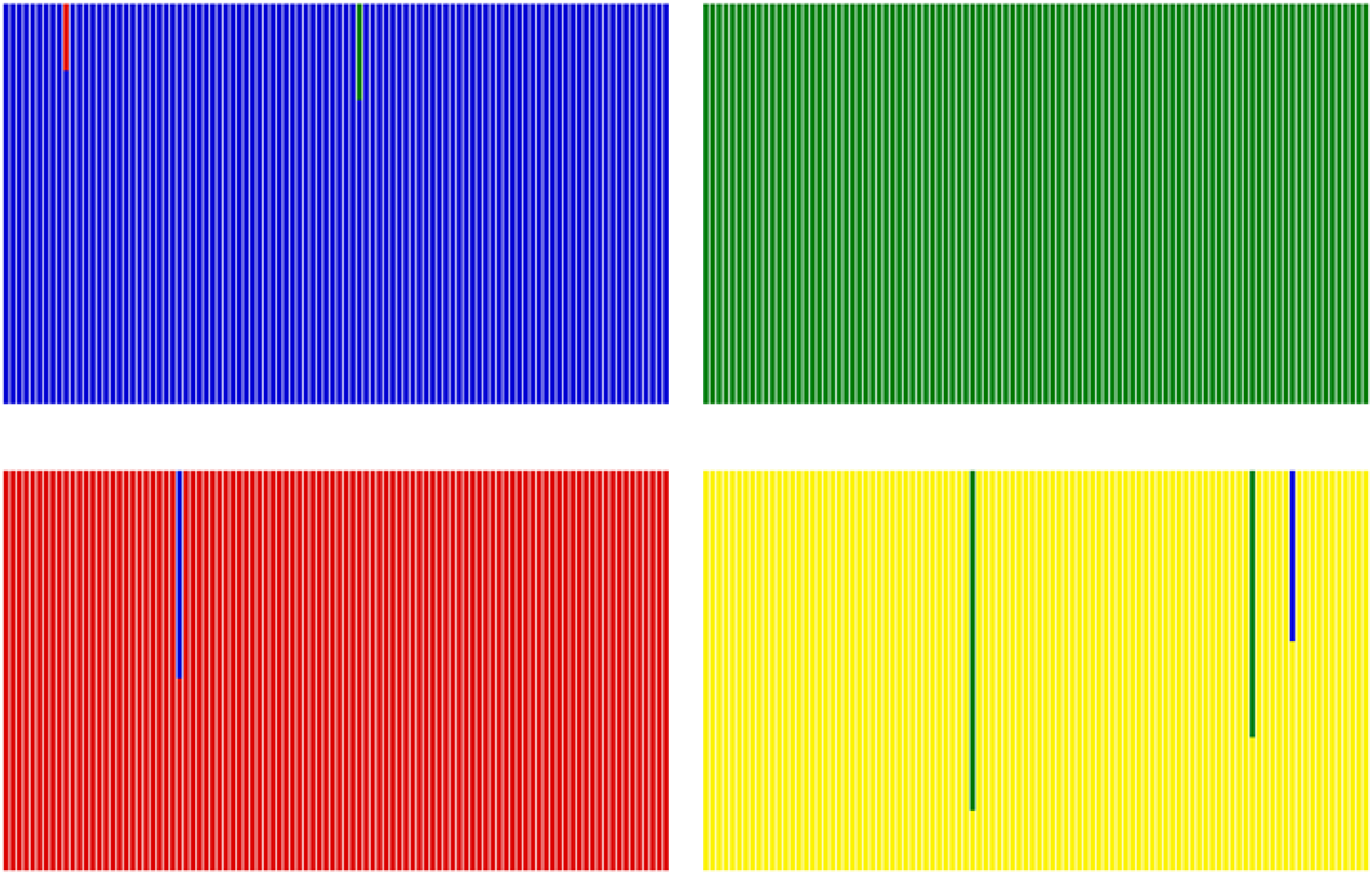
Transition probabilities for any rate matrix, **Q**, can be calculated as

$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$



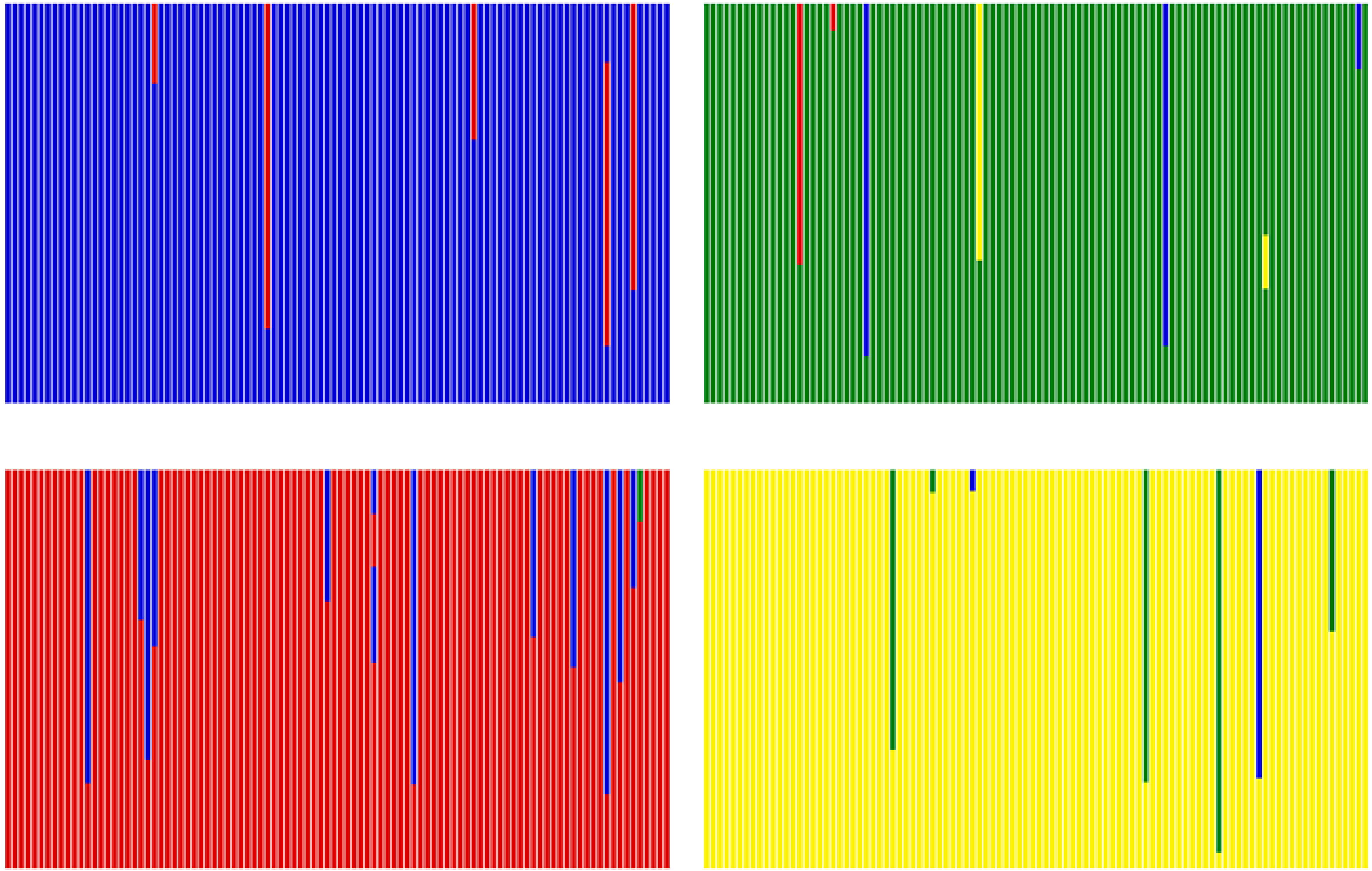
$P(0.00) =$

	A	C	G	T
A	1.0000	0.0000	0.0000	0.0000
C	0.0000	1.0000	0.0000	0.0000
G	0.0000	0.0000	1.0000	0.0000
T	0.0000	0.0000	0.0000	1.0000



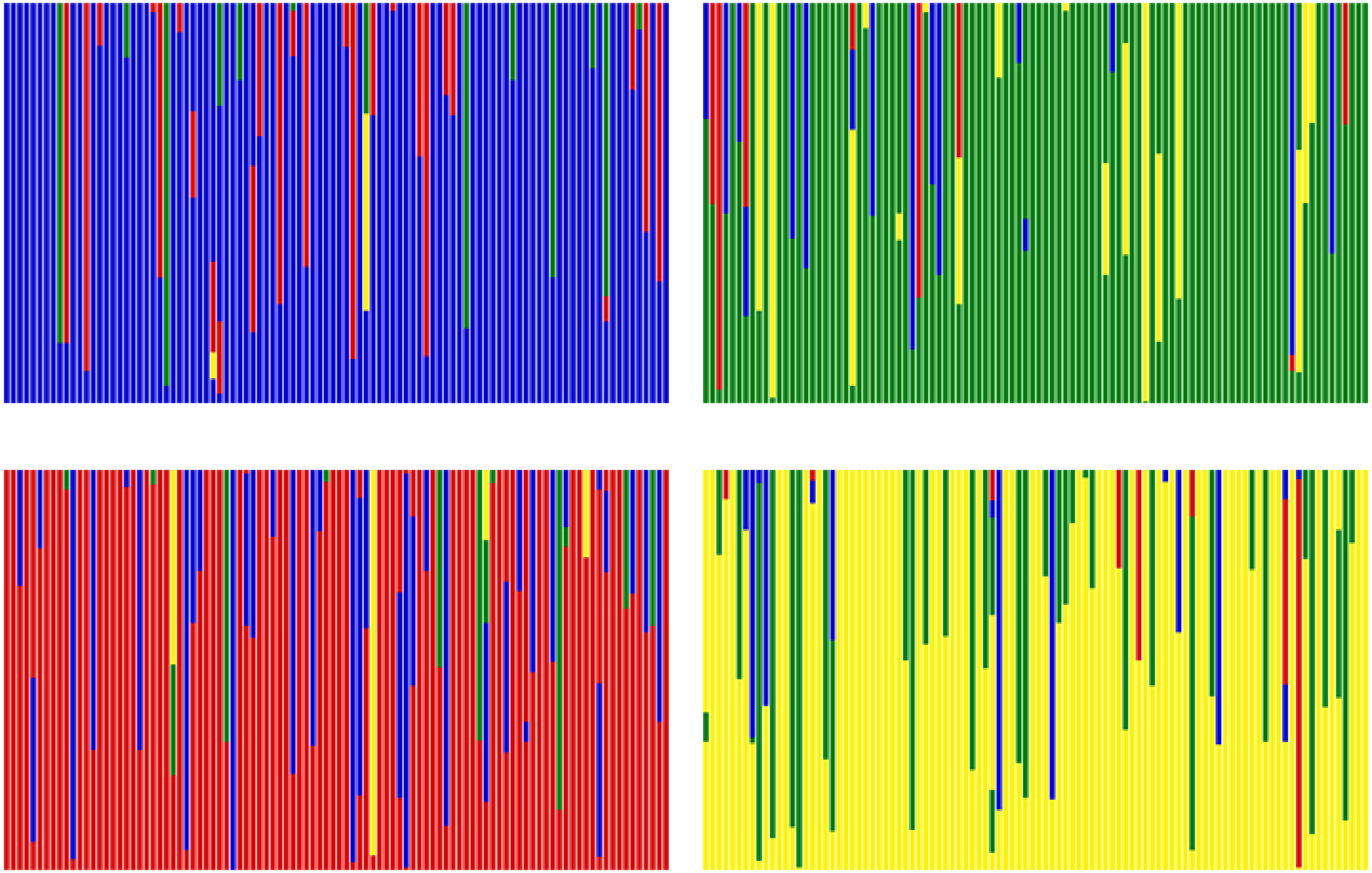
$P(0.01)$ =

	A	C	G	T
A	0.9912	0.0019	0.0062	0.0006
C	0.0025	0.9931	0.0013	0.0031
G	0.0125	0.0019	0.9849	0.0006
T	0.0025	0.0094	0.0013	0.9868



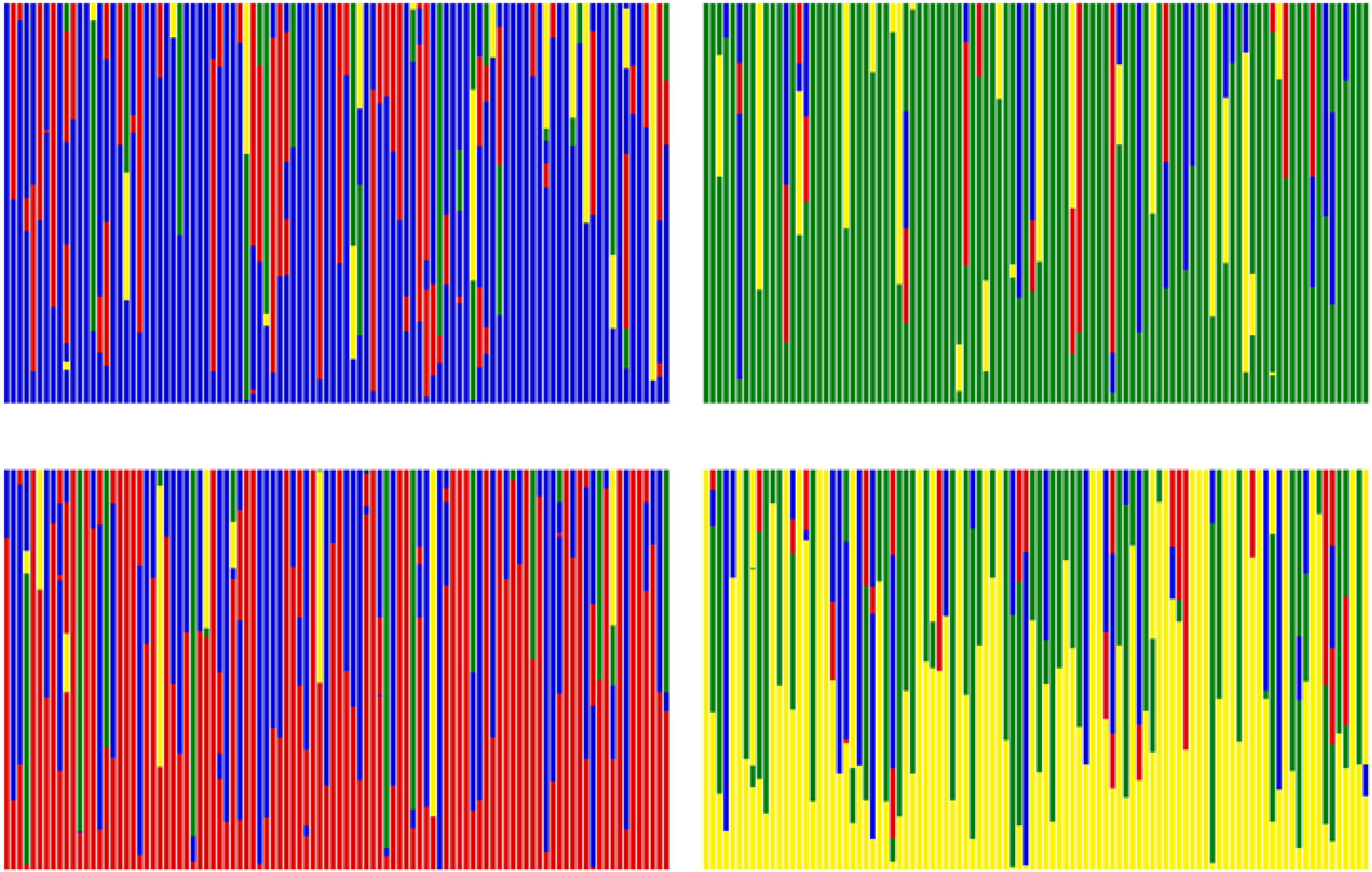
P(0.10) =

	A	C	G	T
A	0.9191	0.0183	0.0563	0.0061
C	0.0243	0.9344	0.0122	0.0287
G	0.1127	0.0184	0.8627	0.0061
T	0.0245	0.0861	0.0122	0.8770



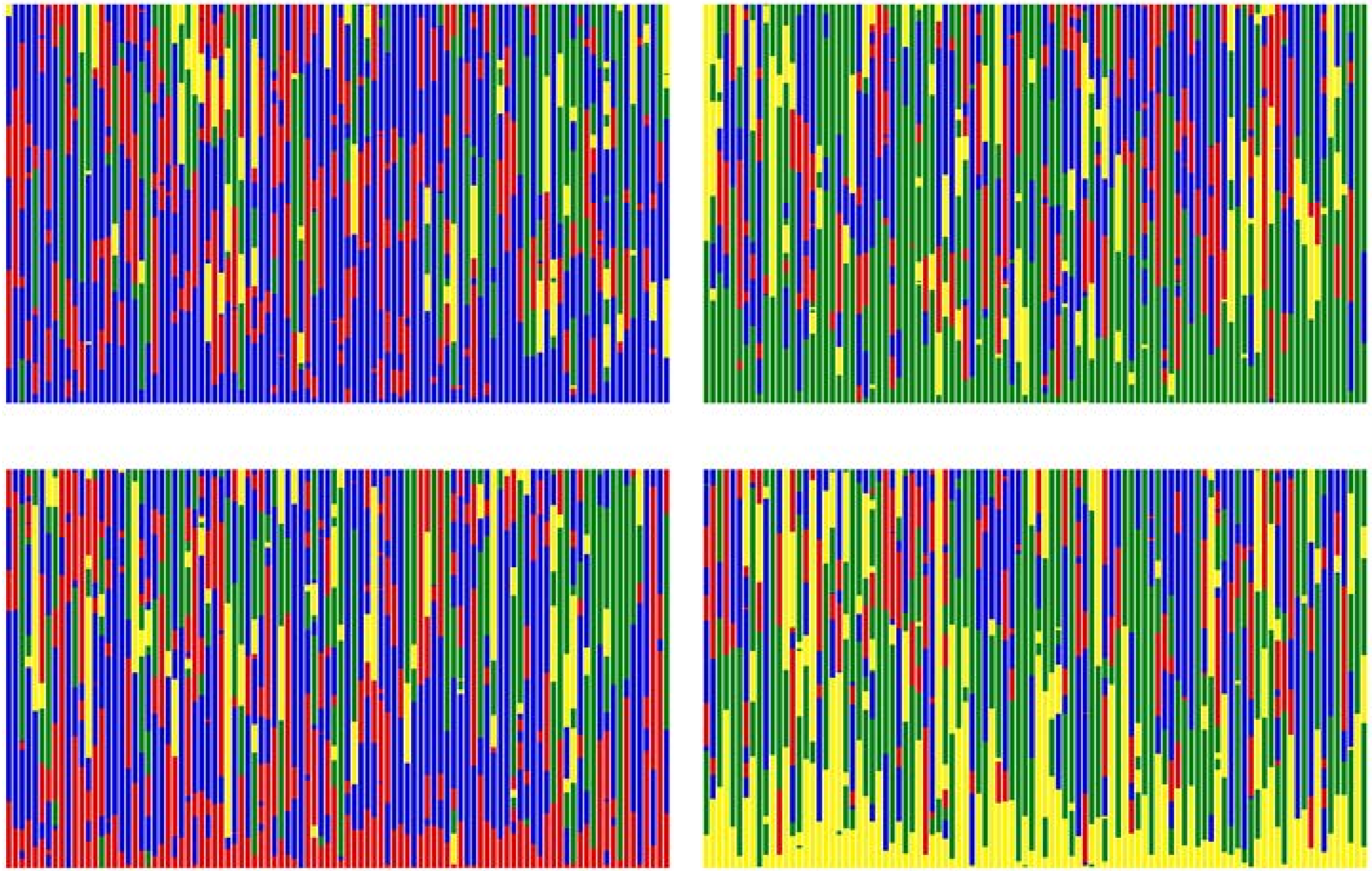
$P(0.50) =$

	A	C	G	T
A	0.7079	0.0813	0.1835	0.0271
C	0.1085	0.7377	0.0542	0.0995
G	0.3670	0.0813	0.5244	0.0271
T	0.1085	0.2985	0.0542	0.5387



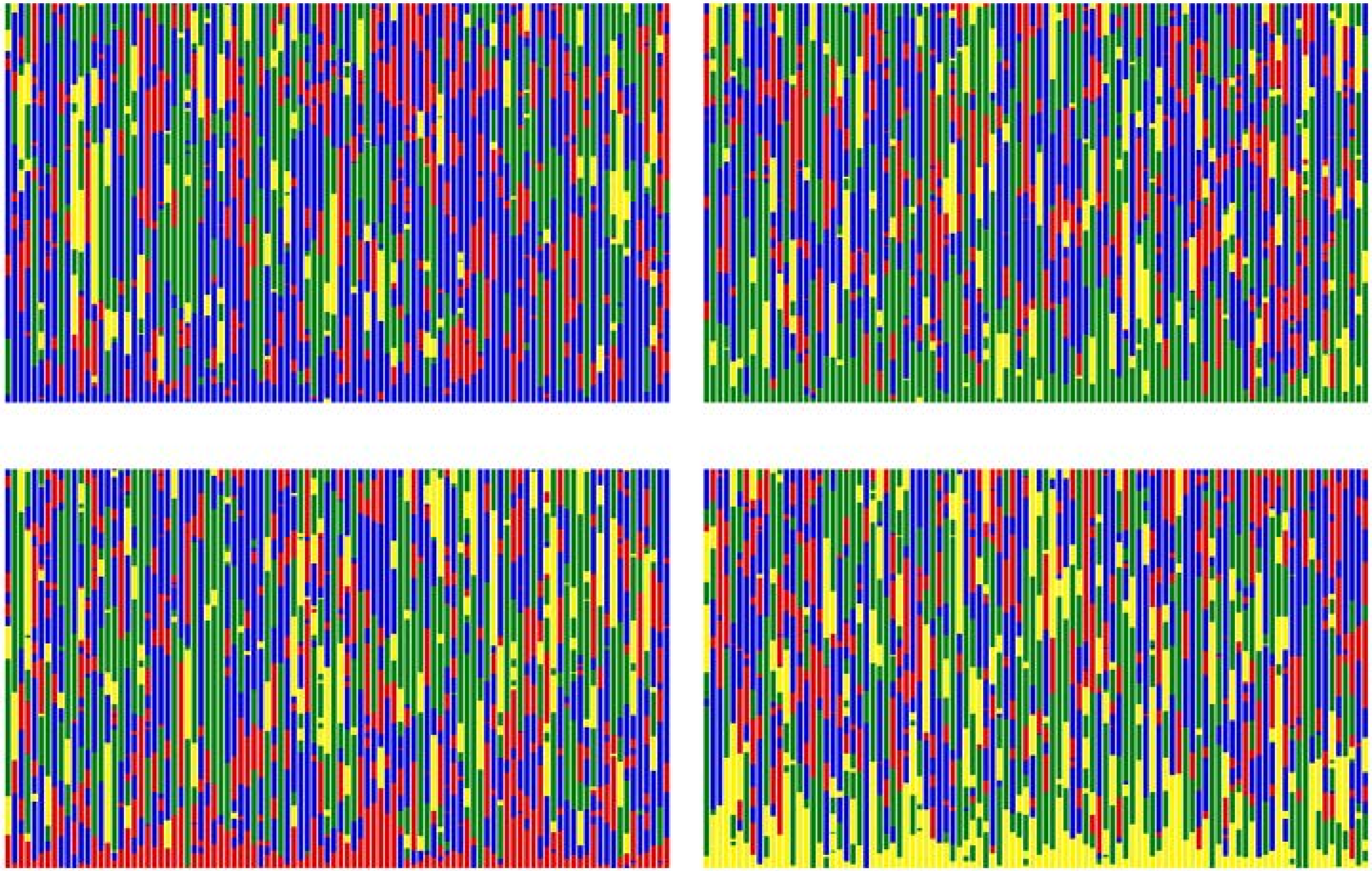
$$P(1.00) =$$

	A	C	G	T
A	0.5803	0.1406	0.2320	0.0468
C	0.1875	0.5871	0.0937	0.1314
G	0.4641	0.1406	0.3483	0.0468
T	0.1875	0.3942	0.0937	0.3243

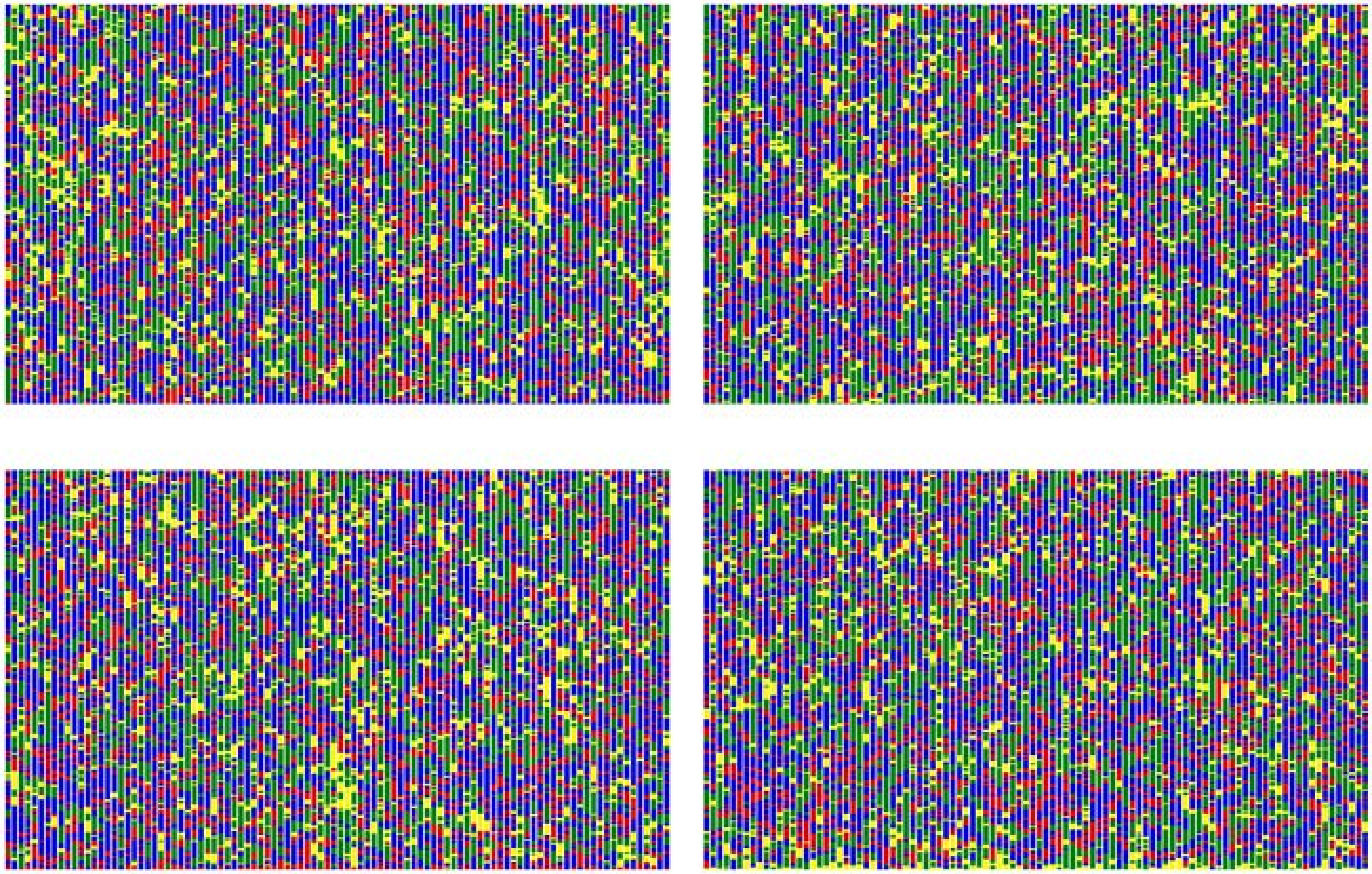


$P(5.00) =$

	A	C	G	T
A	0.4113	0.2873	0.2056	0.0957
C	0.3831	0.3190	0.1915	0.1062
G	0.4112	0.2873	0.2056	0.0957
T	0.3831	0.3188	0.1915	0.1065

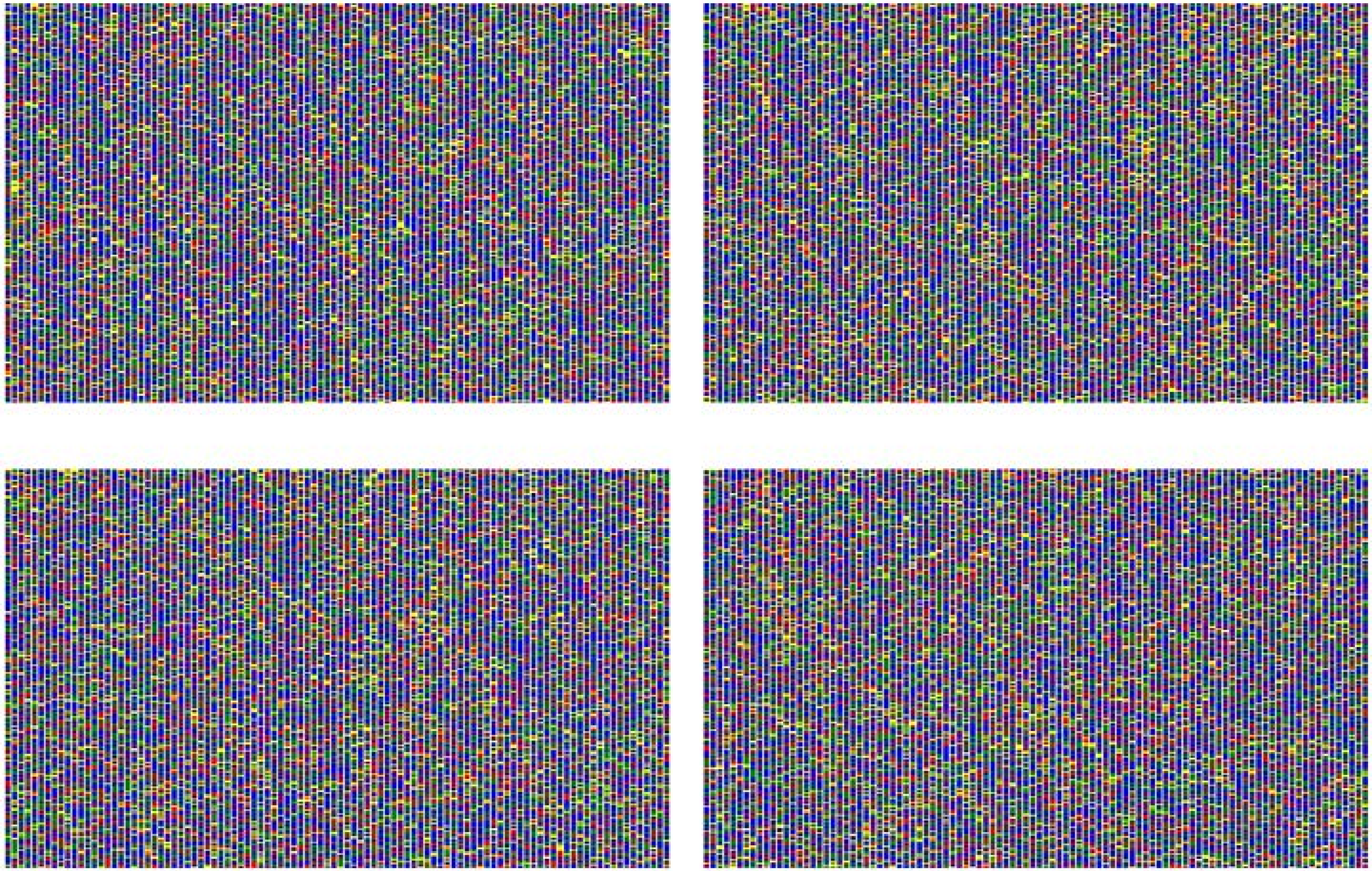


$$P(10.00) = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & 0.4005 & 0.2994 & 0.2002 & 0.0998 \\ C & 0.3992 & 0.3008 & 0.1996 & 0.1002 \\ G & 0.4005 & 0.2994 & 0.2002 & 0.0998 \\ T & 0.3992 & 0.3008 & 0.1996 & 0.1002 \end{array}$$



$P(100.00) =$

	A	C	G	T
A	0.4000	0.3000	0.2000	0.1000
C	0.4000	0.3000	0.2000	0.1000
G	0.4000	0.3000	0.2000	0.1000
T	0.4000	0.3000	0.2000	0.1000



$$\mathbf{P}(1000.00) = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & 0.4000 & 0.3000 & 0.2000 & 0.1000 \\ C & 0.4000 & 0.3000 & 0.2000 & 0.1000 \\ G & 0.4000 & 0.3000 & 0.2000 & 0.1000 \\ T & 0.4000 & 0.3000 & 0.2000 & 0.1000 \end{array}$$

Stationary probabilities (also called equilibrium frequencies, prior probabilities) are the probabilities of finding the process in the different states after an infinite amount of time.

$$\pi_A$$

$$\pi_C$$

$$\pi_G$$

$$\pi_T$$

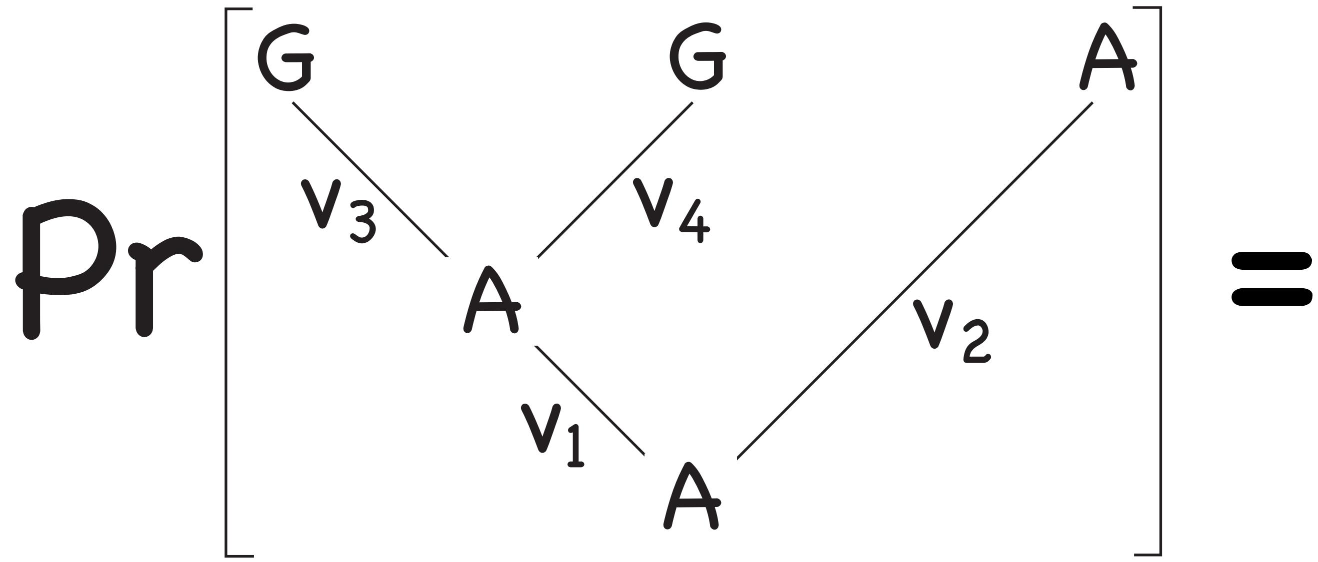
Stationary probabilities (also called equilibrium frequencies, prior probabilities) are the probabilities of finding the process in the different states after an infinite amount of time.

$$\pi_A = 0.4$$

$$\pi_C = 0.3$$

$$\pi_G = 0.2$$

$$\pi_T = 0.1$$



$$\pi_A \times p_{AA}(v_1) \times p_{AA}(v_2) \times p_{AG}(v_3) \times p_{AG}(v_4)$$

π_i – Stationary frequencies

$p_{ij}(v)$ – Transition probabilities

$$\mathbf{Q} = \begin{pmatrix} - & \pi_C & \kappa\,\pi_G & \pi_T \\ \pi_A & - & \pi_G & \kappa\,\pi_T \\ \kappa\,\pi_A & \pi_C & - & \pi_T \\ \pi_A & \kappa\,\pi_C & \pi_G & - \end{pmatrix} \mu$$

$$\kappa = 5$$

$$\pi_A = 0.4$$

$$\pi_C = 0.3$$

$$\pi_G = 0.2$$

$$\pi_T = 0.1$$

$$\mathbf{Q} = \begin{pmatrix} -0.886 & 0.190 & 0.633 & 0.063 \\ 0.253 & -0.696 & 0.127 & 0.316 \\ 1.266 & 0.190 & -1.519 & 0.063 \\ 0.253 & 0.949 & 0.127 & -1.329 \end{pmatrix}$$

Jukes & Cantor (1969)

$$Q = \begin{pmatrix} -1 & 1/3 & 1/3 & 1/3 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{pmatrix}$$

Kimura (1980)

$$Q = \begin{pmatrix} -1 & 1/(\kappa+2) & \kappa/(\kappa+2) & 1/(\kappa+2) \\ 1/(\kappa+2) & -1 & 1/(\kappa+2) & \kappa/(\kappa+2) \\ \kappa/(\kappa+2) & 1/(\kappa+2) & -1 & 1/(\kappa+2) \\ 1/(\kappa+2) & \kappa/(\kappa+2) & 1/(\kappa+2) & -1 \end{pmatrix}$$

Hasegawa, Kishino, and Yano (1985)

$$Q = \begin{pmatrix} - & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & - & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & - & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & - \end{pmatrix}\mu$$

GTR (Tavare, 1986)

$$Q = \begin{pmatrix} - & r_{AC}\pi_C & r_{AG}\pi_G & r_{AT}\pi_T \\ r_{AC}\pi_A & - & r_{CG}\pi_G & r_{CT}\pi_T \\ r_{AG}\pi_A & r_{CG}\pi_C & - & \pi_T \\ r_{AT}\pi_A & r_{CT}\pi_C & \pi_G & - \end{pmatrix}\mu$$

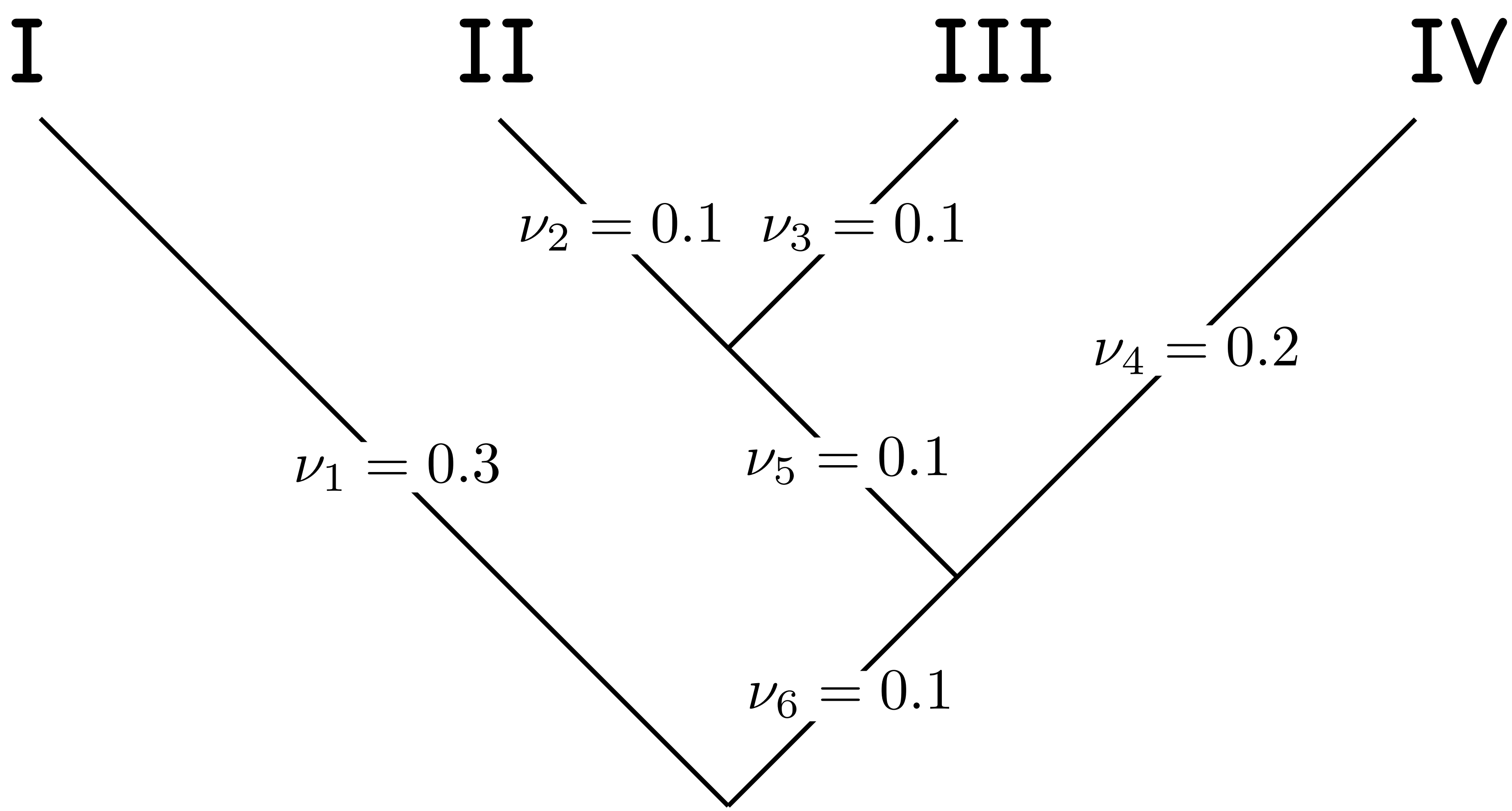
The most general nucleotide model possible is not necessarily time-reversible

$$Q = \begin{pmatrix} - & r_{AC} & r_{AG} & r_{AT} \\ r_{CA} & - & r_{CG} & r_{CT} \\ r_{GA} & r_{GC} & - & 1 \\ r_{TA} & r_{TC} & r_{TG} & - \end{pmatrix} \mu$$

and has 11 parameters

Dice

		To			
		A	C	G	T
From	A	−0.886	0.190	0.633	0.063
	C	0.253	−0.696	0.127	0.316
	G	1.266	0.190	−1.519	0.063
	T	0.253	0.949	0.127	−1.329



Pattern Probabilities (I)

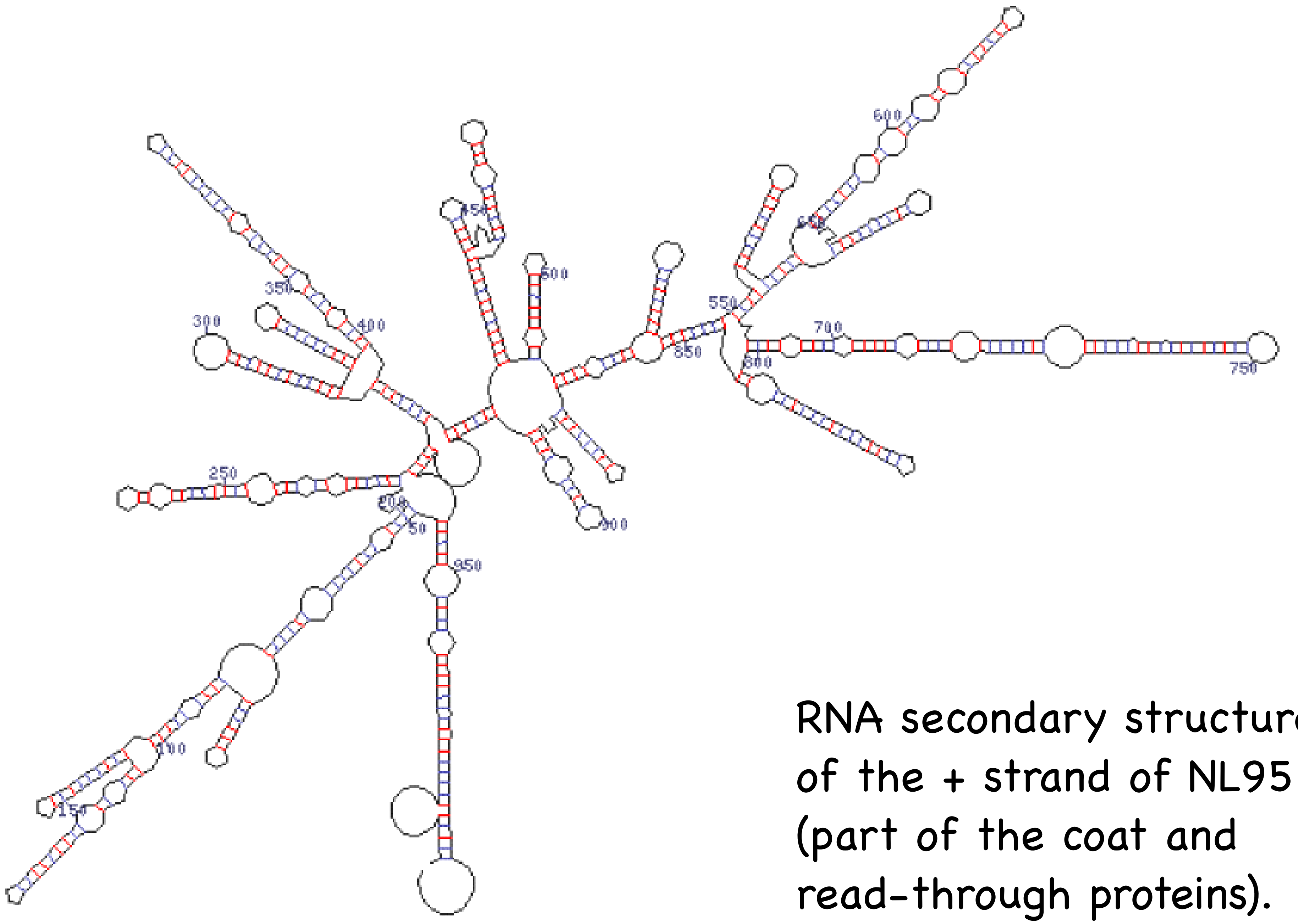
AAAA -- 0.199465	AGAA -- 0.014711	CAAA -- 0.018317	CGAA -- 0.001490
AAAC -- 0.004185	AGAC -- 0.000725	CAAC -- 0.000628	CGAC -- 0.000210
AAAG -- 0.014711	AGAG -- 0.019868	CAAG -- 0.001490	CGAG -- 0.002878
AAAT -- 0.001395	AGAT -- 0.000242	CAAT -- 0.000166	CGAT -- 0.000048
AACA -- 0.009075	AGCA -- 0.000843	CACA -- 0.005277	CGCA -- 0.000669
AACC -- 0.000703	AGCC -- 0.000315	CACC -- 0.004524	CGCC -- 0.002262
AACG -- 0.000843	AGCG -- 0.002202	CACG -- 0.000669	CGCG -- 0.002304
AACT -- 0.000121	AGCT -- 0.000048	CACT -- 0.000375	CGCT -- 0.000188
AAGA -- 0.028625	AGGA -- 0.005985	CAGA -- 0.003304	CGGA -- 0.001065
AAGC -- 0.000702	AGGC -- 0.000755	CAGC -- 0.000210	CGGC -- 0.000209
AAGG -- 0.005985	AGGG -- 0.032738	CAGG -- 0.001065	CGGG -- 0.006655
AAGT -- 0.000234	AGGT -- 0.000252	CAGT -- 0.000048	CGGT -- 0.000059
AATA -- 0.003025	AGTA -- 0.000281	CATA -- 0.000959	CGTA -- 0.000120
AATC -- 0.000121	AGTC -- 0.000048	CATC -- 0.000360	CGTC -- 0.000180
AATG -- 0.000281	AGTG -- 0.000734	CATG -- 0.000120	CGTG -- 0.000420
AATT -- 0.000154	AGTT -- 0.000073	CATT -- 0.000404	CGTT -- 0.000202
ACAA -- 0.004185	ATAA -- 0.001395	CCAA -- 0.000628	CTAA -- 0.000166
ACAC -- 0.005482	ATAC -- 0.000350	CCAC -- 0.009592	CTAC -- 0.000415
ACAG -- 0.000725	ATAG -- 0.000242	CCAG -- 0.000210	CTAG -- 0.000048
ACAT -- 0.000350	ATAT -- 0.001594	CCAT -- 0.000415	CTAT -- 0.001214
ACCA -- 0.000703	ATCA -- 0.000121	CCCA -- 0.004524	CTCA -- 0.000375
ACCC -- 0.019527	ATCC -- 0.000752	CCCC -- 0.167489	CTCC -- 0.005866
ACCG -- 0.000315	ATCG -- 0.000048	CCCG -- 0.002262	CTCG -- 0.000188
ACCT -- 0.000752	ATCT -- 0.001546	CCCT -- 0.005866	CTCT -- 0.007452
ACGA -- 0.000702	ATGA -- 0.000234	CCGA -- 0.000210	CTGA -- 0.000048
ACGC -- 0.001837	ATGC -- 0.000116	CCGC -- 0.004796	CTGC -- 0.000208
ACGG -- 0.000755	ATGG -- 0.000252	CCGG -- 0.000209	CTGG -- 0.000059
ACGT -- 0.000116	ATGT -- 0.000535	CCGT -- 0.000208	CTGT -- 0.000607
ACTA -- 0.000121	ATTA -- 0.000154	CCTA -- 0.000360	CTTA -- 0.000404
ACTC -- 0.001781	ATTC -- 0.000517	CCTC -- 0.011625	CTTC -- 0.001716
ACTG -- 0.000048	ATTG -- 0.000073	CCTG -- 0.000180	CTTG -- 0.000202
ACTT -- 0.000517	ATTT -- 0.004711	CCTT -- 0.001716	CTTT -- 0.013873

Pattern Probabilities (II)

GAAA -- 0.045565	GGAA -- 0.005060	TAAA -- 0.006106	TGAA -- 0.000497
GAAC -- 0.001004	GGAC -- 0.000453	TAAC -- 0.000166	TGAC -- 0.000048
GAAG -- 0.005060	GGAG -- 0.017648	TAAG -- 0.000497	TGAG -- 0.000959
GAAT -- 0.000335	GGAT -- 0.000151	TAAT -- 0.000099	TGAT -- 0.000038
GACA -- 0.002514	GGCA -- 0.000532	TACA -- 0.000959	TGCA -- 0.000120
GACC -- 0.000315	GGCC -- 0.000194	TACC -- 0.000548	TGCC -- 0.000274
GACG -- 0.000532	GGCG -- 0.002904	TACG -- 0.000120	TGCG -- 0.000420
GACT -- 0.000048	GGCT -- 0.000036	TACT -- 0.000215	TGCT -- 0.000108
GAGA -- 0.014437	GGGA -- 0.008240	TAGA -- 0.001101	TGGA -- 0.000355
GAGC -- 0.000476	GGGC -- 0.001251	TAGC -- 0.000048	TGGC -- 0.000059
GAGG -- 0.008240	GGGG -- 0.056794	TAGG -- 0.000355	TGGG -- 0.002218
GAGT -- 0.000159	GGGT -- 0.000417	TAGT -- 0.000038	TGGT -- 0.000030
GATA -- 0.000838	GGTA -- 0.000177	TATA -- 0.001119	TGTA -- 0.000143
GATC -- 0.000048	GGTC -- 0.000036	TATC -- 0.000231	TGTC -- 0.000116
GATG -- 0.000177	GGTG -- 0.000968	TATG -- 0.000143	TGTG -- 0.000488
GATT -- 0.000073	GGTT -- 0.000040	TATT -- 0.000893	TGTT -- 0.000447
GCAA -- 0.001004	GTAA -- 0.000335	TCAA -- 0.000166	TTAA -- 0.000099
GCAC -- 0.001837	GTAC -- 0.000116	TCAC -- 0.001389	TTAC -- 0.000240
GCAG -- 0.000453	GTAG -- 0.000151	TCAG -- 0.000048	TTAG -- 0.000038
GCAT -- 0.000116	GTAT -- 0.000535	TCAT -- 0.000240	TTAT -- 0.002009
GCCA -- 0.000315	GTCA -- 0.000048	TCCA -- 0.000548	TTCA -- 0.000215
GCCC -- 0.009764	GTCC -- 0.000376	TCCC -- 0.019456	TTCC -- 0.001275
GCCG -- 0.000194	GTCG -- 0.000036	TCCG -- 0.000274	TTCG -- 0.000108
GCCT -- 0.000376	GTCT -- 0.000773	TCCT -- 0.001275	TTCT -- 0.006924
GCGA -- 0.000476	GTGA -- 0.000159	TCGA -- 0.000048	TTGA -- 0.000038
GCGC -- 0.001823	GTGC -- 0.000117	TCGC -- 0.000694	TTGC -- 0.000120
GCGG -- 0.001251	GTGG -- 0.000417	TCGG -- 0.000059	TTGG -- 0.000030
GCGT -- 0.000117	GTGT -- 0.000530	TCGT -- 0.000120	TTGT -- 0.001005
GCTA -- 0.000048	GTTA -- 0.000073	TCTA -- 0.000231	TTTA -- 0.000893
GCTC -- 0.000891	GTTC -- 0.000258	TCTC -- 0.004935	TTTC -- 0.003240
GCTG -- 0.000036	GTTG -- 0.000040	TCTG -- 0.000116	TTTG -- 0.000447
GCTT -- 0.000258	GTTT -- 0.002355	TCTT -- 0.003240	TTTT -- 0.031522

Exotic models of substitution

- Expand model around the sequence
- Allow the substitution process to vary at a single site in the sequence
- Allow the substitution process to vary over a tree at shared sites



RNA secondary structure of the + strand of NL95 (part of the coat and read-through proteins).

	AA	AC	AG	AU	CA	CC	CG	CU	GA	GC	GG	GU	UA	UC	UG	UU
AA																
AC																
AG																
AU																
CA																
CC																
CG																
CU																
GA																
GC																
GG																
GU																
UA																
UC																
UG																
UU																

	AA	AC	AG	AU	CA	CC	CG	CU	GA	GC	GG	GU	UA	UC	UG	UU
AA	–															
AC		–														
AG			–													
AU				–												
CA					–											
CC						–										
CG							–									
CU								–								
GA									–							
GC										–						
GG											–					
GU												–				
UA													–			
UC														–		
UG															–	
UU																–

	AA	AC	AG	AU	CA	CC	CG	CU	GA	GC	GG	GU	UA	UC	UG	UU
AA	–					0	0	0		0	0	0		0	0	0
AC		–			0		0	0	0		0	0	0		0	0
AG			–		0	0		0	0	0		0	0	0		0
AU				–	0	0	0		0	0	0		0	0	0	
CA		0	0	0	–					0	0	0		0	0	0
CC	0		0	0		–			0		0	0	0		0	0
CG	0	0		0			–		0	0		0	0	0		0
CU	0	0	0					–	0	0	0		0	0	0	
GA		0	0	0		0	0	0	–					0	0	0
GC	0		0	0	0		0	0		–			0		0	0
GG	0	0		0	0	0		0			–		0	0		0
GU	0	0	0		0	0	0					–	0	0	0	
UA		0	0	0		0	0	0		0	0	0	–			
UC	0		0	0	0		0	0	0		0	0		–		
UG	0	0		0	0	0		0	0	0		0			–	
UU	0	0	0		0	0	0		0	0	0					–

	AA	AC	AG	AU	CA	CC	CG	CU	GA	GC	GG	GU	UA	UC	UG	UU
AA	–	?	?	?	?	0	0	0	?	0	0	0	?	0	0	0
AC	?	–	?	?	0	?	0	0	0	?	0	0	0	?	0	0
AG	?	?	–	?	0	0	?	0	0	0	?	0	0	0	?	0
AU	?	?	?	–	0	0	0	?	0	0	0	?	0	0	0	?
CA	?	0	0	0	–	?	?	?	?	0	0	0	?	0	0	0
CC	0	?	0	0	?	–	?	?	0	?	0	0	0	?	0	0
CG	0	0	?	0	?	?	–	?	0	0	?	0	0	0	?	0
CU	0	0	0	?	?	?	?	–	0	0	0	?	0	0	0	?
GA	?	0	0	0	?	0	0	0	–	?	?	?	?	0	0	0
GC	0	?	0	0	0	?	0	0	?	–	?	?	0	?	0	0
GG	0	0	?	0	0	0	?	0	?	?	–	?	0	0	?	0
GU	0	0	0	?	0	0	0	?	?	?	?	–	0	0	0	?
UA	?	0	0	0	?	0	0	0	?	0	0	0	–	?	?	?
UC	0	?	0	0	0	?	0	0	0	?	0	0	?	–	?	?
UG	0	0	?	0	0	0	?	0	0	0	?	0	?	?	–	?
UU	0	0	0	?	0	0	0	?	0	0	0	?	?	?	?	–

Doublet Model
(Schöniger and von Haeseler, 1994)

$$q_{ij} = \begin{cases} \kappa\pi_j & : \text{transition} \\ \pi_j & : \text{transversion} \\ 0 & : i \text{ and } j \text{ differ at two positions} \end{cases}$$

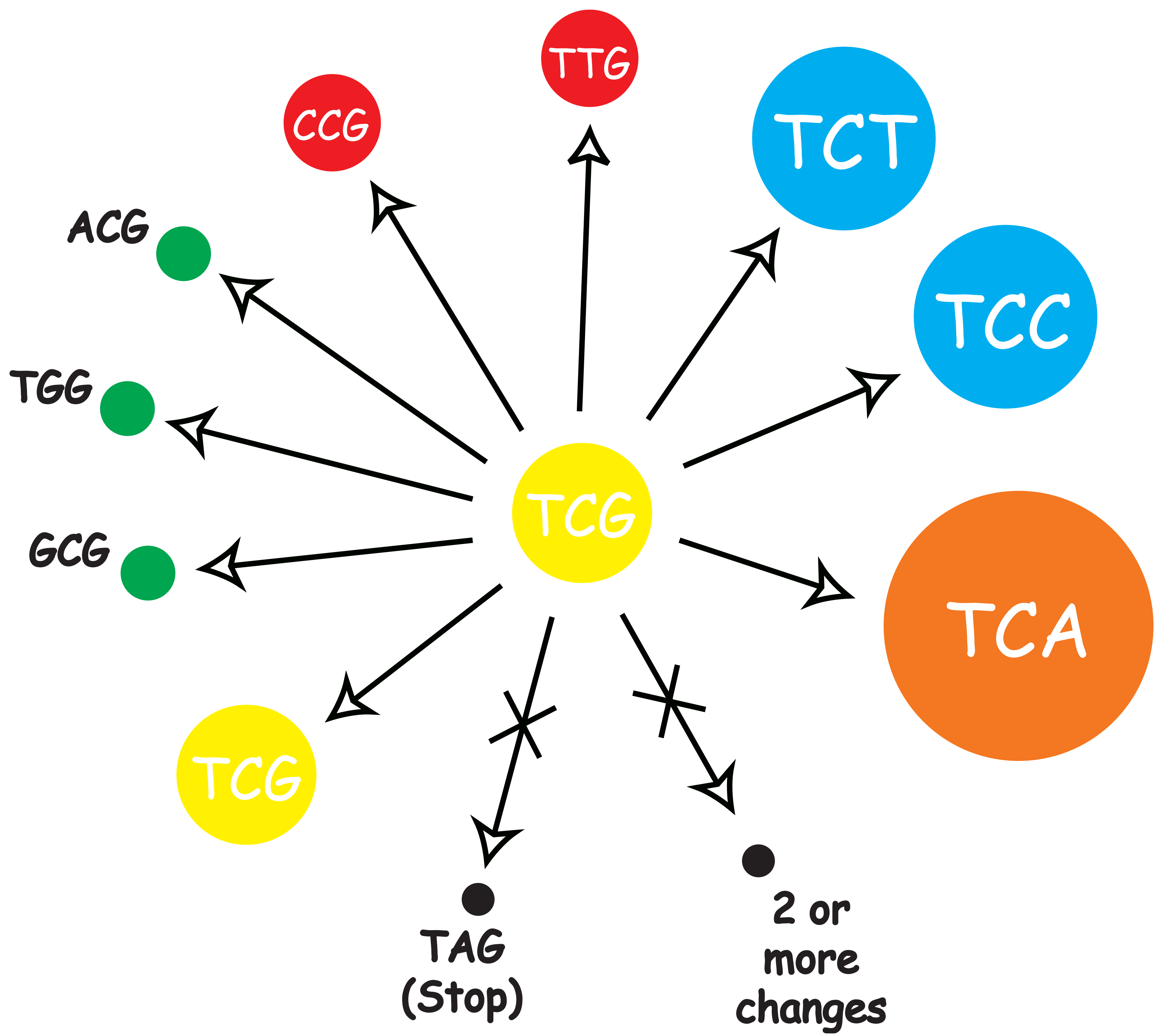
	AAA	AAC	AAG	AAT	• • • • •	TTA	TTC	TTG	TTT
AAA	–	?	?	?		0	0	0	0
AAC	?	–	?	?		0	0	0	0
AAG	?	?	–	?		0	0	0	0
AAT	?	?	?	–		0	0	0	0
⋮									
TTA	0	0	0	0		–	?	?	?
TTC	0	0	0	0		?	–	?	?
TTG	0	0	0	0		?	?	–	?
TTT	0	0	0	0		?	?	?	–

53 states not shown

	AAA	AAC	AAG	AAT	• • • • •	TTA	TTC	TTG	TTT
AAA	–	?	?	?		0	0	0	0
AAC	?	–	?	?		0	0	0	0
AAG	?	?	–	?		0	0	0	0
AAT	?	?	?	–		0	0	0	0
• • • • •									
TTA	0	0	0	0		–	?	?	?
TTC	0	0	0	0		?	–	?	?
TTG	0	0	0	0		?	?	–	?
TTT	0	0	0	0		?	?	?	–

Codon Model
(Goldman & Yang, 1994; Muse and Gaut, 1994; Nielsen & Yang, 1998)

$$q_{ij} = \begin{cases} \omega\kappa\pi_j & : \text{nonsynonymous transition} \\ \omega\pi_j & : \text{nonsynonymous transversion} \\ \kappa\pi_j & : \text{synonymous transition} \\ \pi_j & : \text{synonymous transversion} \\ 0 & : i \text{ and } j \text{ differ at 2 or 3 positions} \end{cases}$$


TTG
CCG
TCT
TCC
ACG
TGG
GCG
TCG
TCA
TCG
TAG
(Stop)
2 or
more
changes

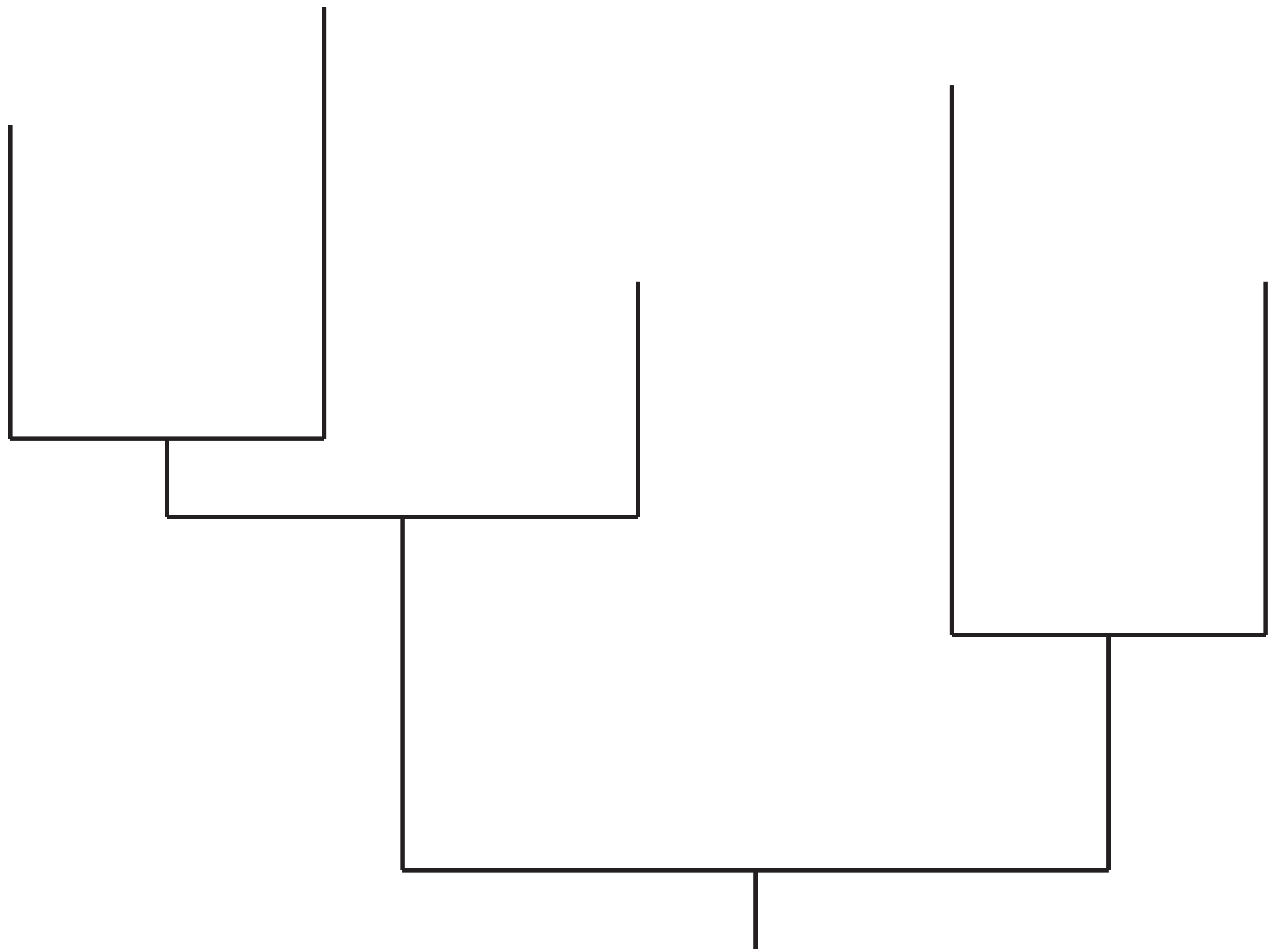
	AAAAAA	AAAAAC	• • • • •	TTTTTG	TTTTTT
AAAAAA	–	?		0	0
AAAAAC	?	–		0	0
• • • • •					
TTTTTG	0	0		–	?
TTTTTT	0	0		?	–

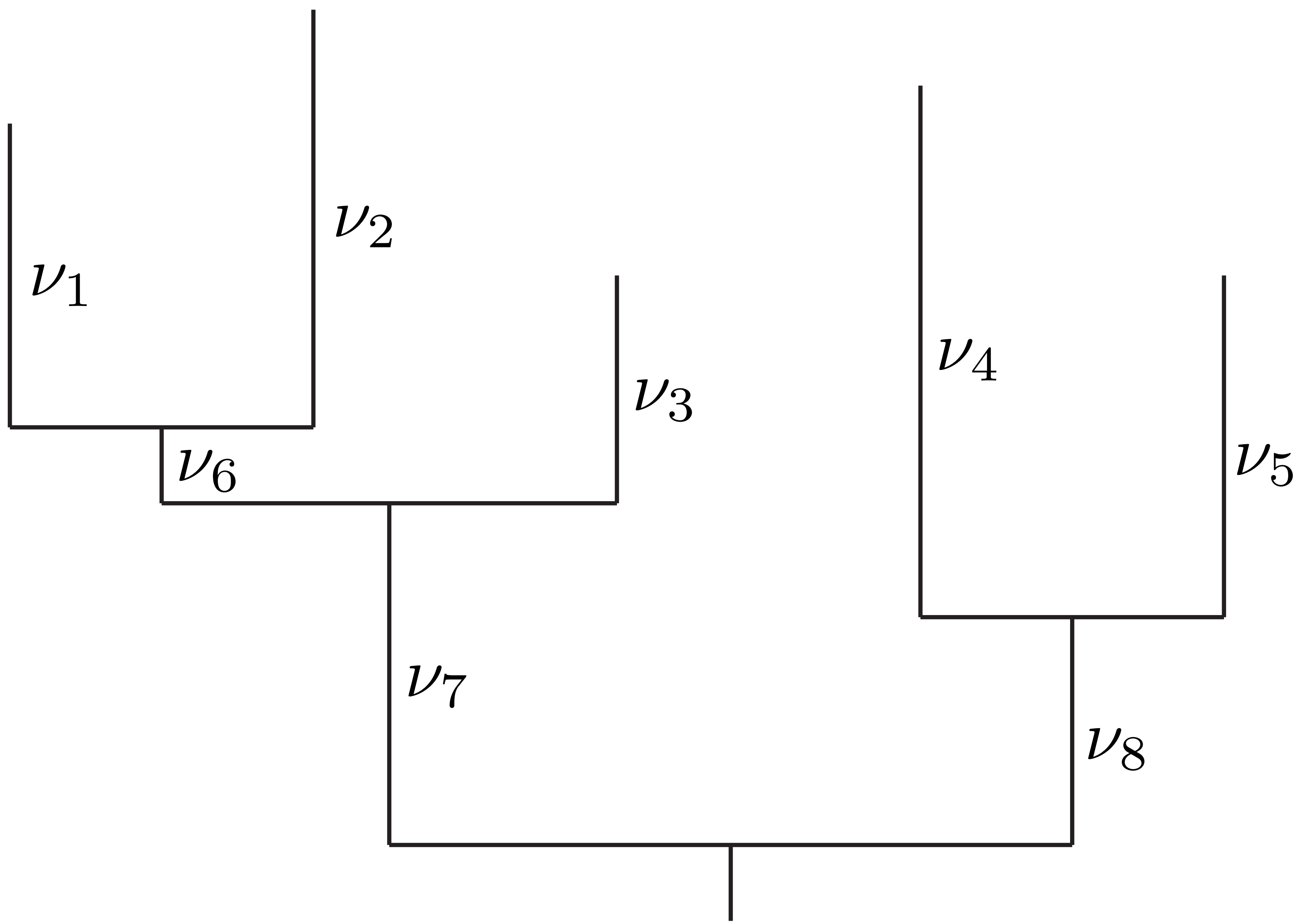
4092 states not shown

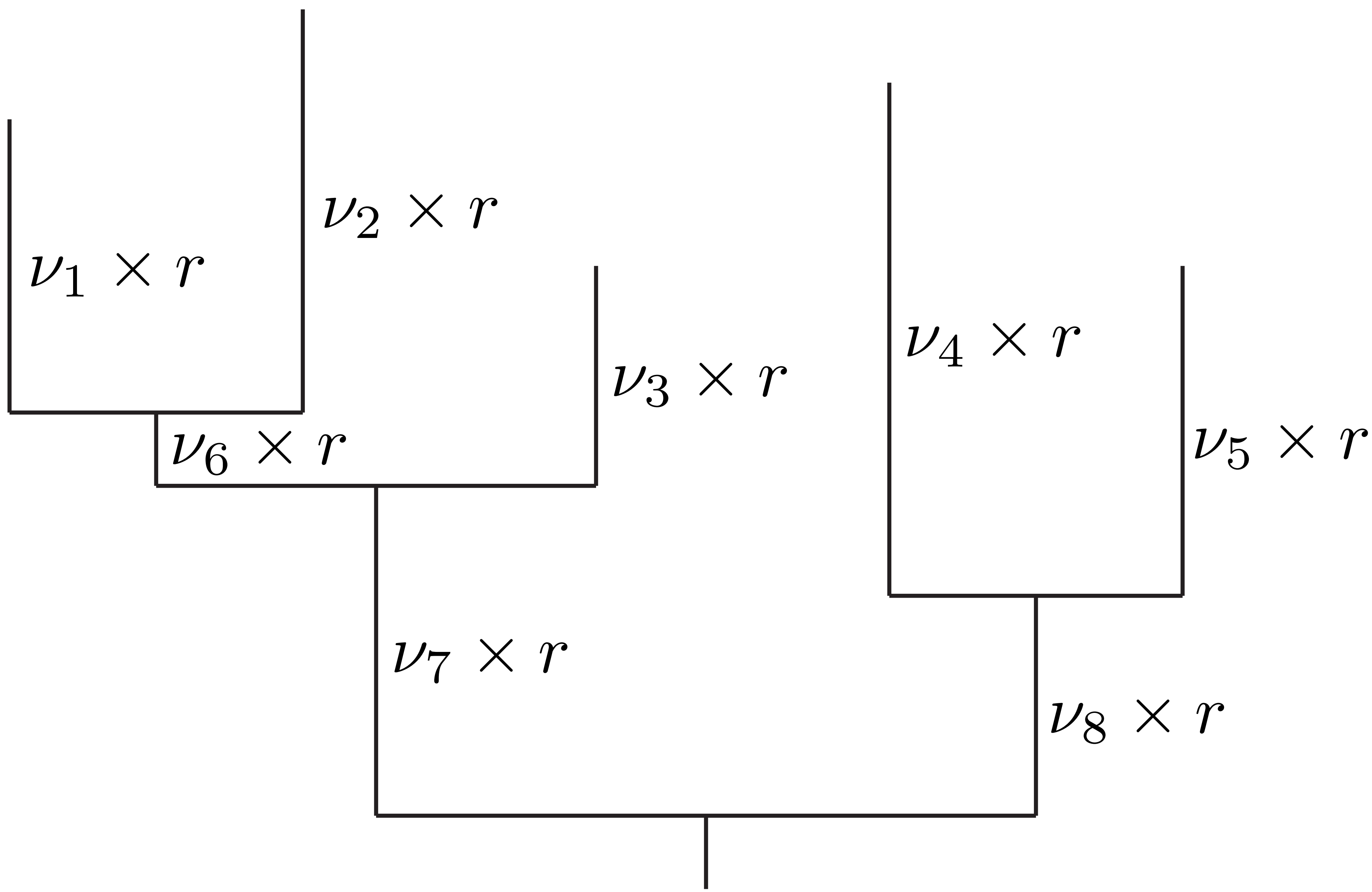
	AAAAAA	AAAAAC	• • • • •	TTTTTG	TTTTTT
AAAAAA	–	?		0	0
AAAAAC	?	–		0	0
⋮					
TTTTTG	0	0		–	?
TTTTTT	0	0		?	–

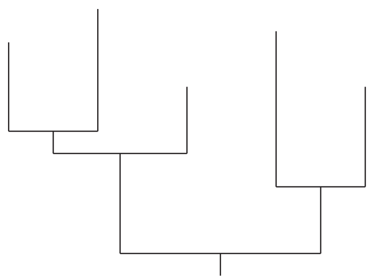
‘Sequence’ Model

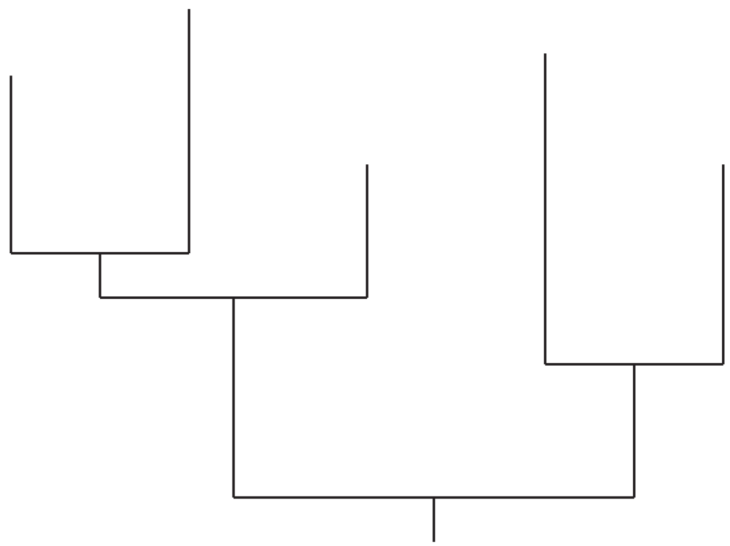
(Robinson et al., 2003)

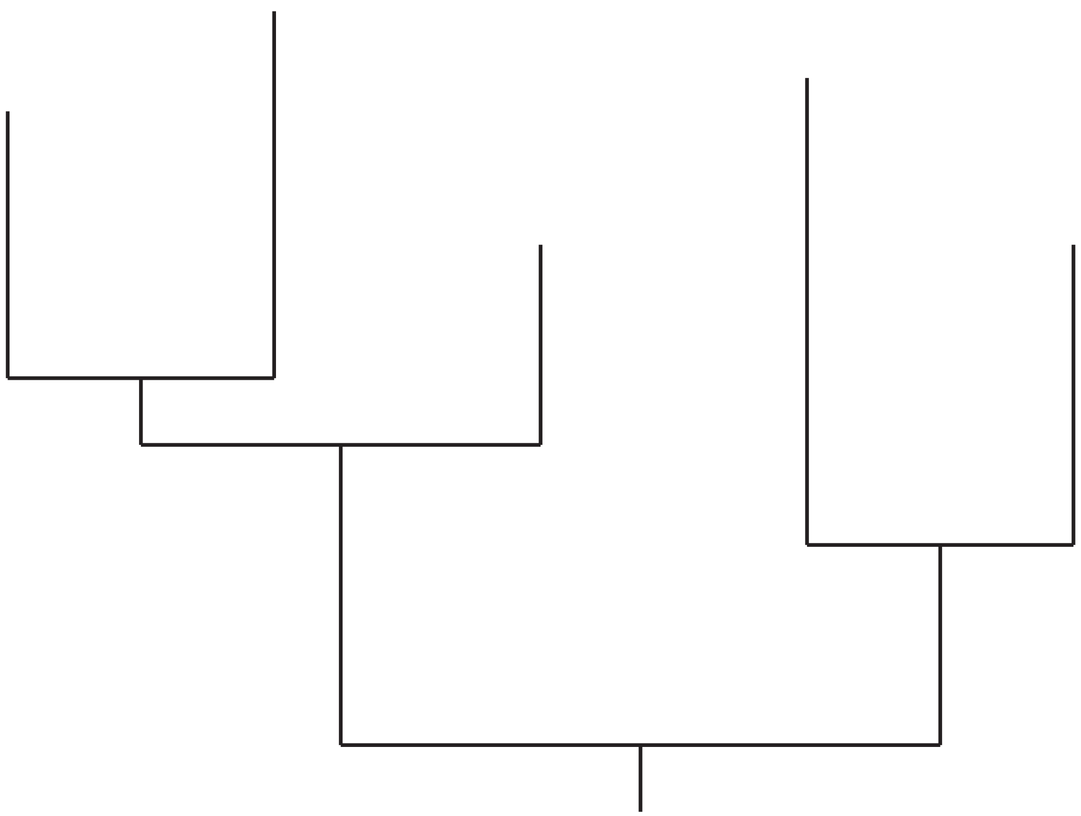


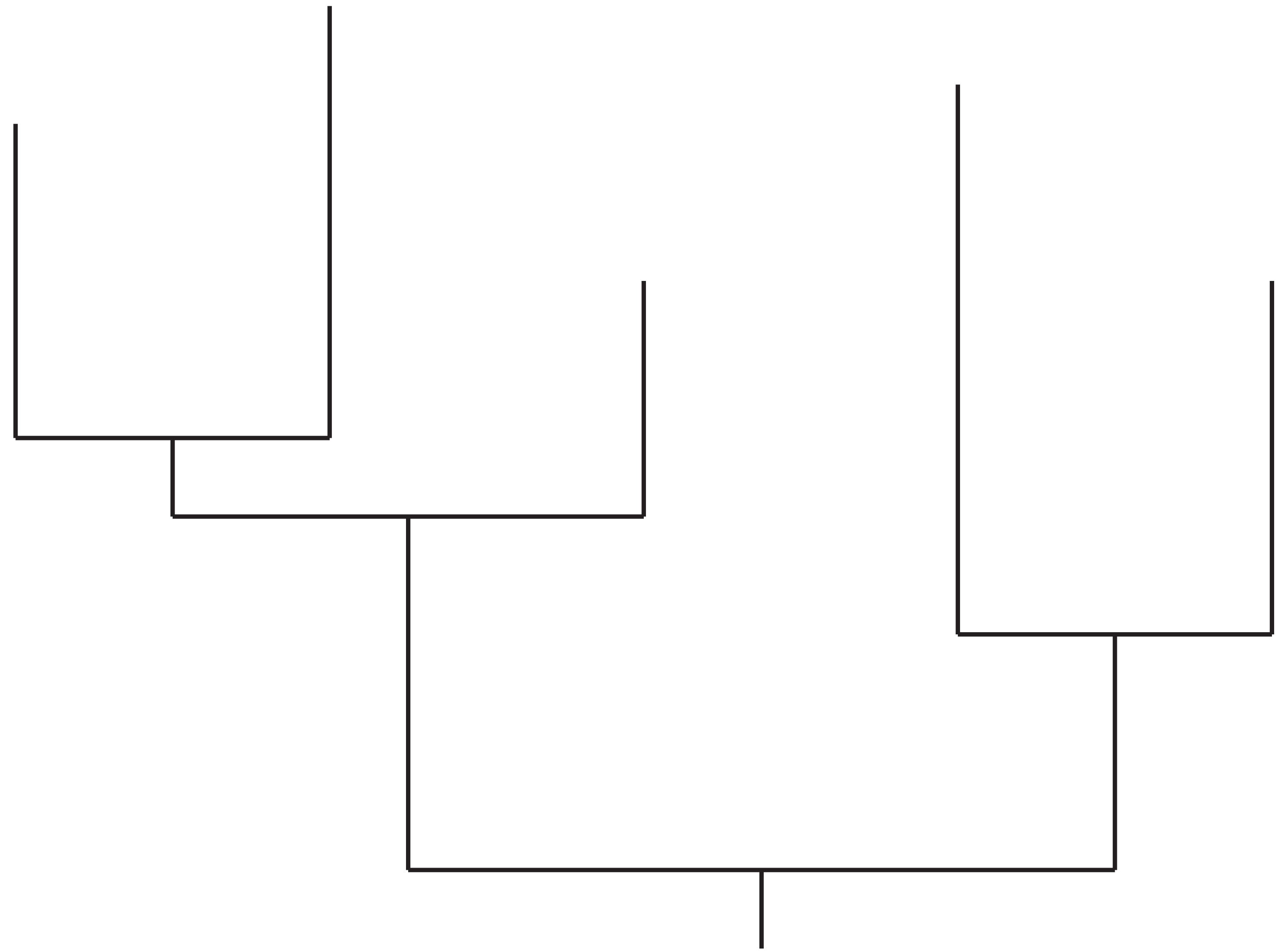

ν_1
ν_2
ν_3
ν_4
ν_5
ν_6
ν_7
ν_8

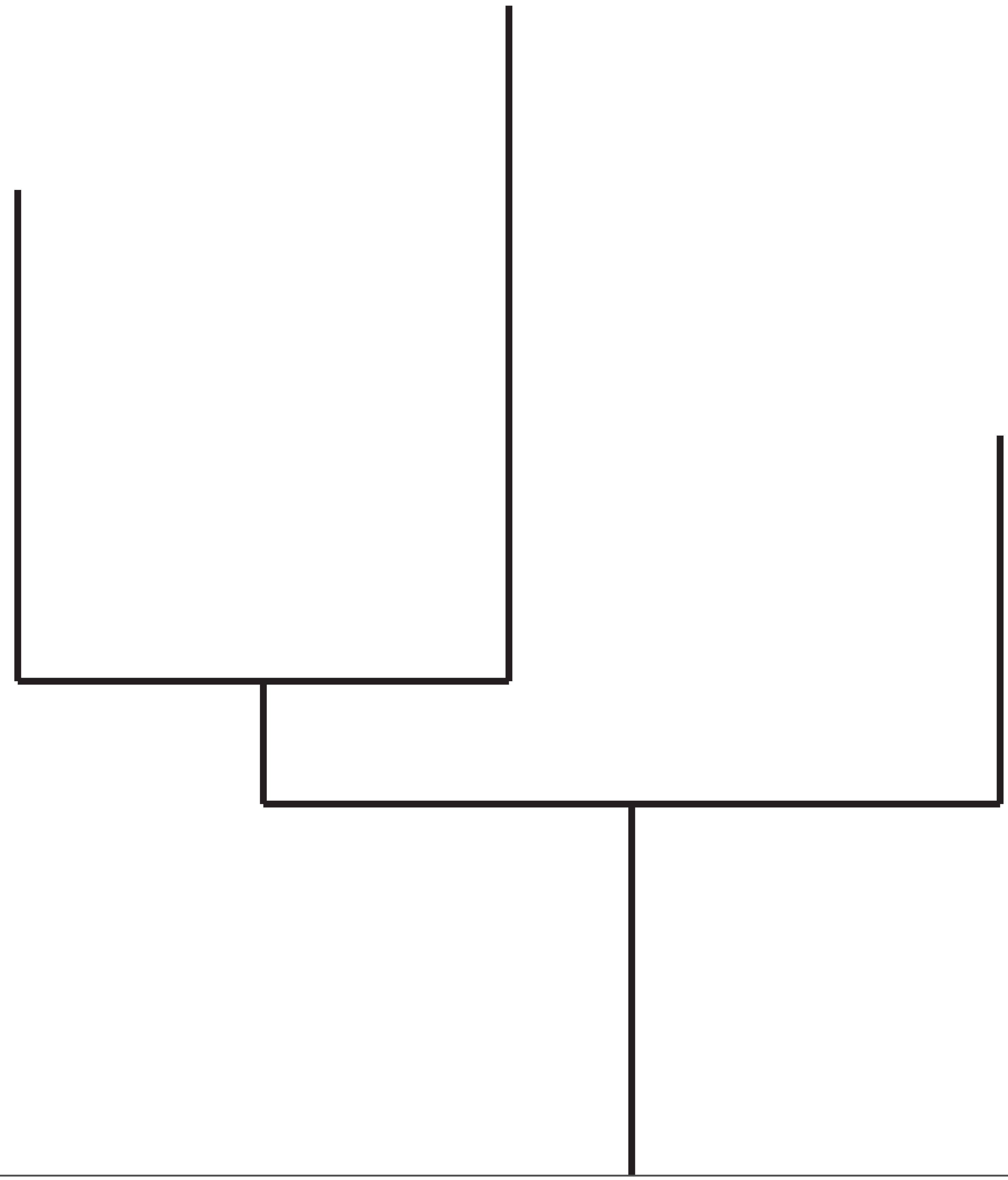

$\nu_2 \times r$
$\nu_1 \times r$
$\nu_4 \times r$
$\nu_3 \times r$
$\nu_6 \times r$
$\nu_5 \times r$
$\nu_7 \times r$
$\nu_8 \times r$

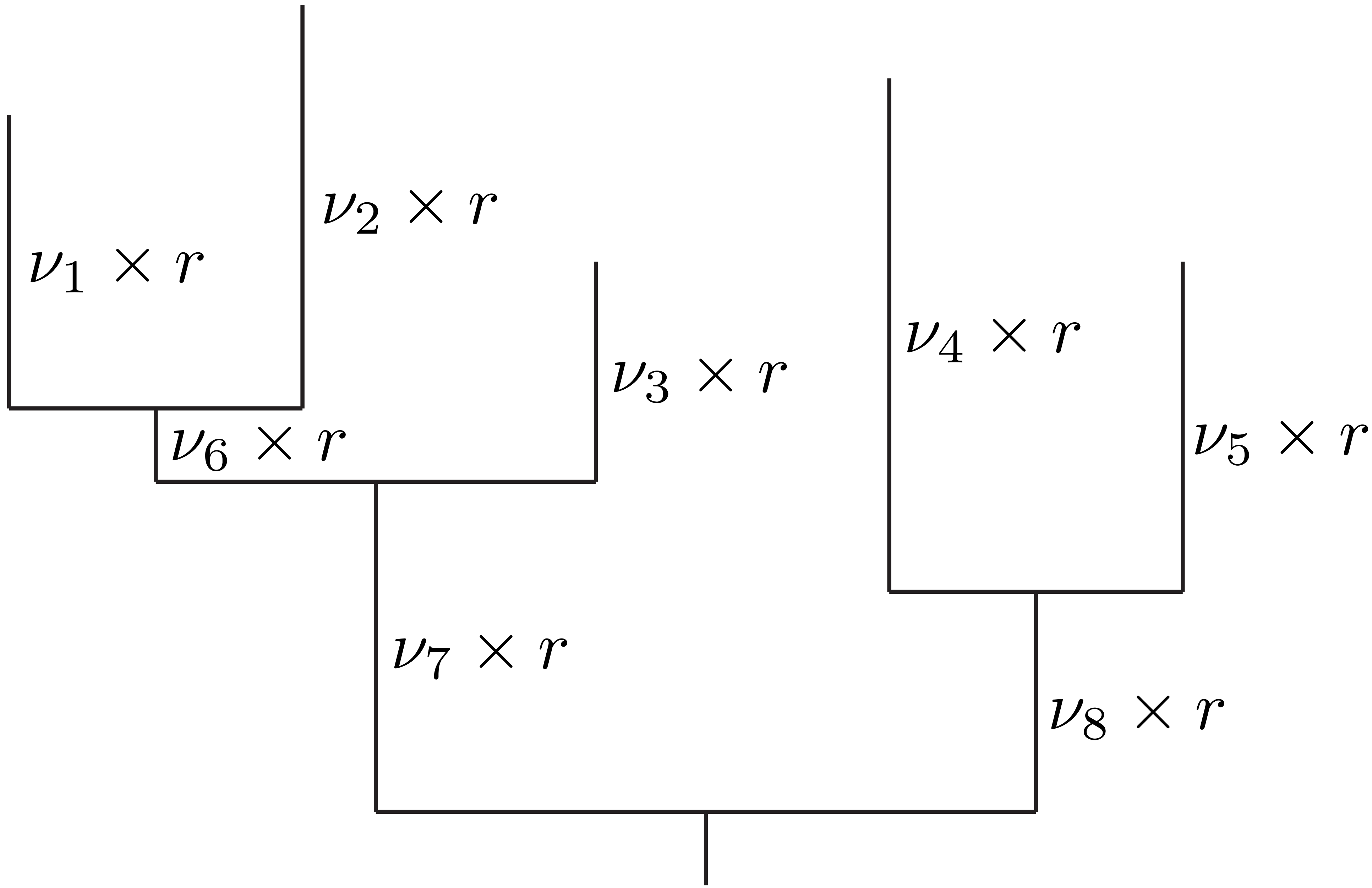








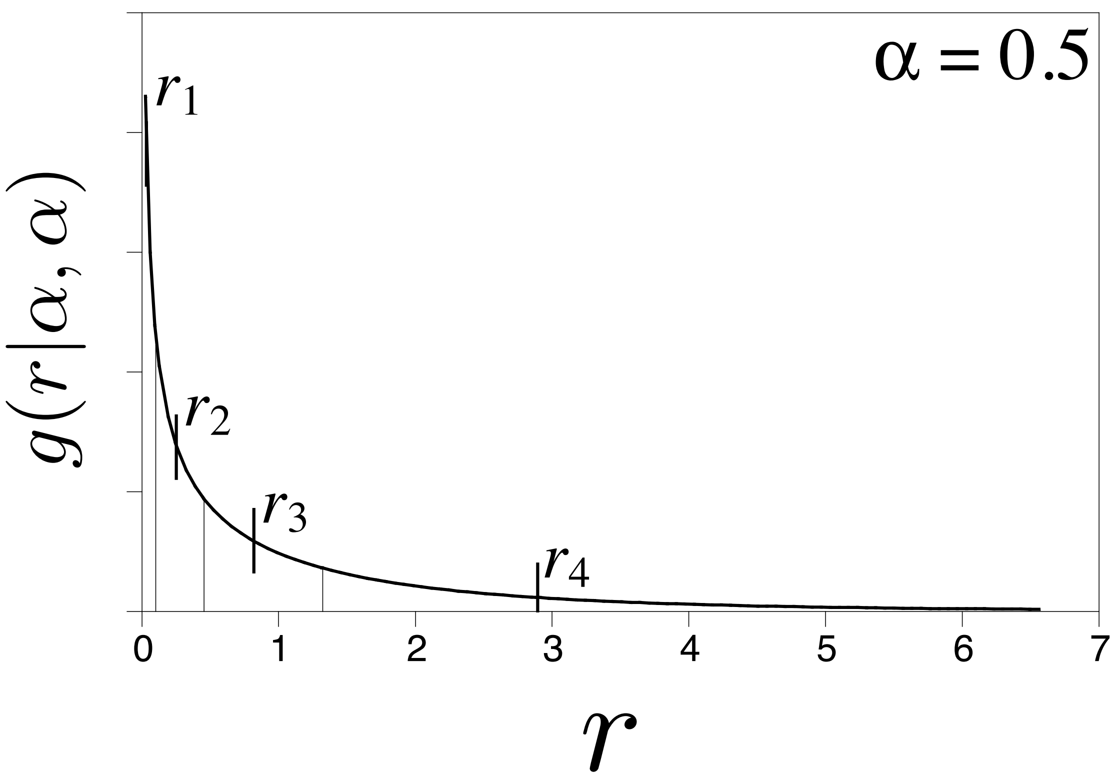


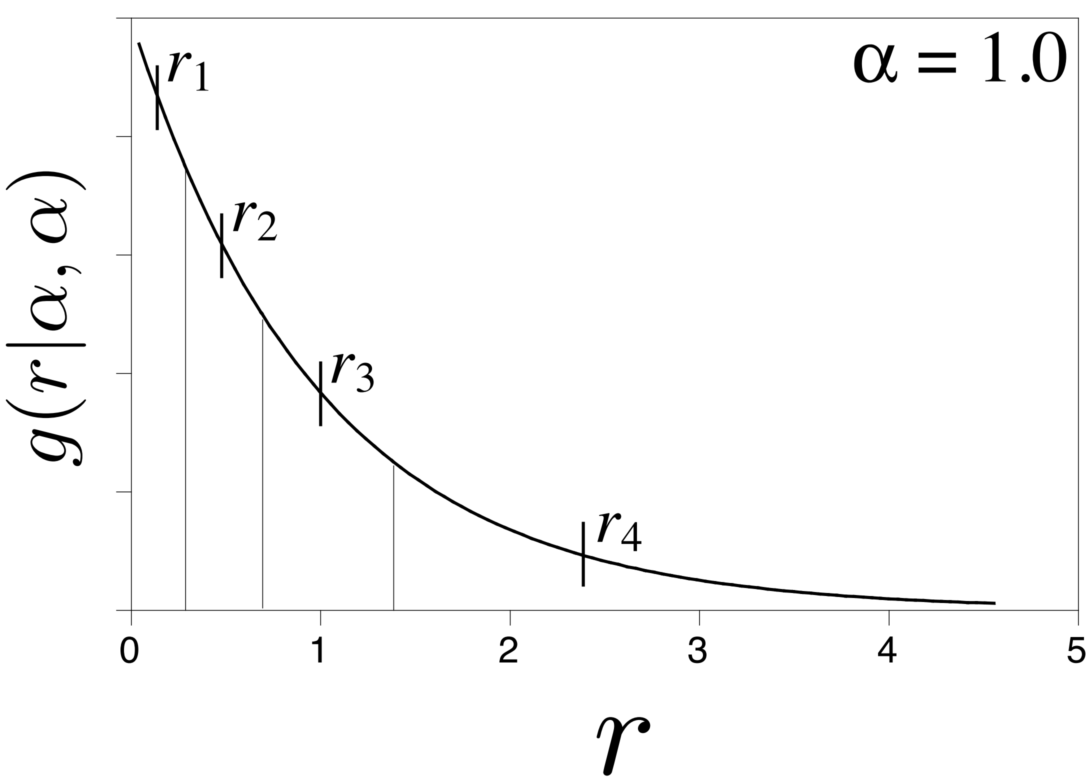


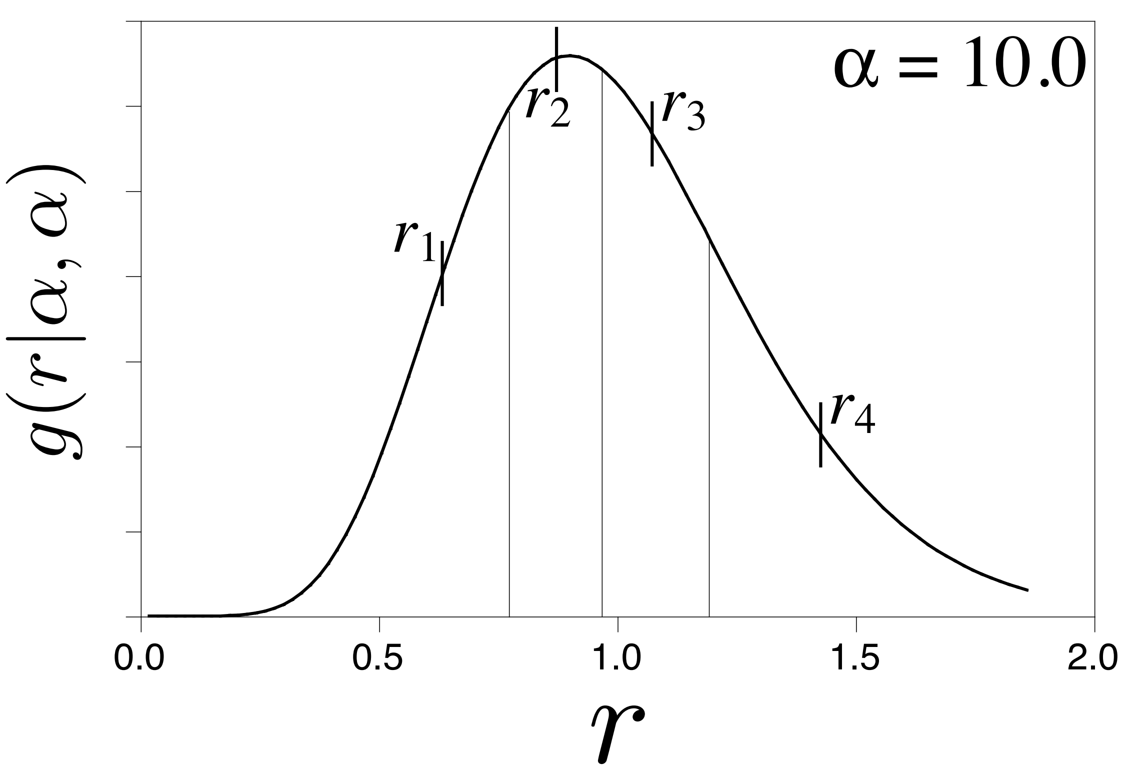
$$r \sim \text{Gamma}(\alpha, \alpha)$$

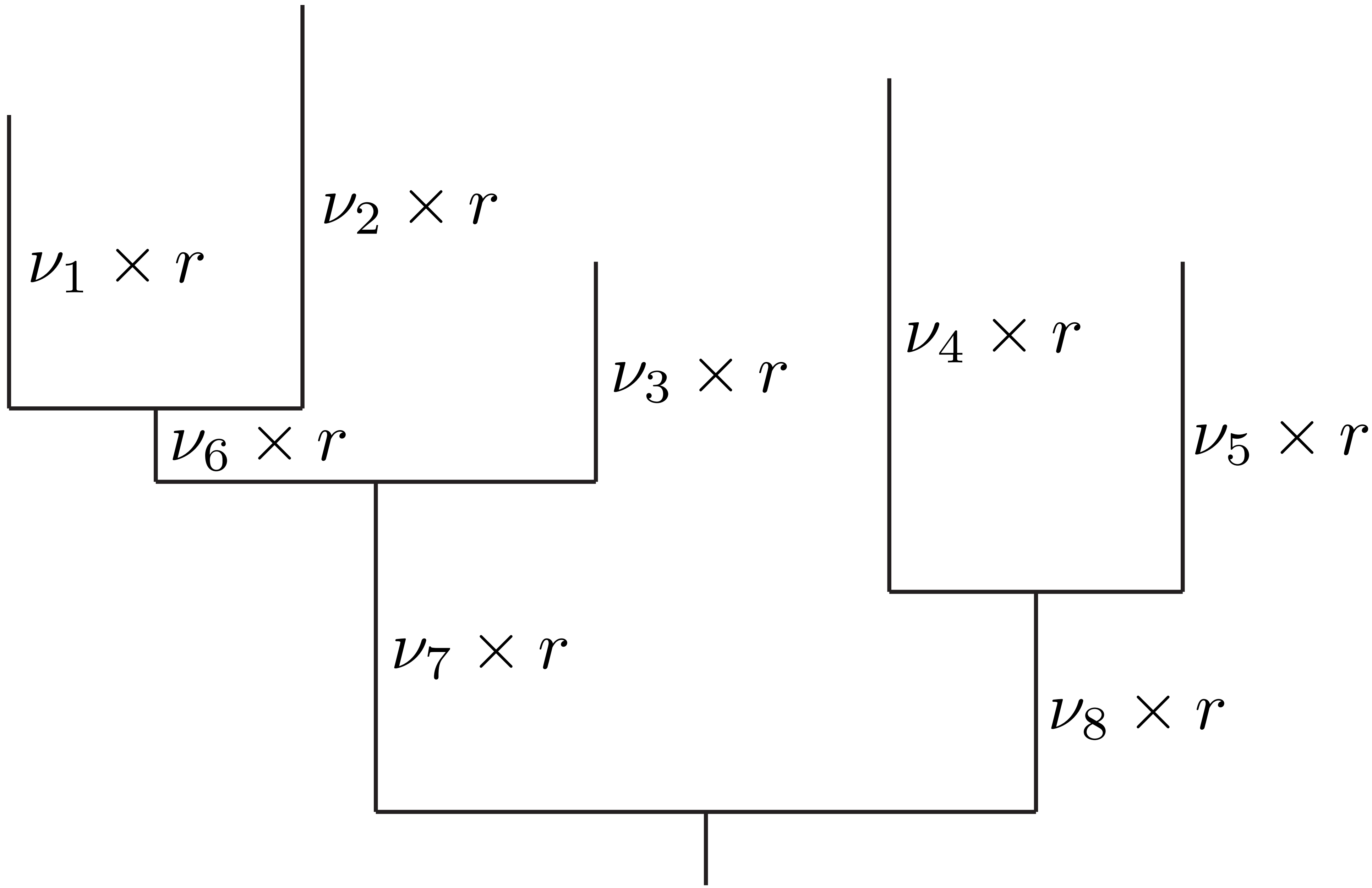
$$\Pr(\text{site}|\alpha, \text{other stuff}) = \int_0^\infty \Pr(\text{site}|r, \text{other stuff}) g(r|\alpha, \alpha) dr$$

Yang, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.


α = 0.5
r_1
r_2
r_3
r_4
$g(r|\alpha, \alpha)$
0
1
2
3
4
5
6
7
r

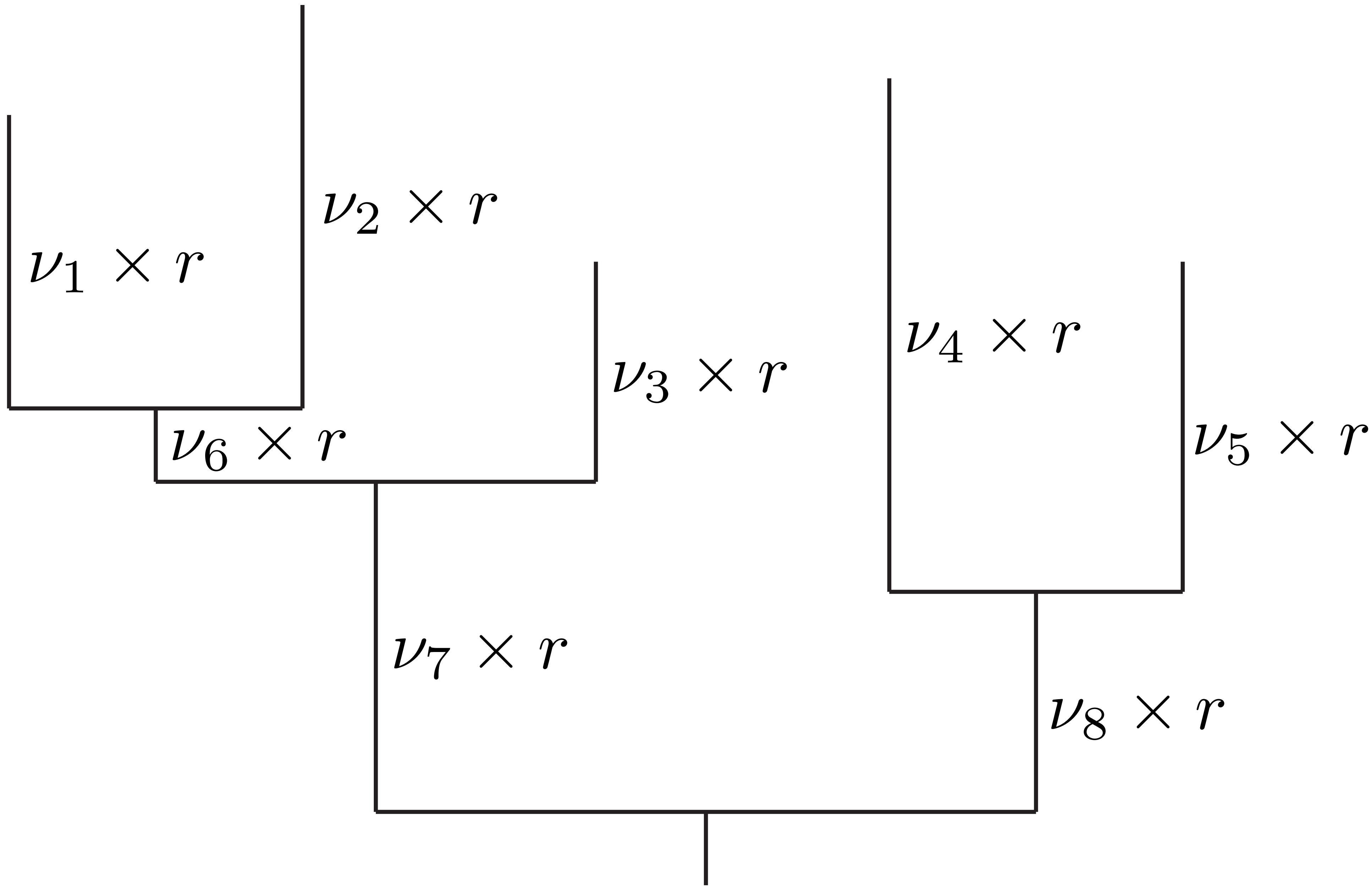

α = 1.0
r_1
r_2
r_3
r_4
$g(r|\alpha, \alpha)$
0
1
2
3
4
5
r


α = 10.0
r_1
r_2
r_3
r_4
$g(r|\alpha, \alpha)$
0.0
0.5
1.0
1.5
2.0
r



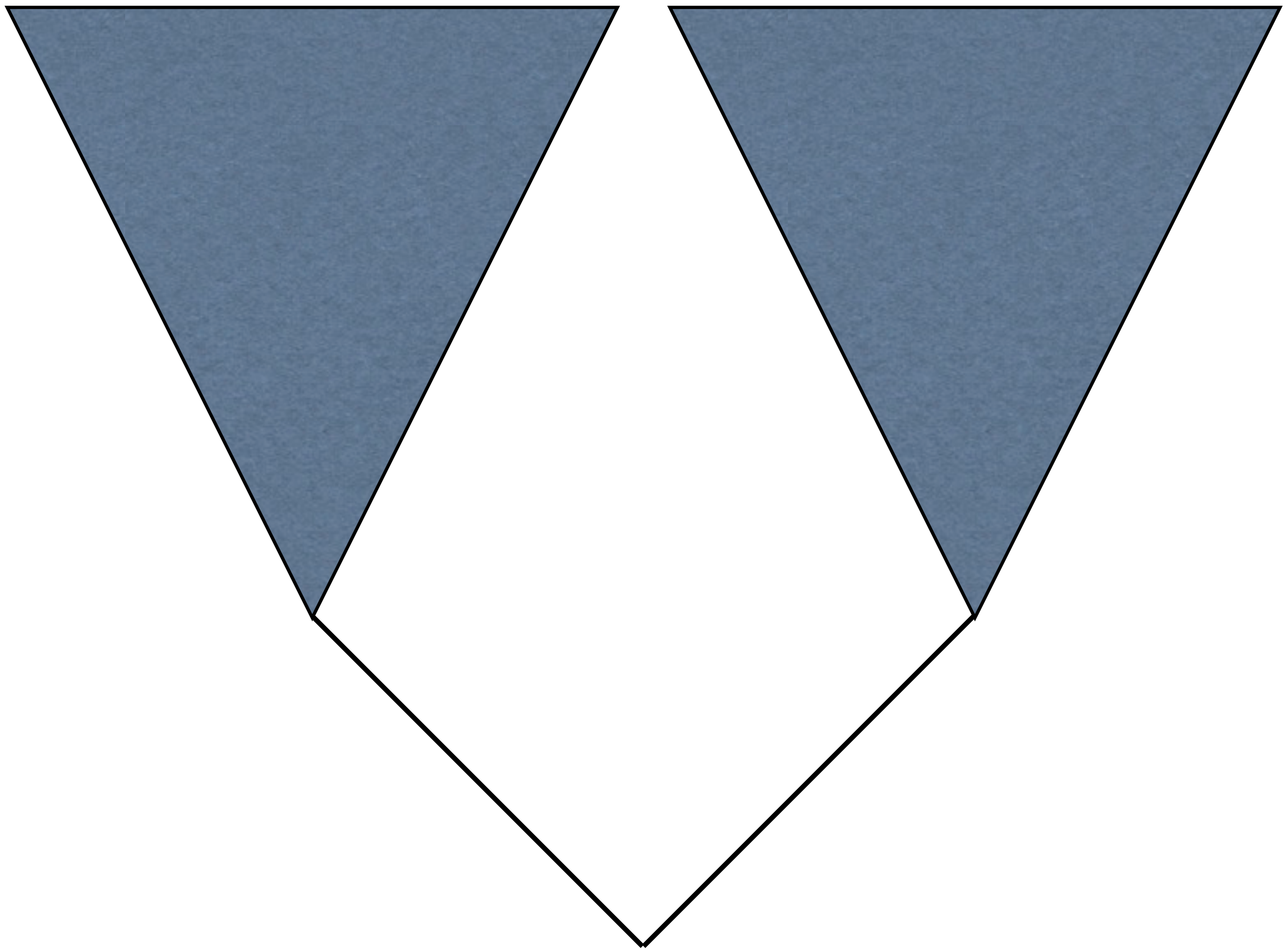
$$\Pr(\text{site}|\alpha, \text{other stuff}) = \sum_{k=1}^{K} \Pr(\text{site}|r_k, \text{other stuff}) \frac{1}{K}$$

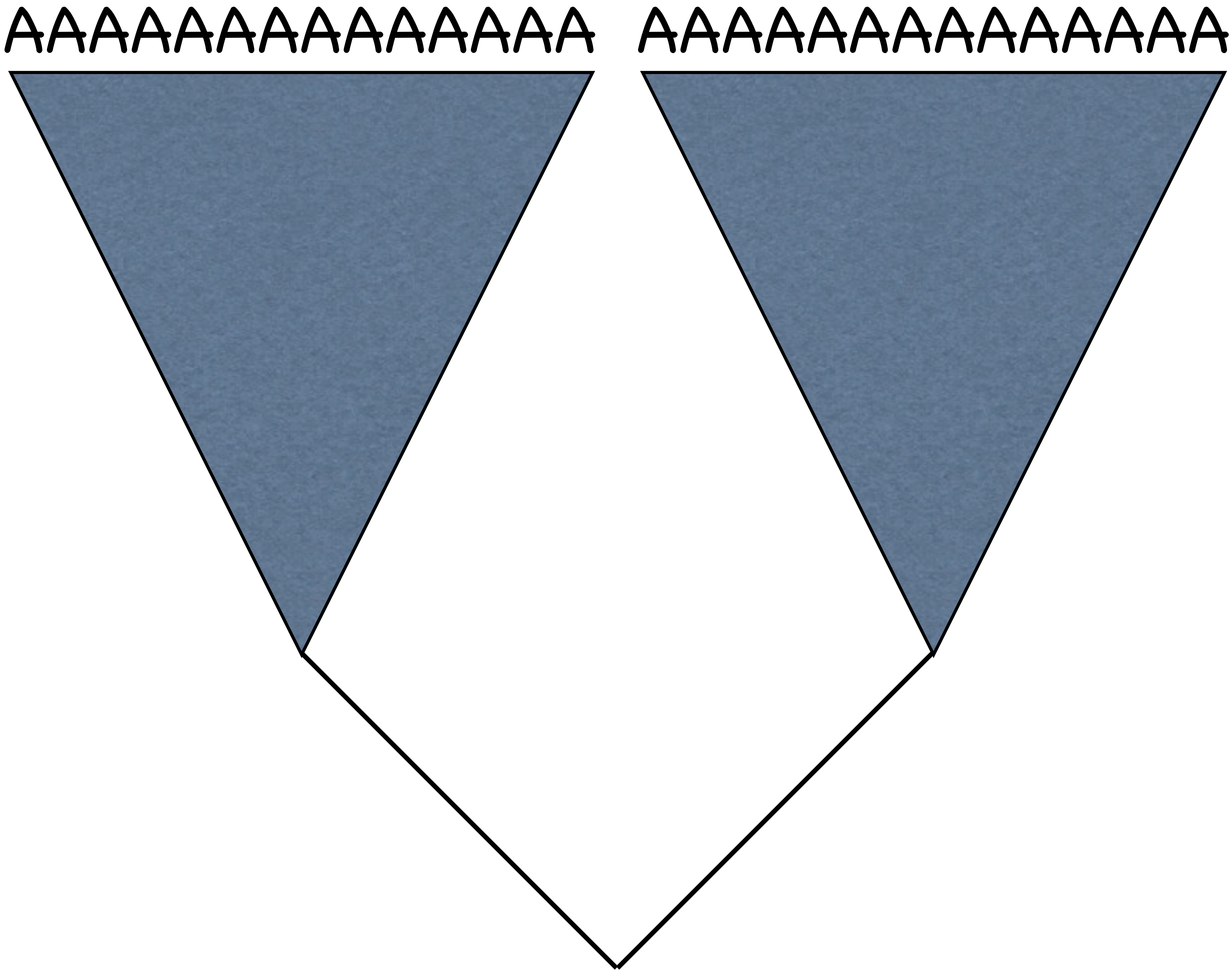
Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39:306–314.

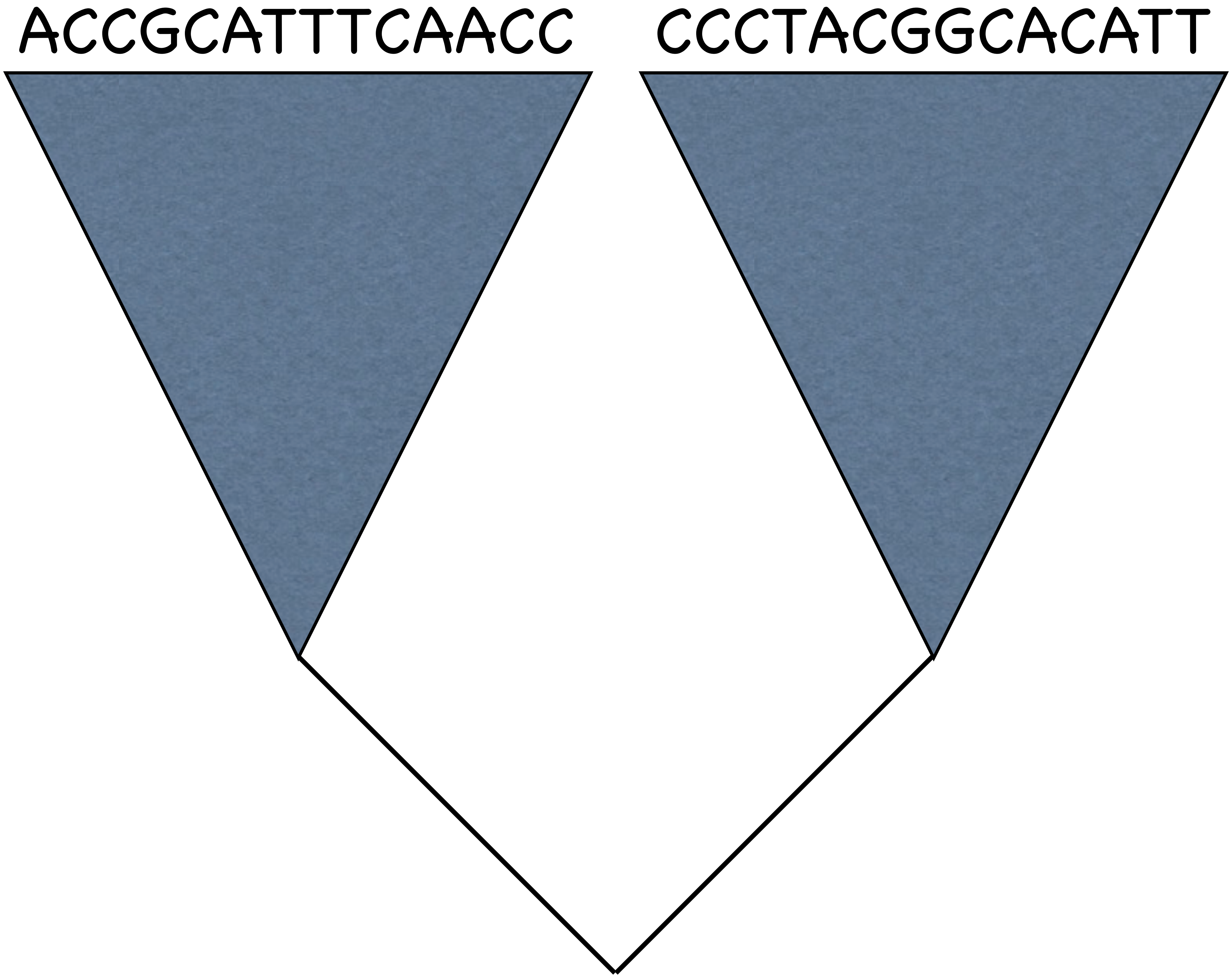


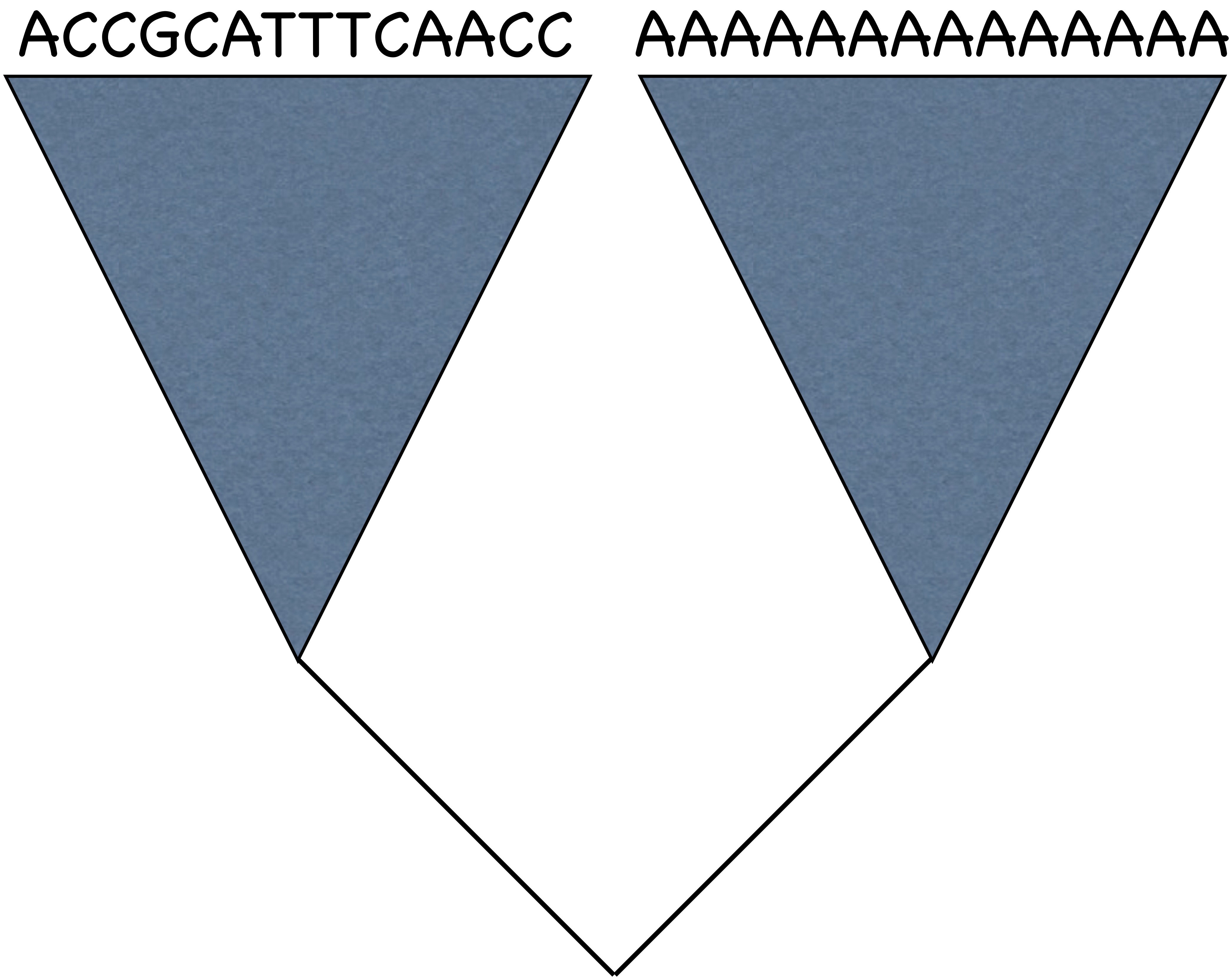
$$r \sim \begin{cases} 0 & : \quad \text{with probability } p \\ 1/(1-p) & : \quad \text{with probability } 1-p \end{cases}$$

$$\begin{aligned} \Pr(\text{site}|p, \text{other stuff}) &= \Pr(\text{site}|r = 0, \text{other stuff}) \times p \\ &+ \Pr(\text{site}|r = 1/(1-p), \text{other stuff}) \times (1-p) \end{aligned}$$




AAAAAAAAAAAAAA AAAAAAAAAAAAAA


ACCGCATTTCAACC
CCCTACGGCACATT


ACCGCATTTCAACC
AAAAAAAAAAAAAA

	A_0	C_0	G_0	T_0	A_1	C_1	G_1	T_1
A_0								
C_0								
G_0								
T_0								
A_1								
C_1								
G_1								
T_1								

	A_0	C_0	G_0	T_0	A_1	C_1	G_1	T_1
A_0	–							
C_0		–						
G_0			–					
T_0				–				
A_1					–			
C_1						–		
G_1							–	
T_1								–

	A_0	C_0	G_0	T_0	A_1	C_1	G_1	T_1
A_0	–					0	0	0
C_0		–			0		0	0
G_0			–		0	0		0
T_0				–	0	0	0	
A_1		0	0	0	–			
C_1	0		0	0		–		
G_1	0	0		0			–	
T_1	0	0	0					–

	A_0	C_0	G_0	T_0	A_1	C_1	G_1	T_1
A_0	−	0	0	0		0	0	0
C_0	0	−	0	0	0		0	0
G_0	0	0	−	0	0	0		0
T_0	0	0	0	−	0	0	0	
A_1		0	0	0	−			
C_1	0		0	0		−		
G_1	0	0		0			−	
T_1	0	0	0					−

	A_0	C_0	G_0	T_0	A_1	C_1	G_1	T_1
A_0	–	0	0	0	λq	0	0	0
C_0	0	–	0	0	0	λq	0	0
G_0	0	0	–	0	0	0	λq	0
T_0	0	0	0	–	0	0	0	λq
A_1	λp	0	0	0	–			
C_1	0	λp	0	0		–		
G_1	0	0	λp	0			–	
T_1	0	0	0	λp				–

$q = 1-p$

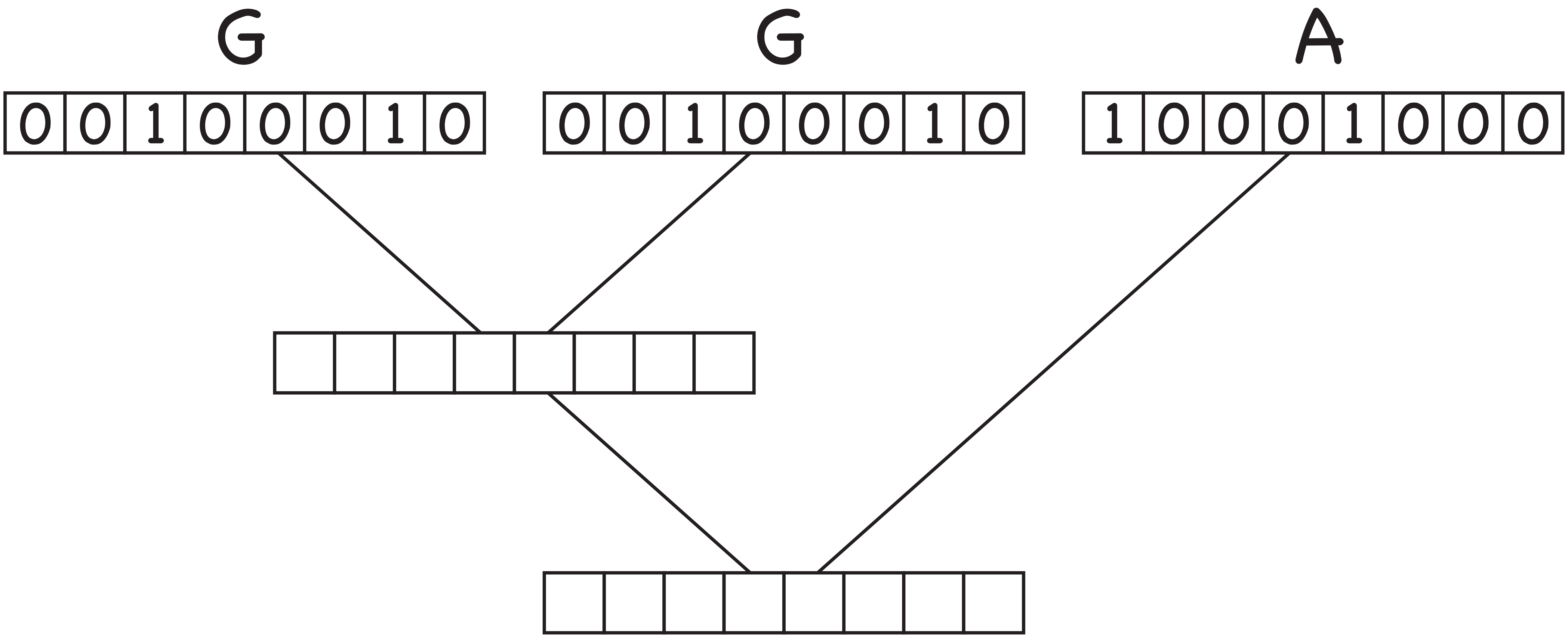
	A_0	C_0	G_0	T_0	A_1	C_1	G_1	T_1
A_0	–	0	0	0	λq	0	0	0
C_0	0	–	0	0	0	λq	0	0
G_0	0	0	–	0	0	0	λq	0
T_0	0	0	0	–	0	0	0	λq
A_1	λp	0	0	0	–			
C_1	0	λp	0	0		–		
G_1	0	0	λp	0			–	
T_1	0	0	0	λp				–

	A_0	C_0	G_0	T_0	A_1	C_1	G_1	T_1
A_0	–	0	0	0	λq	0	0	0
C_0	0	–	0	0	0	λq	0	0
G_0	0	0	–	0	0	0	λq	0
T_0	0	0	0	–	0	0	0	λq
A_1	λp	0	0	0	–	?	?	?
C_1	0	λp	0	0	?	–	?	?
G_1	0	0	λp	0	?	?	–	?
T_1	0	0	0	λp	?	?	?	–

Covariotide-like model of Tuffley & Steel (1997)

$$
Q = \begin{pmatrix}
- & 0 & 0 & 0 & \lambda_{01} & 0 & 0 & 0 \\
0 & - & 0 & 0 & 0 & \lambda_{01} & 0 & 0 \\
0 & 0 & - & 0 & 0 & 0 & \lambda_{01} & 0 \\
0 & 0 & 0 & - & 0 & 0 & 0 & \lambda_{01} \\
\lambda_{10} & 0 & 0 & 0 & - & r_{AC}\pi_C & r_{AG}\pi_G & r_{AT}\pi_T \\
0 & \lambda_{10} & 0 & 0 & r_{AC}\pi_A & - & r_{CG}\pi_G & r_{CT}\pi_T \\
0 & 0 & \lambda_{10} & 0 & r_{AG}\pi_A & r_{CG}\pi_C & - & \pi_T \\
0 & 0 & 0 & \lambda_{10} & r_{AT}\pi_A & r_{CT}\pi_C & \pi_G & -
\end{pmatrix}
$$

$$
Q = \left(\begin{array}{c|c}
\begin{array}{c}\text{Process is}\\ \text{off (no substitutions}\\ \text{are possible)}\end{array} & \begin{array}{c}\text{Switching from}\\ \text{off to on}\end{array} \\
\hline
\begin{array}{c}\text{Switching from}\\ \text{on to off}\end{array} & \begin{array}{c}\text{Process is}\\ \text{on (substitutions}\\ \text{may occur)}\end{array}
\end{array}\right)
$$


G
G
A
0 0 1 0 0 0 1 0
0 0 1 0 0 0 1 0
1 0 0 0 1 0 0 0

MCMC

Why I like likelihood

- Good for phylogeny estimation (good models lead to good trees?)
- Allows us to learn about the pattern and, to some extent, the process of molecular evolution (model comparison)
- Coherent methodology that uses all of the information in the data

Why I like Bayes

- Allows us to examine quite complicated models (e.g., sequence models)
- Easy interpretation of results
- Allows us to marginalize over things we should be marginalizing (e.g., trees, substitution parameters, partitions, alignments)
- I like to think that scientists operate in a Bayesian manner

Caveats

- How complicated can our models become before they are unidentifiable?
- MCMC allows us to do things that are impossible to do any other way. That said, the method is complicated and not guaranteed to work for any particular problem.
- How sensitive are results to the prior?