Kingman's coalescent

Random collision of lineages as go back in time (sans recombination)

Collision is faster the smaller the effective population size

Average time for

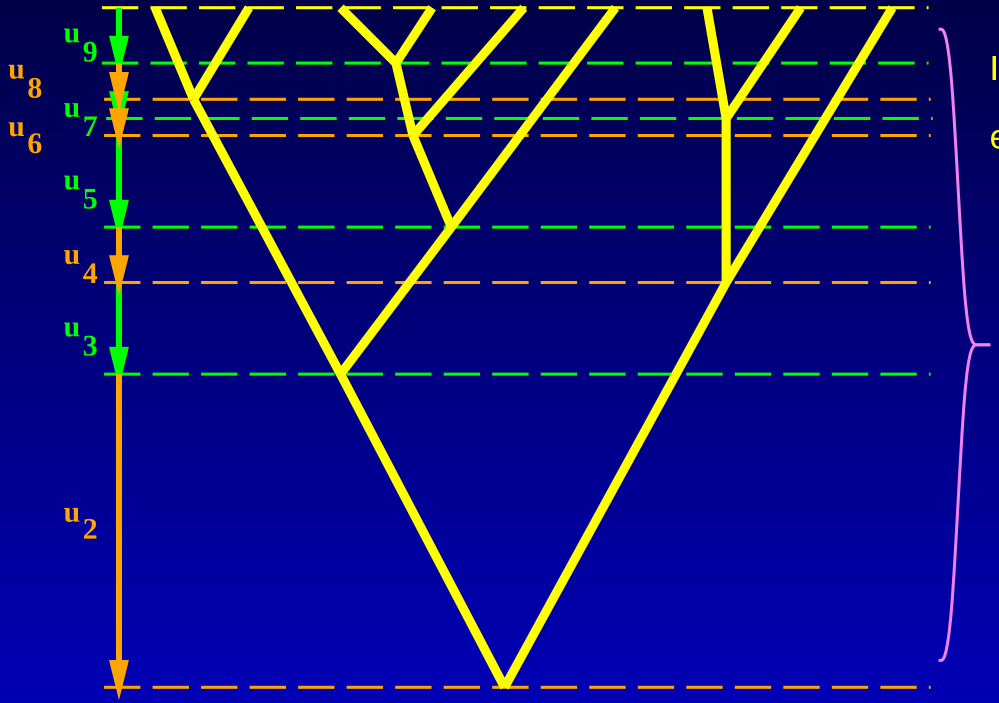
k copies to coalesce to

$$k-1 \quad = \quad \frac{4N}{k(k-1)}$$

Average time for

two copies to coalesce

= 2N generations



In a diploid population of

effective population size N,

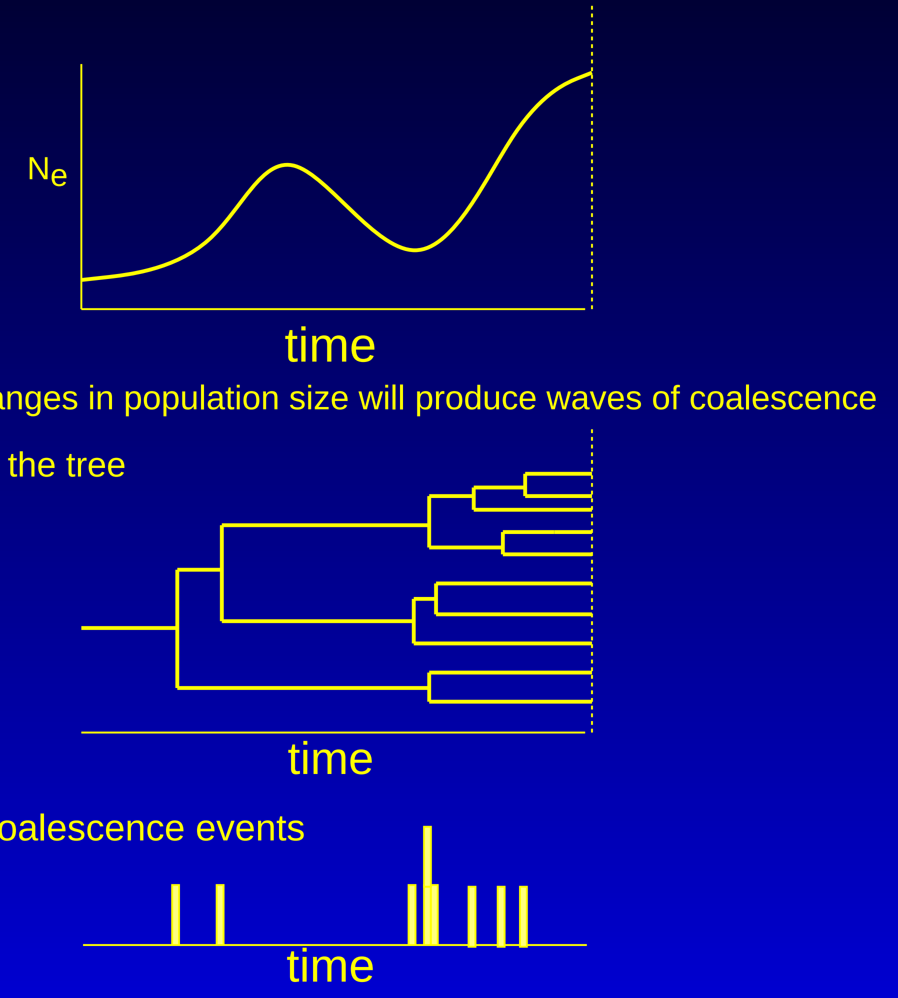
Average time for n

copies to coalesce

$$= \quad 4N \left(1 - \frac{1}{n}\right) \quad \text{generations}$$

Coalescence is faster in small populations

Change of population size and coalescents



the changes in population size will produce waves of coalescence

The parameters of the growth curve for N_e can be inferred by likelihood methods as they affect the prior probabilities of those trees that fit the data.

“Skyline” and “Skyride” plots in BEAST

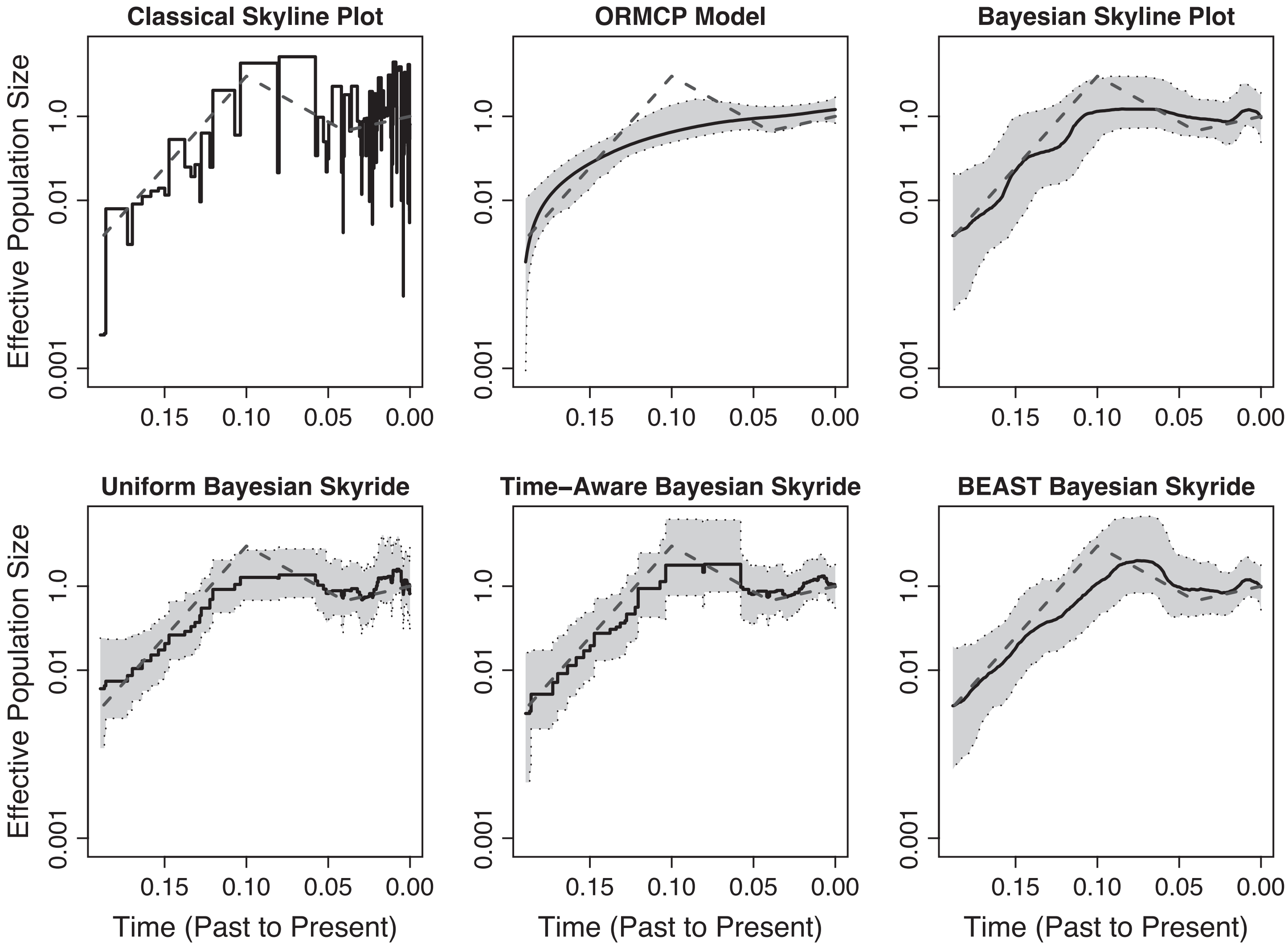


Figure from Minin, Bloomquist, and Suchard 2008

BEST Liu and Pearl (2007); Edwards et al. (2007)

- X – sequence data
- G – a genealogy (gene tree – with branch lengths)
- S – a species tree
- $\boldsymbol{\theta}$ – demographic parameters
- $\boldsymbol{\Lambda}$ – parameters of molecular sequence evolution

$$\begin{aligned}
\Pr(S,\boldsymbol{\theta}|X) &= \frac{\Pr(S,\boldsymbol{\theta})\Pr(X|S,\boldsymbol{\theta})}{\Pr(X)} \\
&= \Pr(S)\Pr(\boldsymbol{\theta})\int \Pr(X|G)\Pr(G|S,\boldsymbol{\theta})dG \\
&\propto \Pr(S)\Pr(\boldsymbol{\theta})\int \left[\int \Pr(X|G,\boldsymbol{\Lambda})\Pr(\boldsymbol{\Lambda})d\boldsymbol{\Lambda}\right]\Pr(G|S,\boldsymbol{\theta})dG
\end{aligned}$$

BEST – importance sampling

1. Generate a collection of gene trees, $\boldsymbol{G}$, using an approximation of the coalescent prior
2. Sample from the distribution of the species trees conditional on the gene trees, $\boldsymbol{G}$.
3. Use “importance weights” to correct the sample for the fact that an approximate prior was used

BEST – importance sampling

1. Generate a collection of gene trees, $\boldsymbol{G}$, using an approximation of the coalescent prior
   (a) Use a tweaked version of MrBayes to sample N sets of gene trees, $\boldsymbol{G}$, from

$$\overset{\dagger}{\Pr}(\boldsymbol{G}|X) = \frac{\Pr^{\dagger}(\boldsymbol{G})\Pr(X|\boldsymbol{G})}{\Pr^{\dagger}(X)}$$

   (b) $Pr^{\dagger}(\boldsymbol{G})$ is an approximate prior on gene trees from using a “maximal” species tree.
2. Sample from the distribution of the species trees conditional on the gene trees, $\boldsymbol{G}$.
3. Use “importance weights” to correct the sample for the fact that an approximate prior was used

BEST – importance sampling

1. Generate a collection of gene trees, $\boldsymbol{G}$, using an approximation of the coalescent prior
2. Sample from the distribution of the species trees conditional on the gene trees, $\boldsymbol{G}$.
   (a) From each set of gene trees (G_j for $1 \leq j \leq N$) generate k species trees using coalescent theory:

$$\Pr(S_i|\boldsymbol{G}_j) = \frac{\Pr(S_i)\Pr(\boldsymbol{G}_j|S_i)}{\Pr(\boldsymbol{G}_j)}$$

3. Use “importance weights” to correct the sample for the fact that an approximate prior was used

BEST – importance sampling

1. Generate a collection of gene trees, $\boldsymbol{G}$, using an approximation of the coalescent prior
2. Sample from the distribution of the species trees conditional on the gene trees, $\boldsymbol{G}$.
3. Use "importance weights" to correct the sample for the fact that an approximate prior was used
   (a) Estimate $\widehat{\Pr}(\boldsymbol{G}_j)$ by using the harmonic mean estimator from the MCMC in step 2.
   (b) Compute a normalization factor

$$\beta = \sum_{j=1}^{N} \frac{\widehat{\Pr}(\boldsymbol{G}_j)}{\Pr(\boldsymbol{G}_j)}$$

   (c) Reweight all sampled species trees by

$$\frac{\widehat{\Pr}(\boldsymbol{G}_j)}{\Pr(\boldsymbol{G}_j)}\beta$$

BEST – conclusions

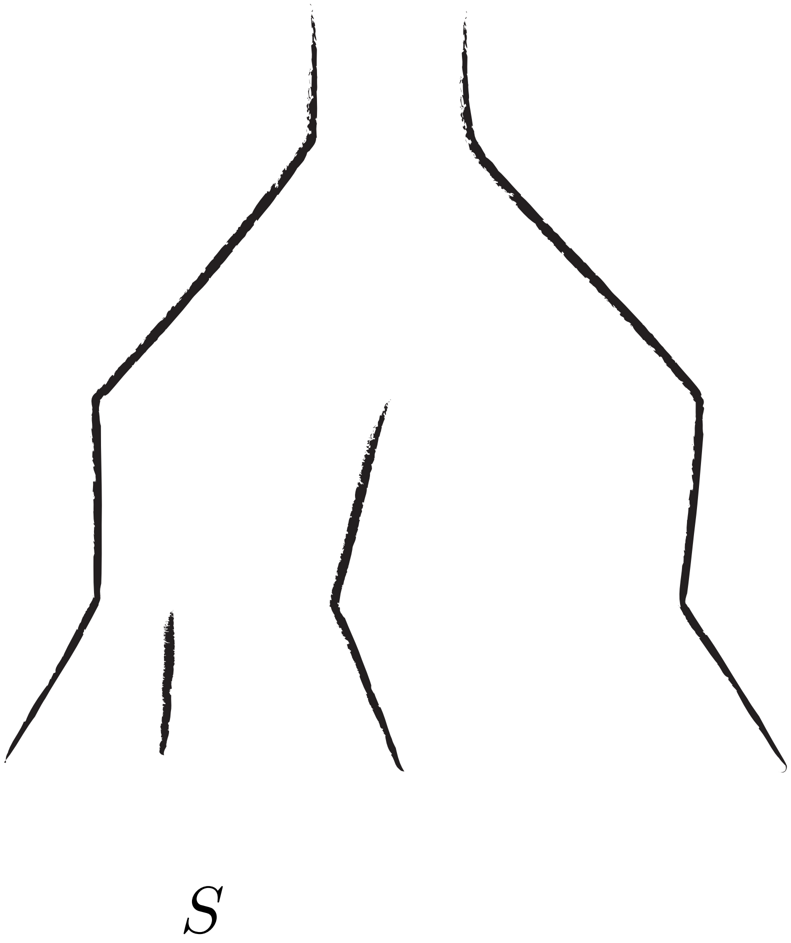
1. very expensive computationally (long MrBayes runs are needed)
2. should correctly deal with the variability in gene tree caused by the coalescent process.

*BEAST overview

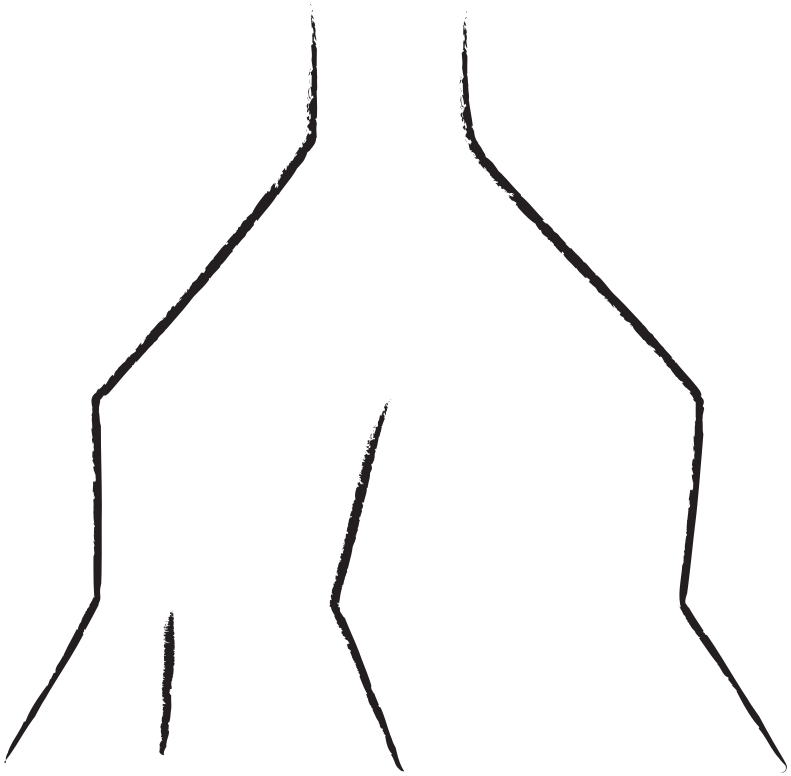
Goal: approximate $\Pr(S|X)$

$$
\begin{aligned}
\Pr(S|X) &\propto \Pr(X|S)\Pr(S) \\
&= \int \Pr(X|G)\Pr(G|S)\Pr(S)dG \\
&= \int\int \Pr(X|G)\Pr(G|S,\boldsymbol{\theta})\Pr(S)dGd\boldsymbol{\theta} \\
&= \int\int\int \Pr(X|G,\boldsymbol{\Lambda})\Pr(G|S,\boldsymbol{\theta})\Pr(S)dGd\boldsymbol{\theta}d\boldsymbol{\Lambda} \\
\boldsymbol{\theta} &= \{N_1, N_2, \ldots,\} \\
\boldsymbol{\Lambda} &= \{\kappa, \boldsymbol{\pi}, \ldots\}
\end{aligned}
$$

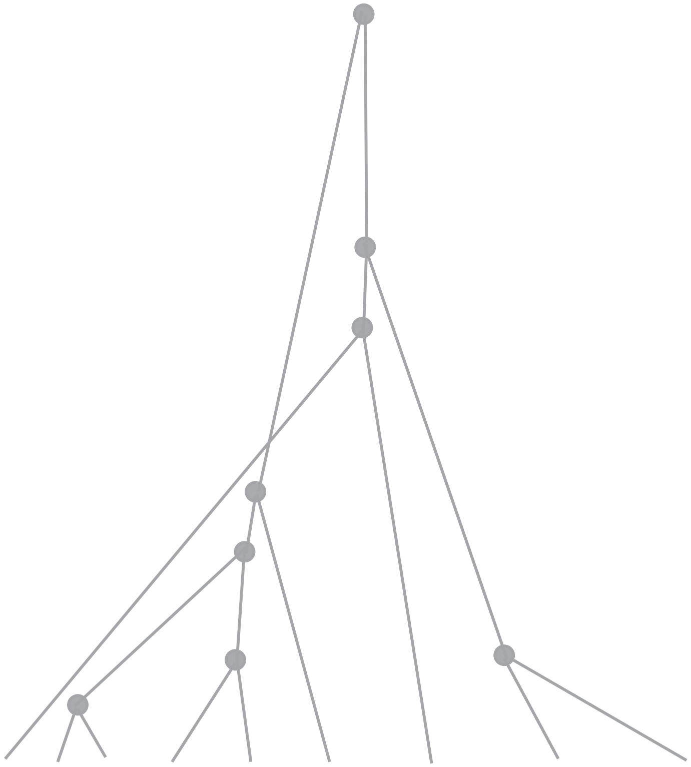
$\Pr(S)$ **from speciation model**



$\Pr(G|S)$



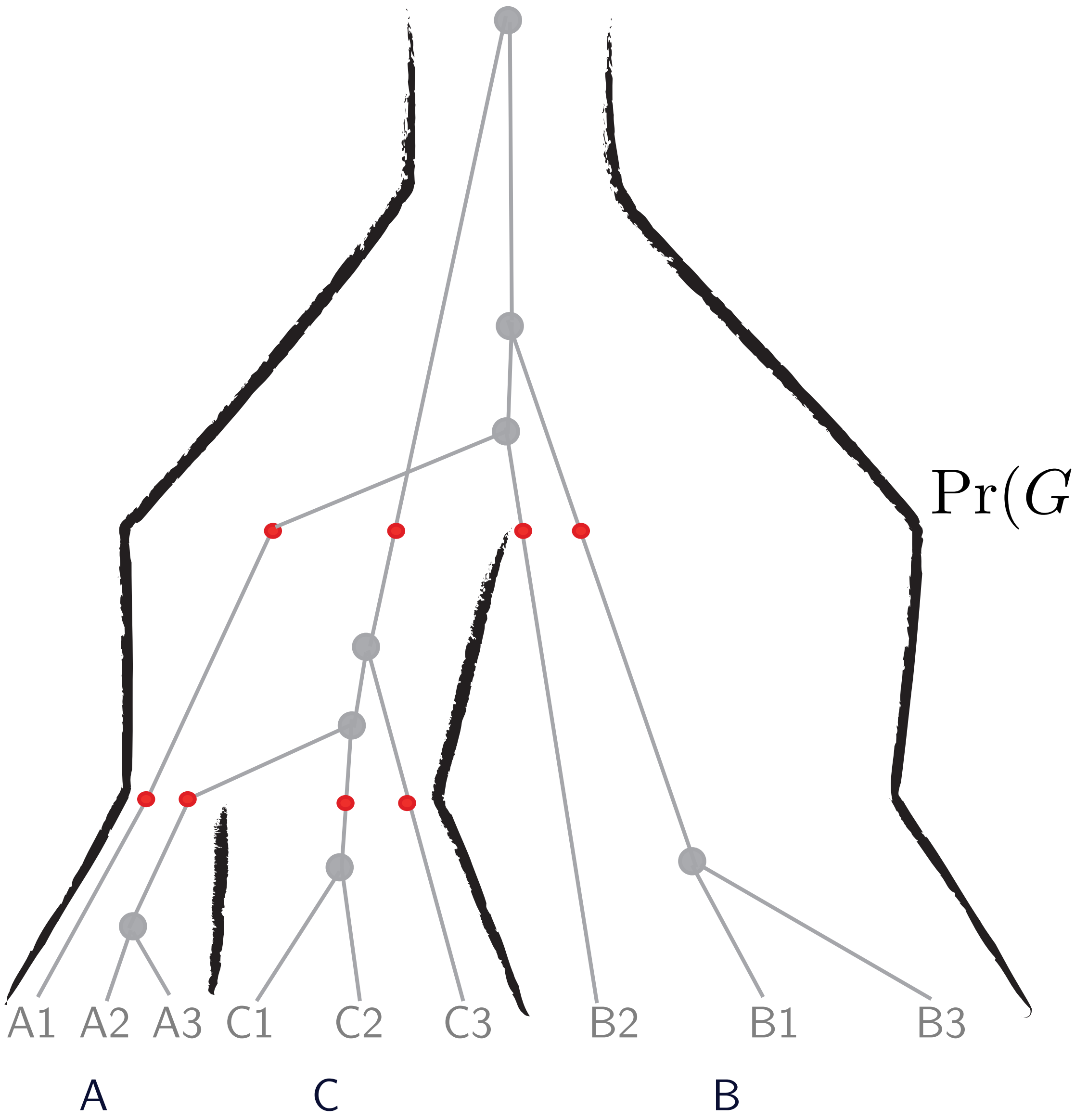
S



G

Gene tree in a species tree w/ variable population size

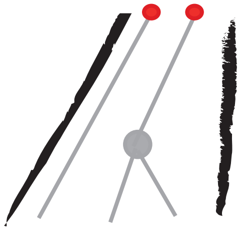
G in grey
S in black



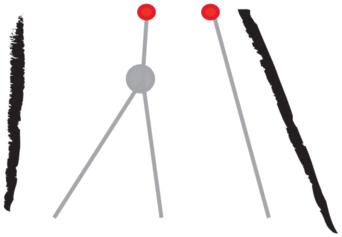
$$\Pr(G|S) = \prod_i^b \Pr(G_i|S_i)$$

Figure modified from Heled and Drummond 2010

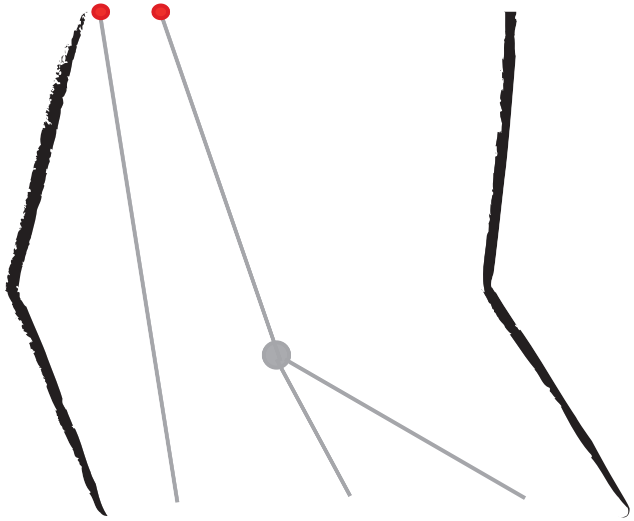
In Species A



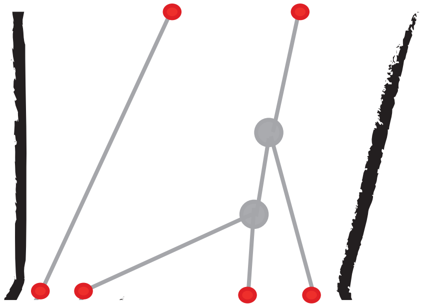
In Species C



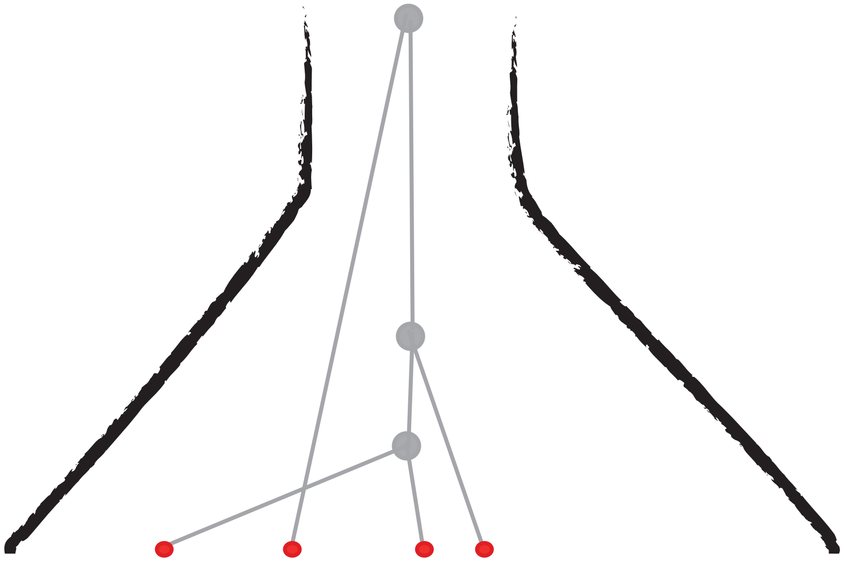
In Species B



In ancestor of AC

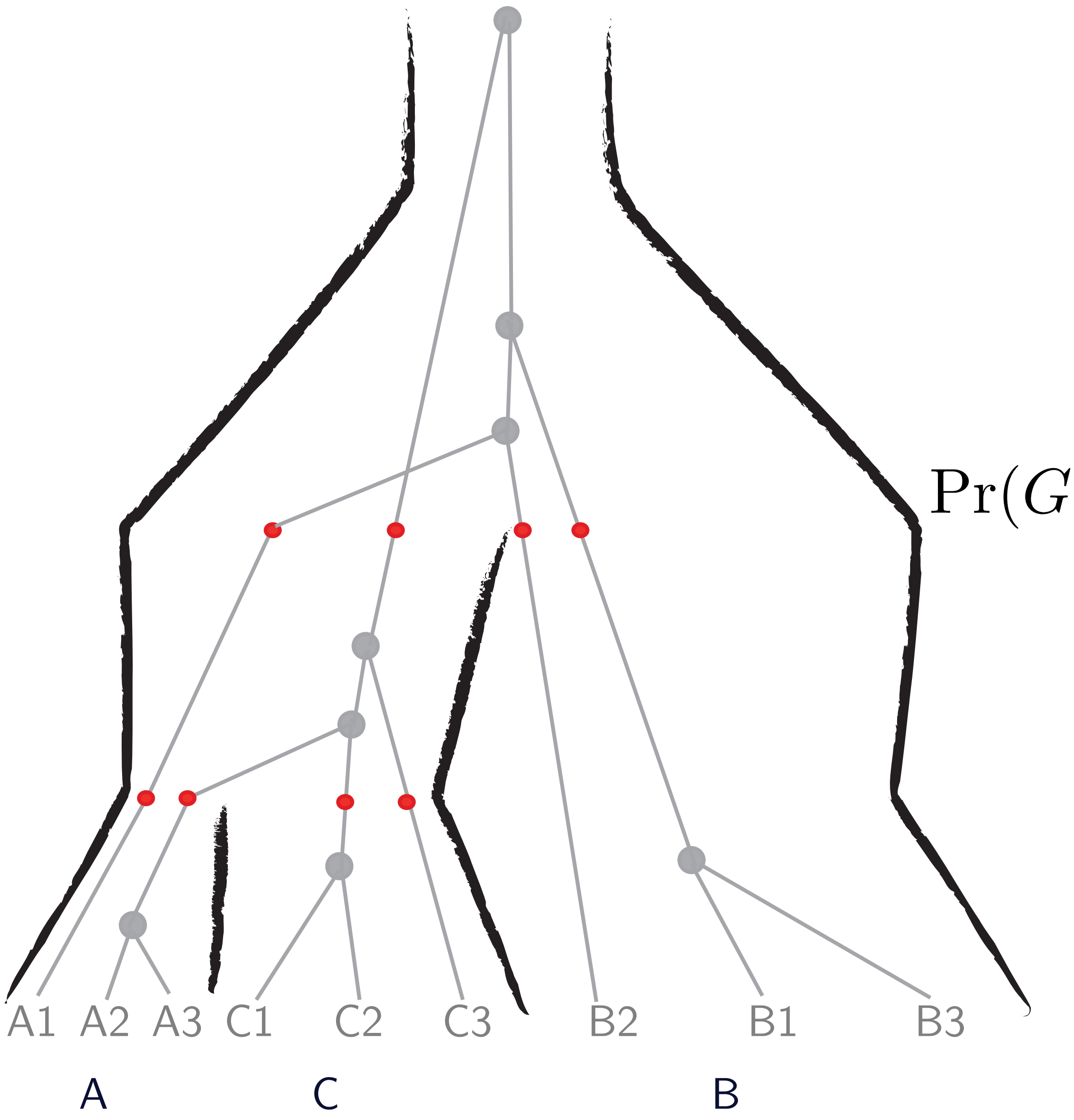


In ancestor of ACB



Gene tree in a species tree w/ variable population size

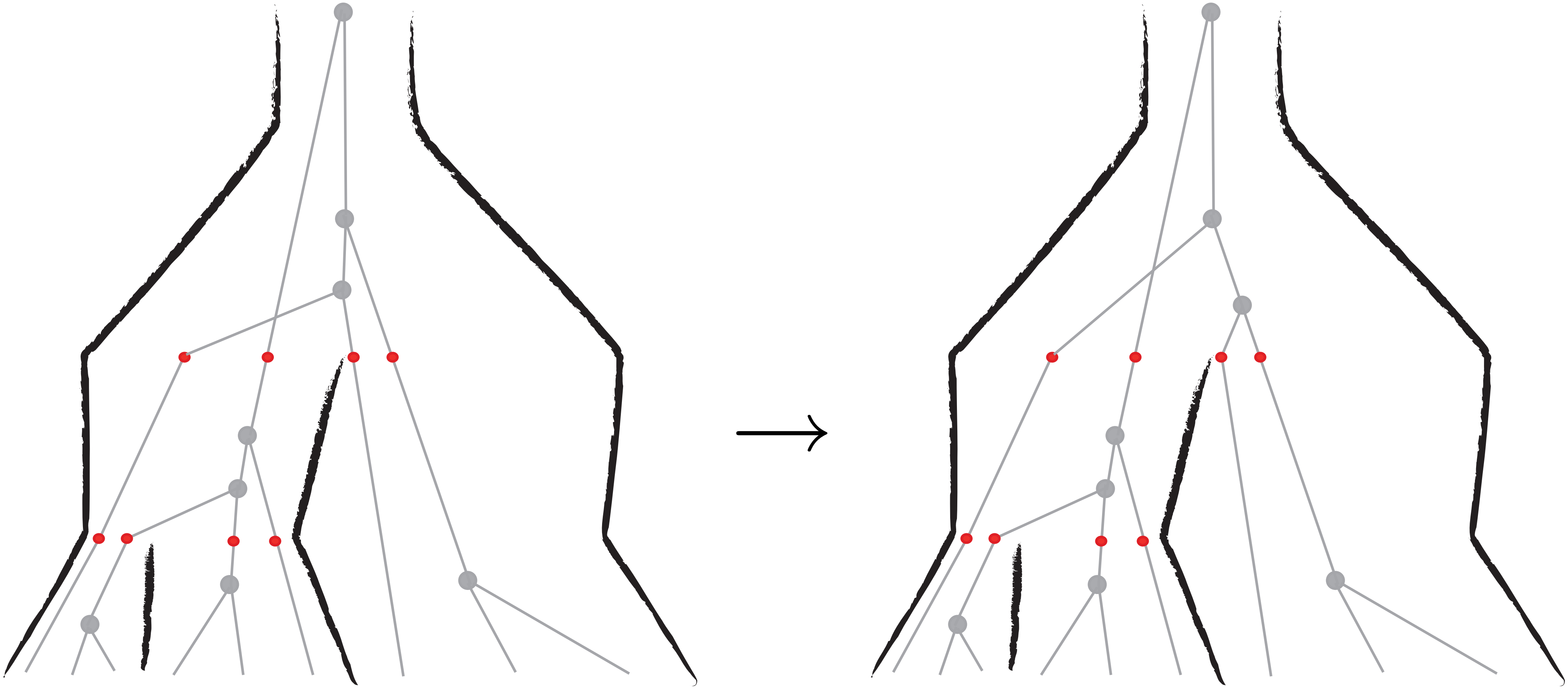
G in grey
S in black



$$\Pr(G|S) = \prod_i^b \Pr(G_i|S_i)$$

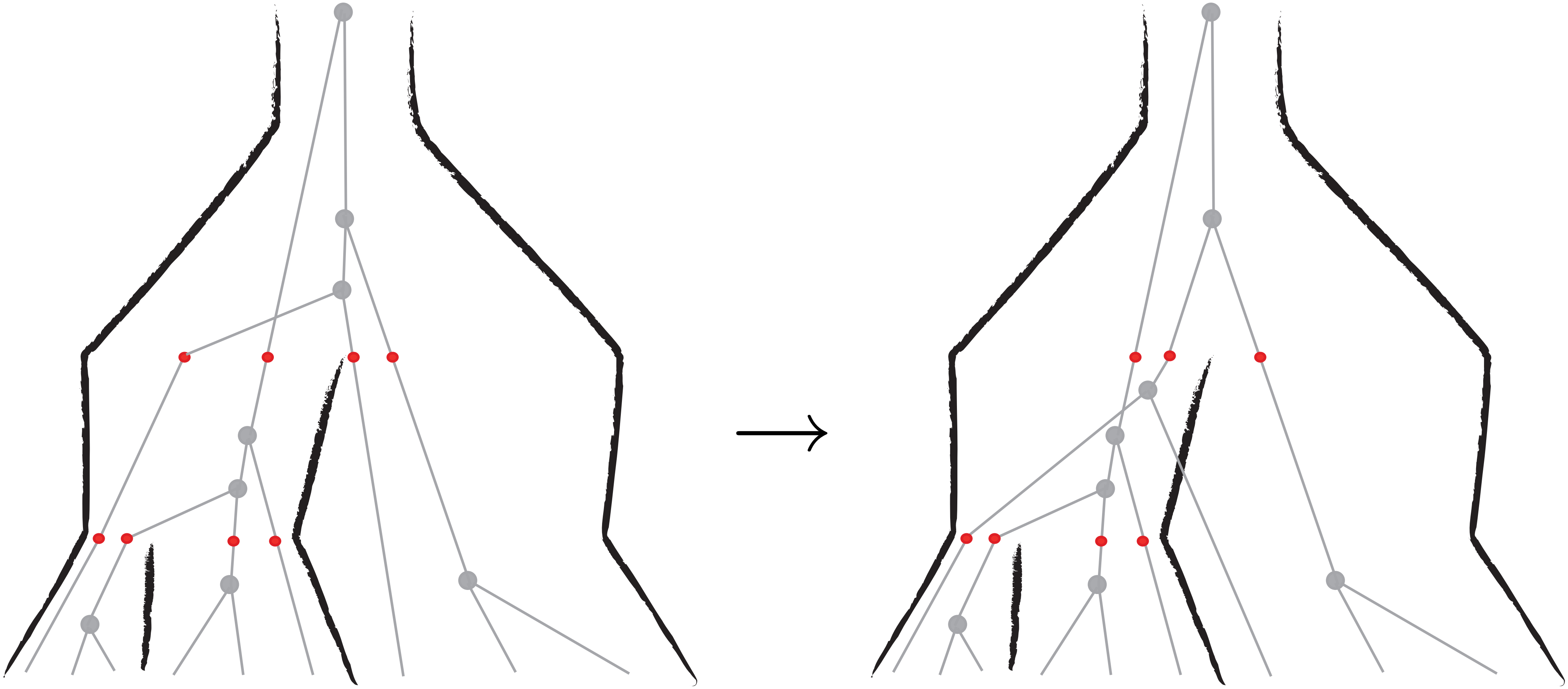
Figure modified from Heled and Drummond 2010

MCMC update to gene tree



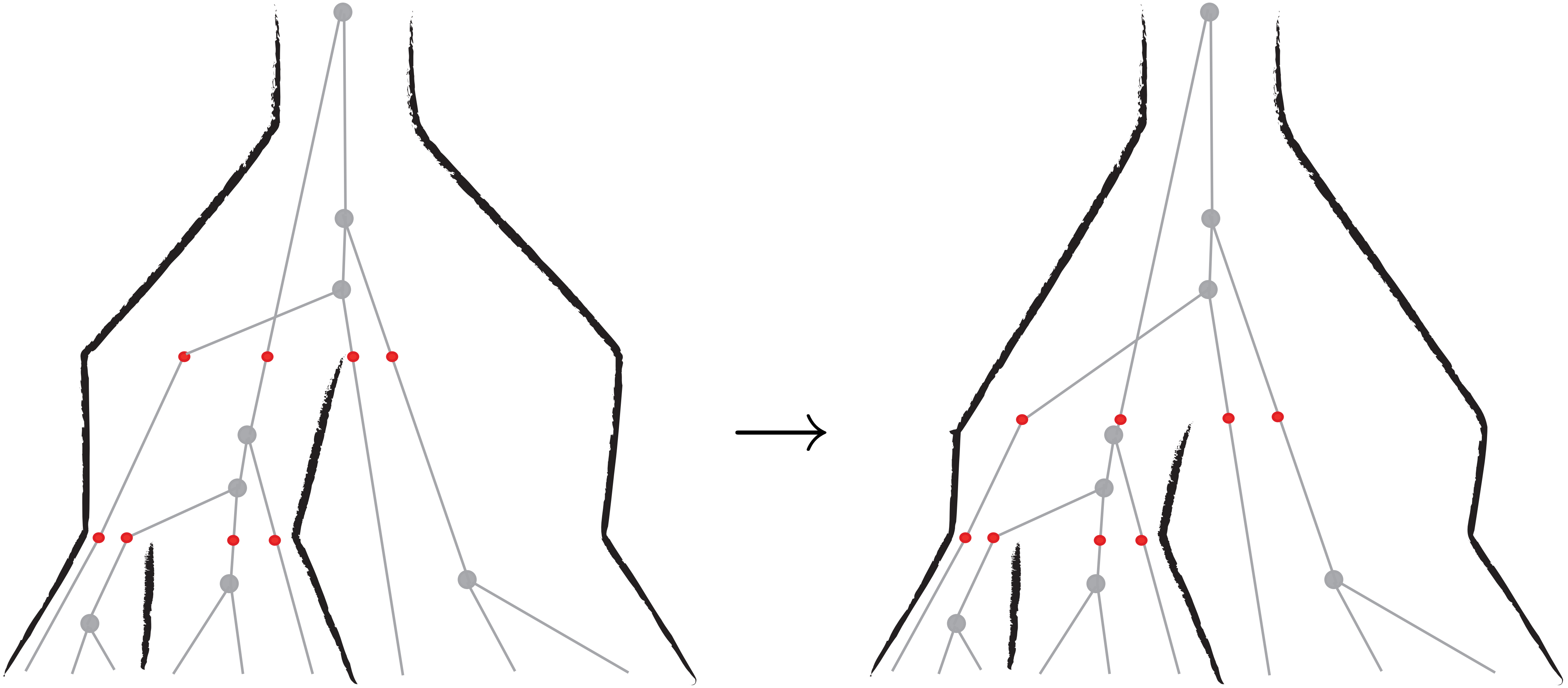
Changing G affects $\Pr(X|G,\boldsymbol{\Lambda})$ and $\Pr(G|S,\boldsymbol{\theta})$

Another MCMC update to gene tree



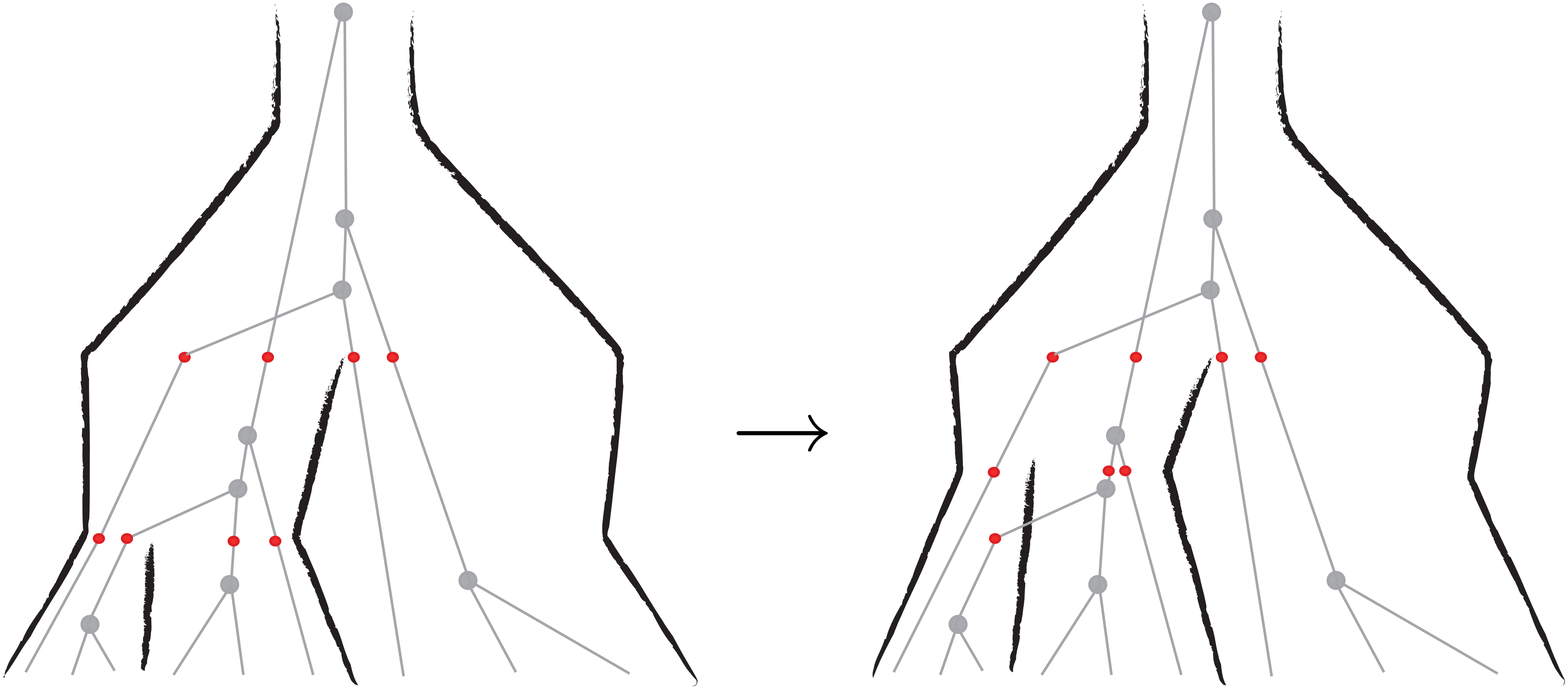
Some changes to G are incompatible with S (and will be rejected).

MCMC update to species tree



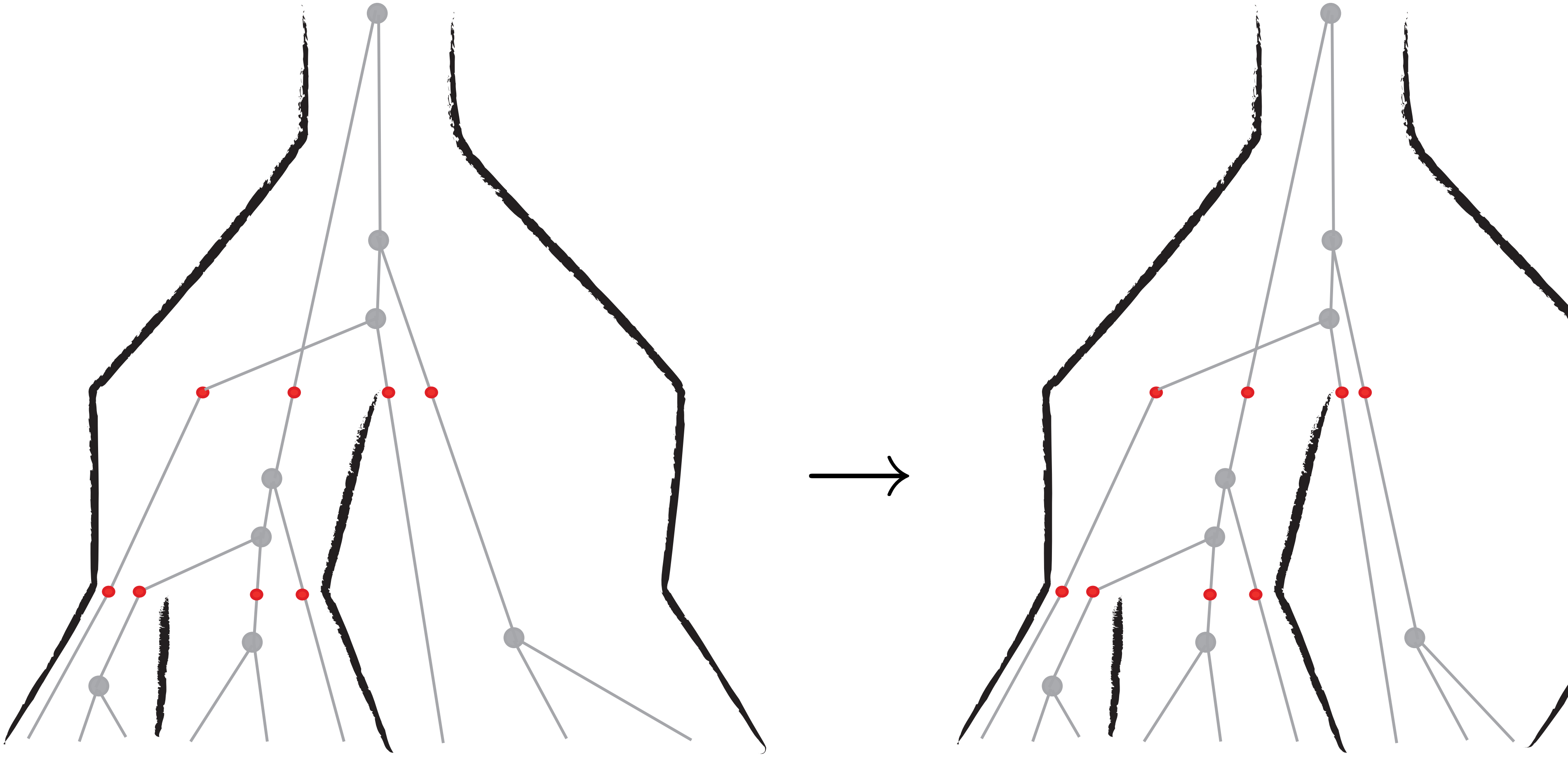
Changing S affects $\Pr(G|S,\boldsymbol{\theta})$, but not $\Pr(X|G,\boldsymbol{\Lambda})$.
Note that the red dots are “flags” for when a lineage enters a new species; the heights are determined by the species tree.

Another MCMC update to species tree



Some changes to S are incompatible with C (and will be rejected).

An MCMC update to the population size



$N_e \in \boldsymbol{\theta}$, so changing N_e affects $\Pr(G|S, \boldsymbol{\theta})$, but not $\Pr(X|G, \boldsymbol{\Lambda})$.

Multiple gene tree in a species tree w/ variable population size

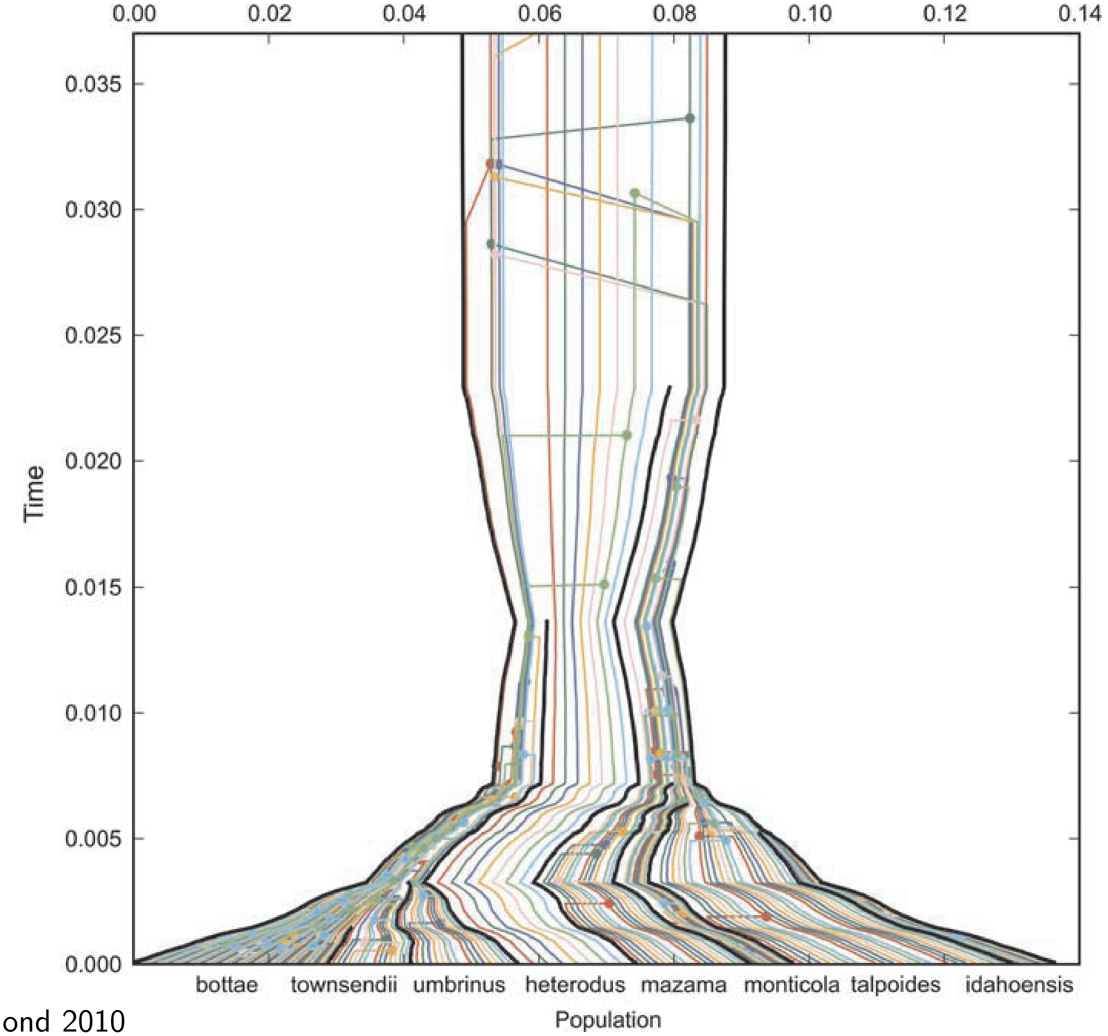


Figure from modified Heled and Drummond 2010

*BEST

Similar model to BEST, but *much* more efficient implementation.

Both attempt to sample the posterior distribution of species trees, gene trees, demographic parameter values and mutational parameter values.

Both will be very sensitive to migration, but they represent the state-of-the-art for estimating species trees from gene trees.

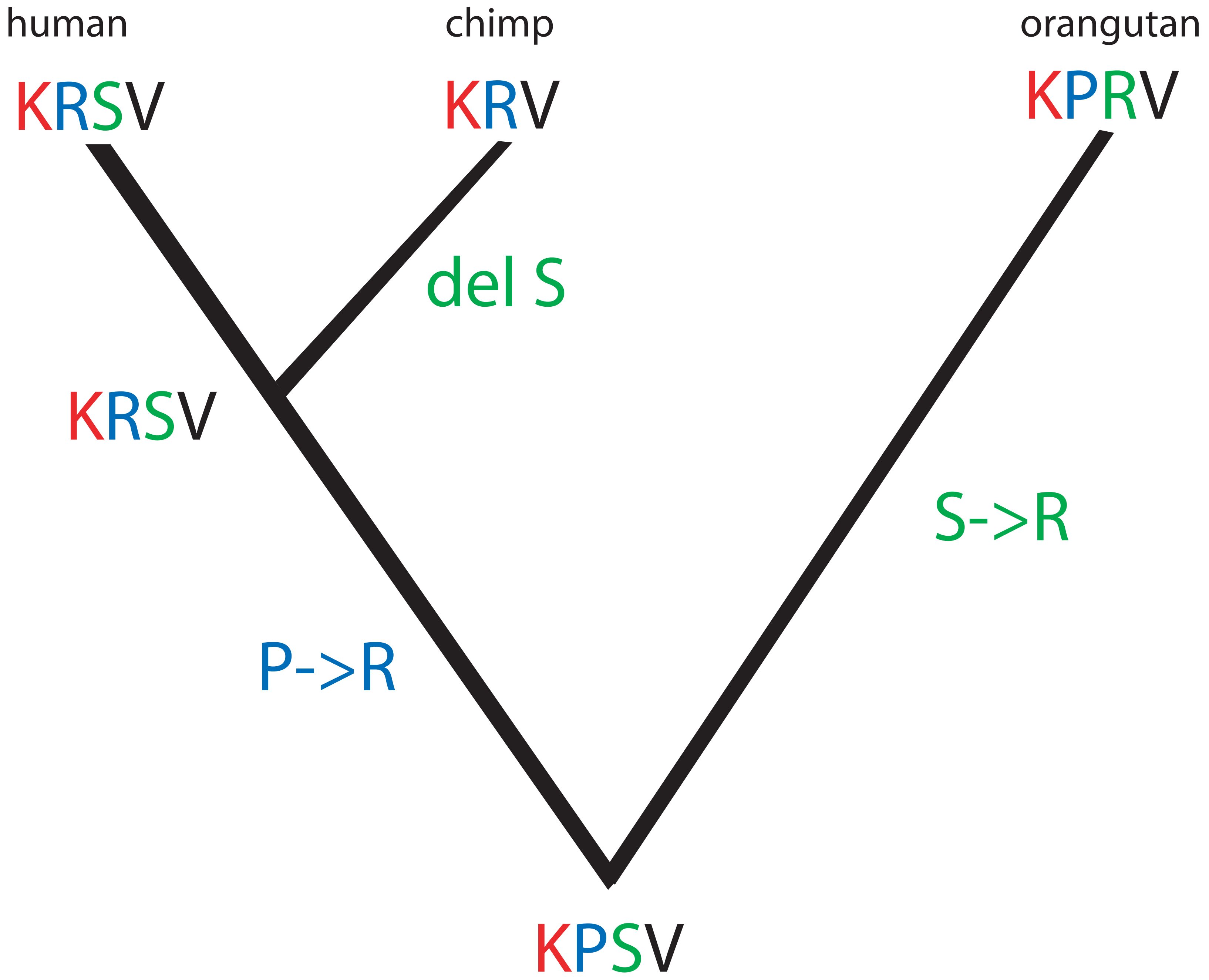
Multiple Sequence Alignment - main points

- The goal of MSA is to introduce gaps such that residues in the same column are homologous (all residues in the column descended from a residue in their common ancestor).
- The problem is recast as:
  - reward matches (+ scores)
  - penalize rare substitutions (- scores),
  - penalize gaps (- scores),
  - try to find an alignment that maximizes the total score
- pairwise alignment is tractable
- MSA is usually done progressively
- progressive alignment algorithms are heuristic, and do not optimize an evolutionary defensible criterion

Multiple Sequence Alignment tools

- clustal variants are popular, but not very reliable.
- simultaneous inference of MSA and tree is the most defensible (but computationally demanding)
- Promising tools for MSA (roughly in order of computational tractability):
  1. Simultaneous MSA + Trees (Handel, BAliPhy, BEAST, AliFritz...)
  2. FSA (fast statistical alignment); Infernal (for rRNA); Prank
  3. MAFFT, Muscle, ProbCons
- Iterative "meta-solutions" (e.g. SATè) allow MSA uncertainty to be incorporated in tree inference.
- GBlocks (and similar tools) cull ambiguously aligned regions.

human	KRSV
chimp	KRV
orang	KPRV


human
chimp
orangutan
KRSV
KRV
KPRV
del S
KRSV
S->R
P->R
KPSV

human	KRSV
chimp	KRV
gorilla	KSV
orang	KPRV

How should we align these sequences?

human	KRSV
chimp	KR-V
gorilla	KS-V
orang	KPRV

OR

human	KRSV
chimp	K-RV
gorilla	K-SV
orang	KPRV

Pairwise alignment

Gap penalties and a substitution matrix imply a score for any alignment. Pairwise alignment involves finding the alignment that maximizes this score.

- substitution matrices assign positive values to matches or similar substitutions (for example Leucine→Isoleucine).

- unlikely substitutions receive negative scores

- gaps are rare and are heavily penalized (given large negative values).

Scoring an alignment. Simplest case

Costs:

Match	1
Mismatch	0
Gap	-5

Alignment:

Pongo	V	D	E	V	G	G	E	L	G	R	L	F	V	V	P	T	Q
Gorilla	V	E	V	A	G	D	L	G	R	L	L	I	V	Y	P	S	R
Score	1	0	0	0	1	0	0	0	0	0	1	0	1	0	1	0	0

Total score = 5

Scoring an different alignment. Simplest case

Match	1
Mismatch	0
Gap	-5

Pongo	V	D	E	V	G	G	E	L	G	R	L	-	F	V	V	P	T	Q
Gorilla	V	-	E	V	A	G	D	L	G	R	L	L	I	V	Y	P	S	R
Score	1	-5	1	1	0	1	0	1	1	1	1	-5	0	1	0	1	0	0

Total score = 0

BLOSUM 62 Substitution matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	−1	5																		
N	−2	0	6																	
D	−2	−2	1	6																
C	0	−3	−3	−3	9															
Q	−1	1	0	0	−3	5														
E	−1	0	0	2	−4	2	5													
G	0	−2	0	−1	−3	−2	−2	6												
H	−2	0	1	−1	−3	0	0	−2	8											
I	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
L	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
K	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
M	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
F	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
P	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
S	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
W	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Y	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
V	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4
	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**

Scoring an alignment with the BLOSUM 62 matrix

Pongo	V	D	E	V	G	G	E	L	G	R	L	F	V	V	P	T	Q
Gorilla	V	E	V	A	G	D	L	G	R	L	L	I	V	Y	P	S	R
Score	4	2	-2	0	6	-6	-3	-4	-2	-2	4	0	4	-1	7	4	1

The score for the alignment is

$$D_{ij} = \sum_k d_{ij}^{(k)}$$

If i indicates *Pongo* and j indicates *Gorilla*

$$D_{ij} = 12$$

Scoring an alignment with gaps

If the GP is -8:

Pongo	V	D	E	V	G	G	E	L	G	R	L	-	F	V	V	P	T	Q
Gorilla	V	-	E	V	A	G	D	L	G	R	L	L	I	V	Y	P	S	R
Score	4	-8	5	5	0	6	2	4	6	5	4	-8	0	4	-1	7	4	1

By introducing gaps we have improved the score:

$$D_{ij} = 40$$

Gap Penalties

Gaps are penalized more heavily than substitutions to avoid alignments like this:

```
Pongo    VDEVGGE-LGRLFVVPTQ
Gorilla  VDEVGG-WLGRLFVVPTQ
```

Gap Penalties

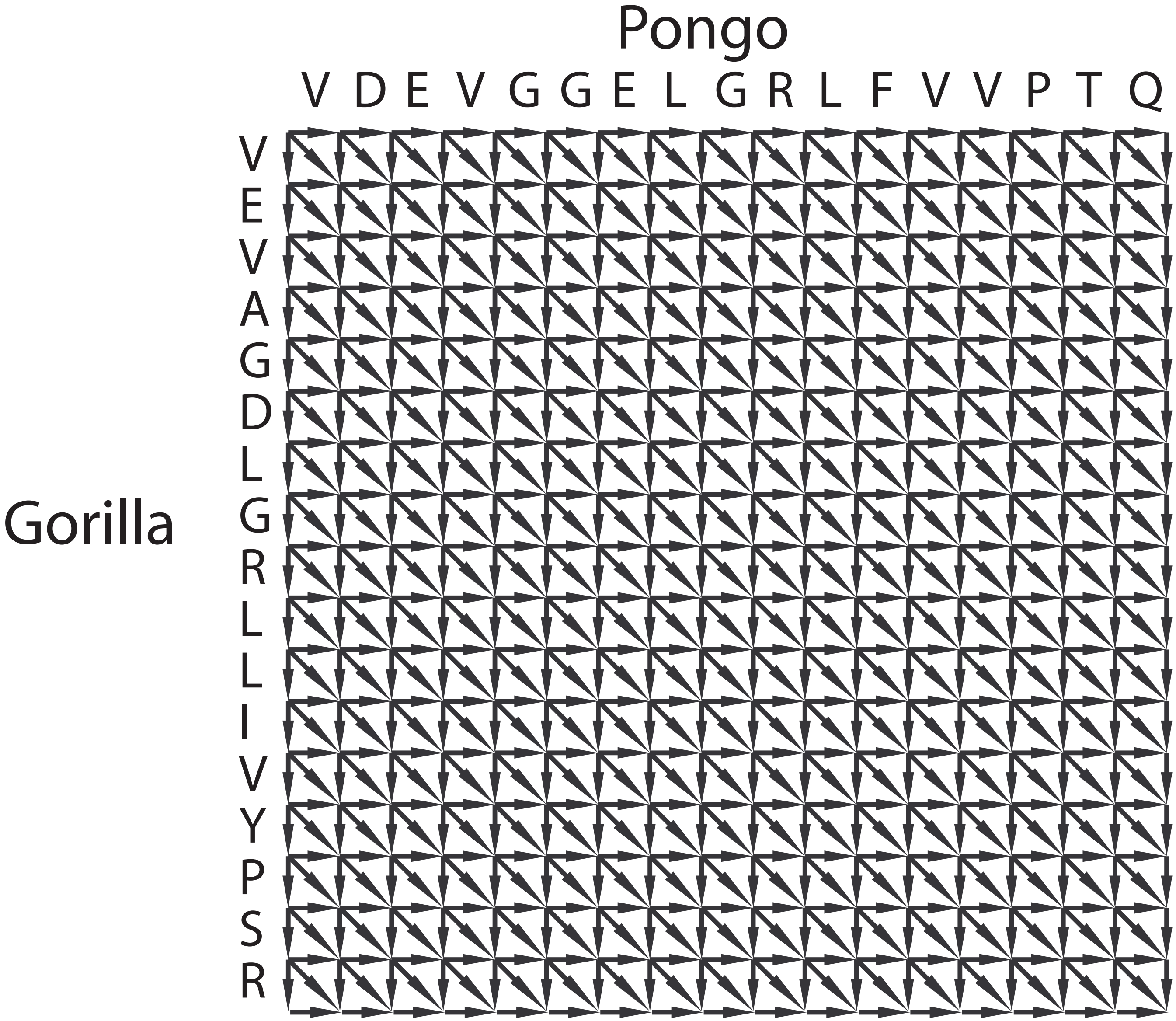
Because multiple residues are often inserted or deleted at the same time, *affine gap penalties* are often used:

$$\text{GP} = \text{GO} + l\text{GE}$$

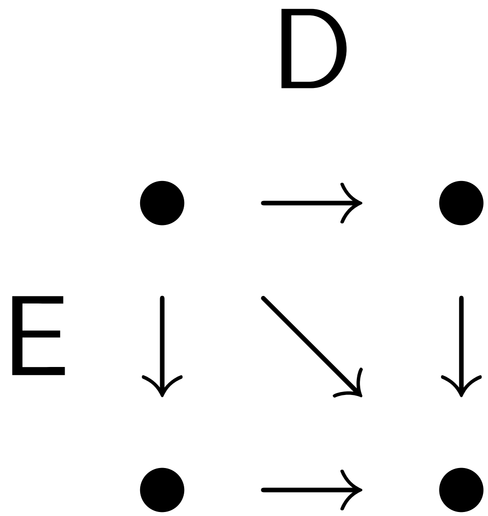
where:

- GP is the gap penalty.
- GO is the “gap-opening penalty”
- GE is the “gap-extension penalty”
- l is the length of the gap

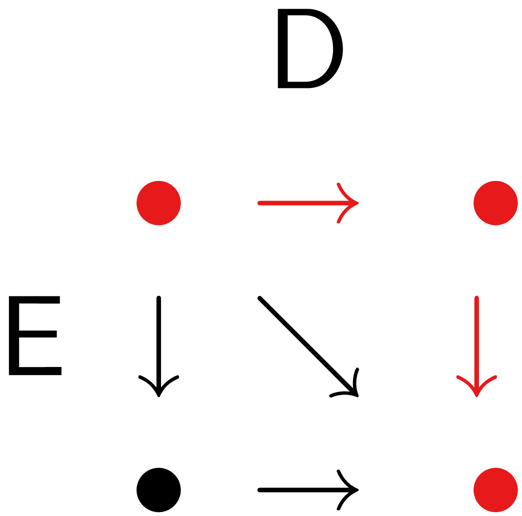
Finding an optimal alignment



Aligning two sequences, each with length = 1

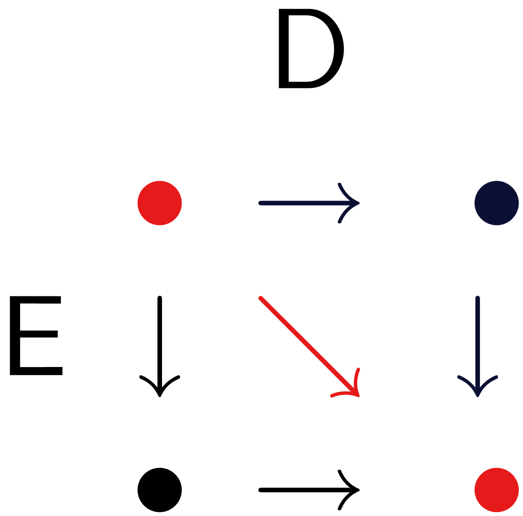


Alignment 1



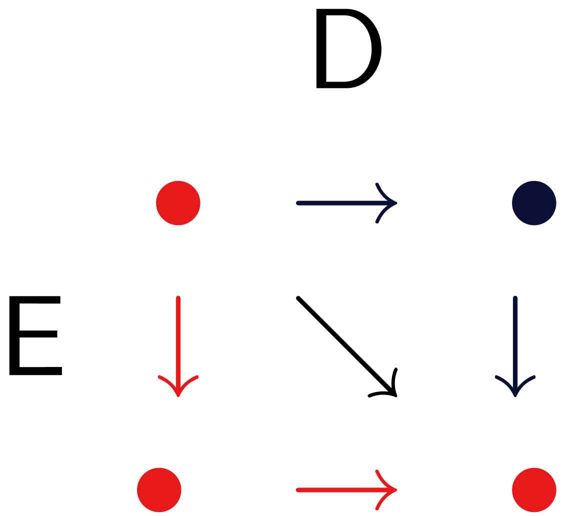
```
D-
-E
```

Alignment 2



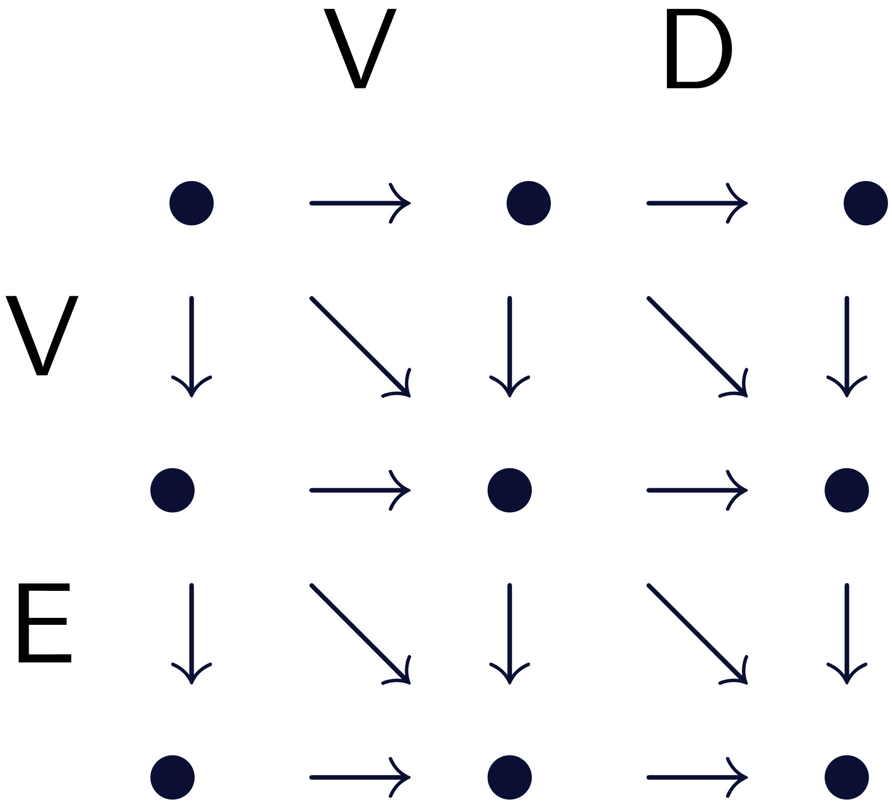
```
D
E
```

Alignment 3

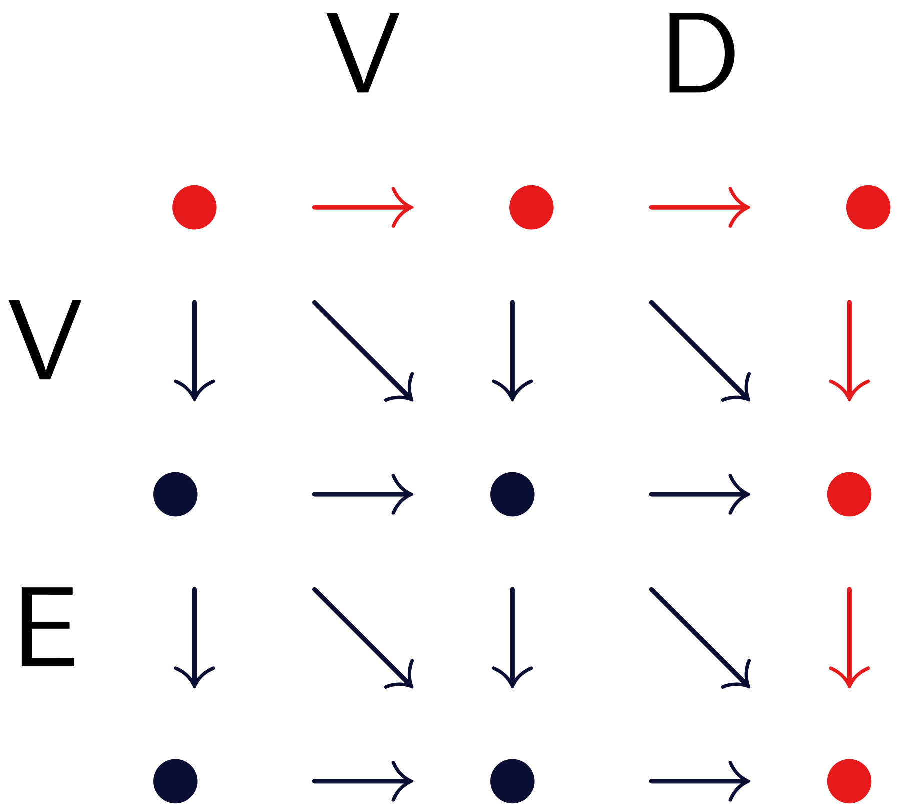


```
-D
E-
```

Longer sequences – up to 2 amino acids!

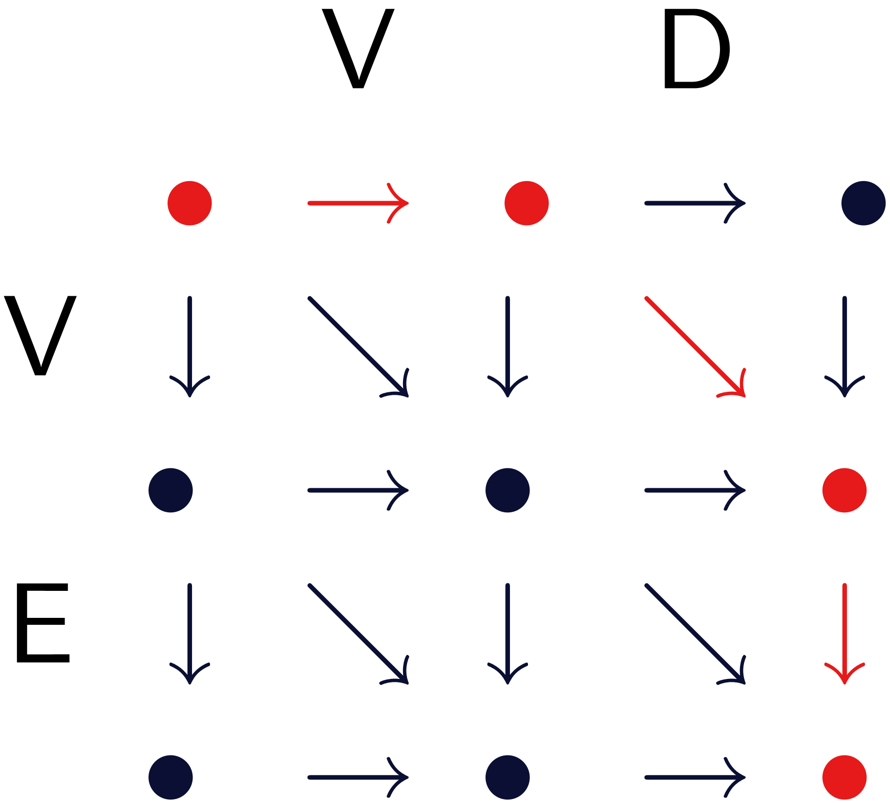


Alignment 1



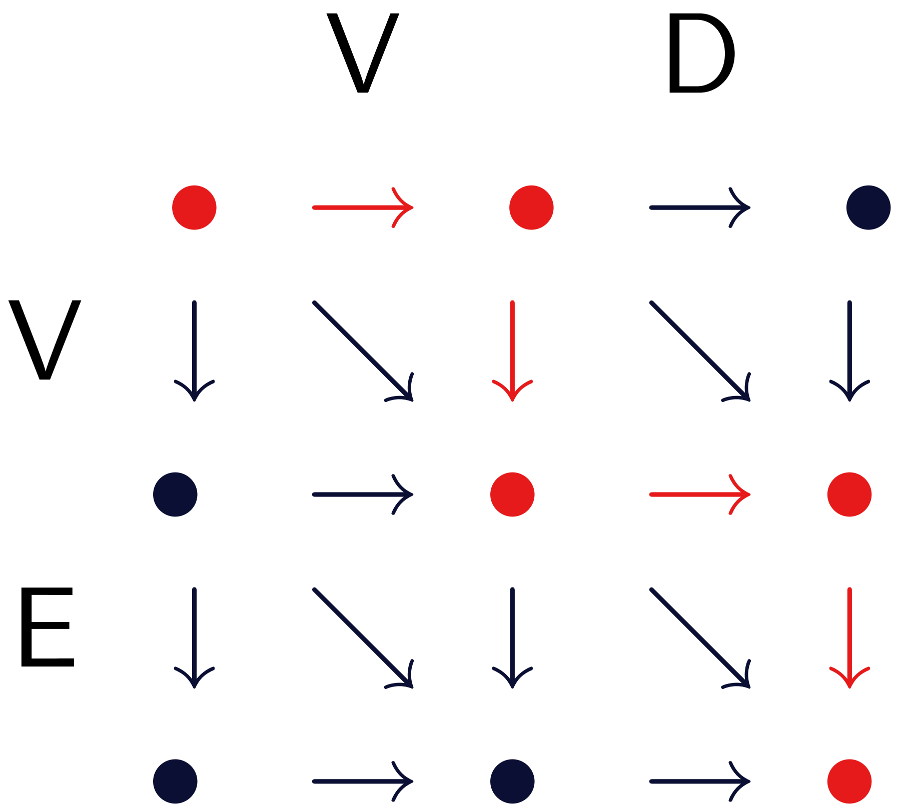
```
VD--
--VE
```

Alignment 2



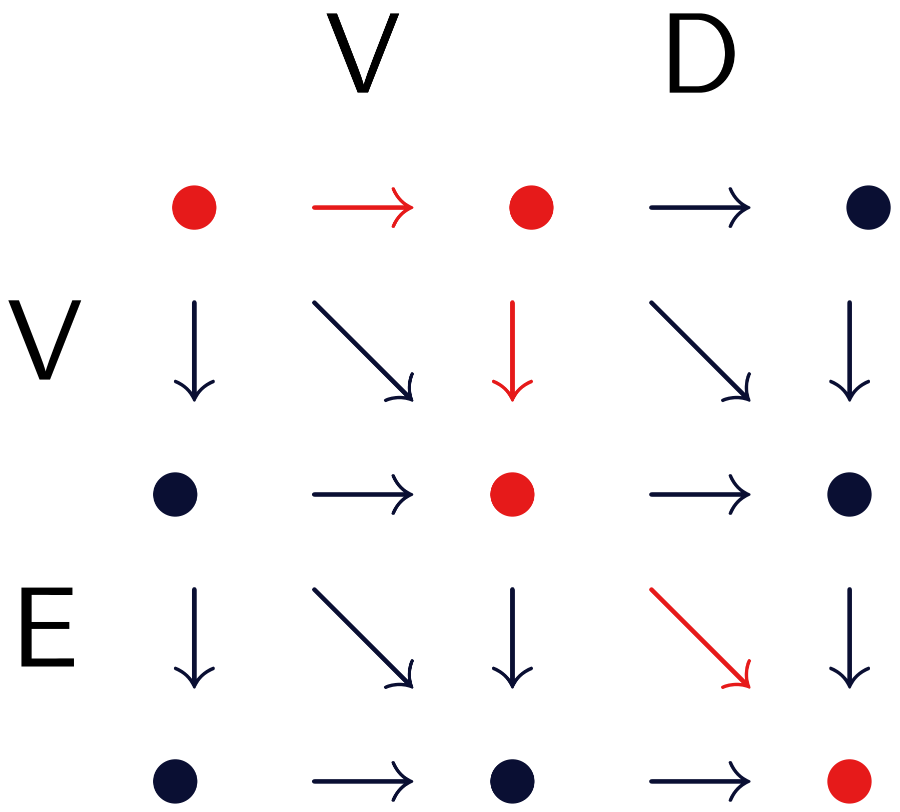
```
VD-
-VE
```

Alignment 3



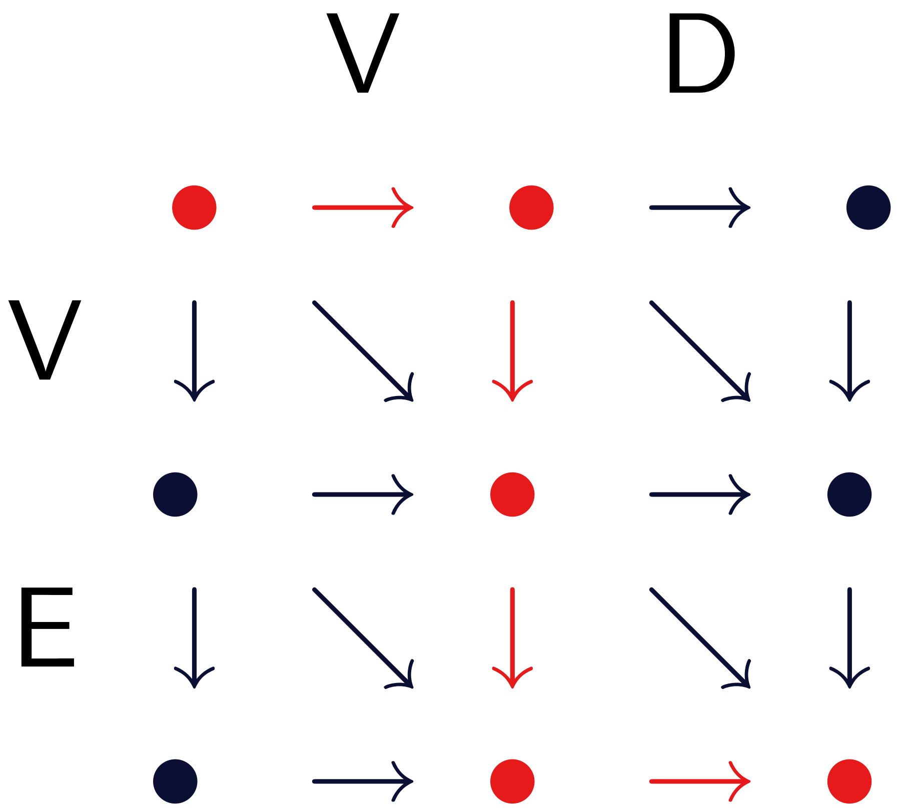
```
V-D-
-V-E
```

Alignment 4



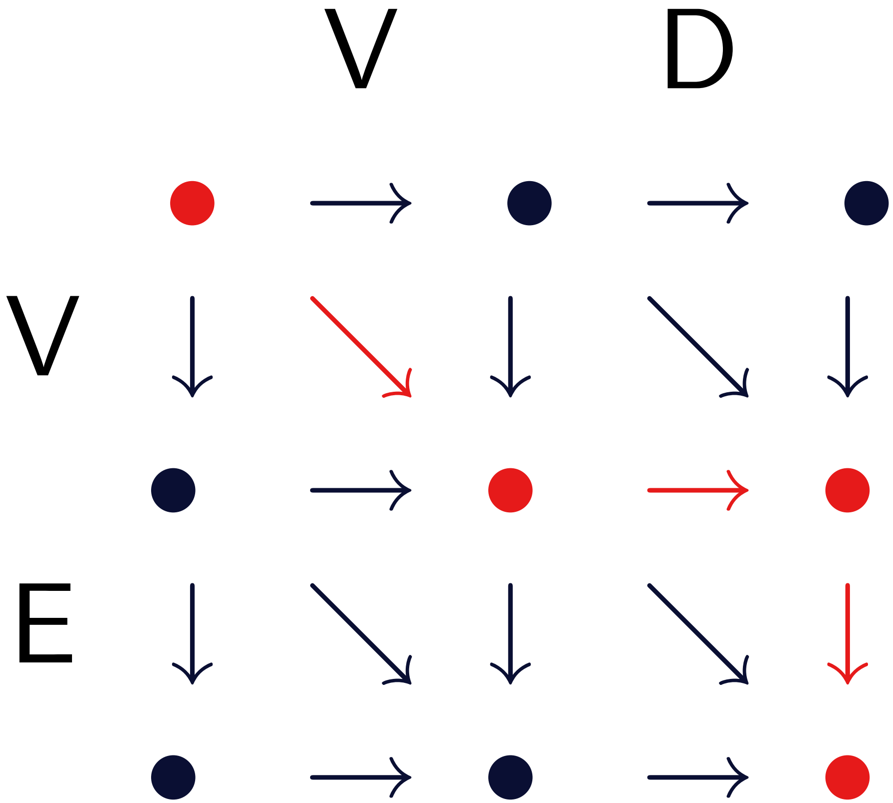
```
V-D
-VE
```

Alignment 5



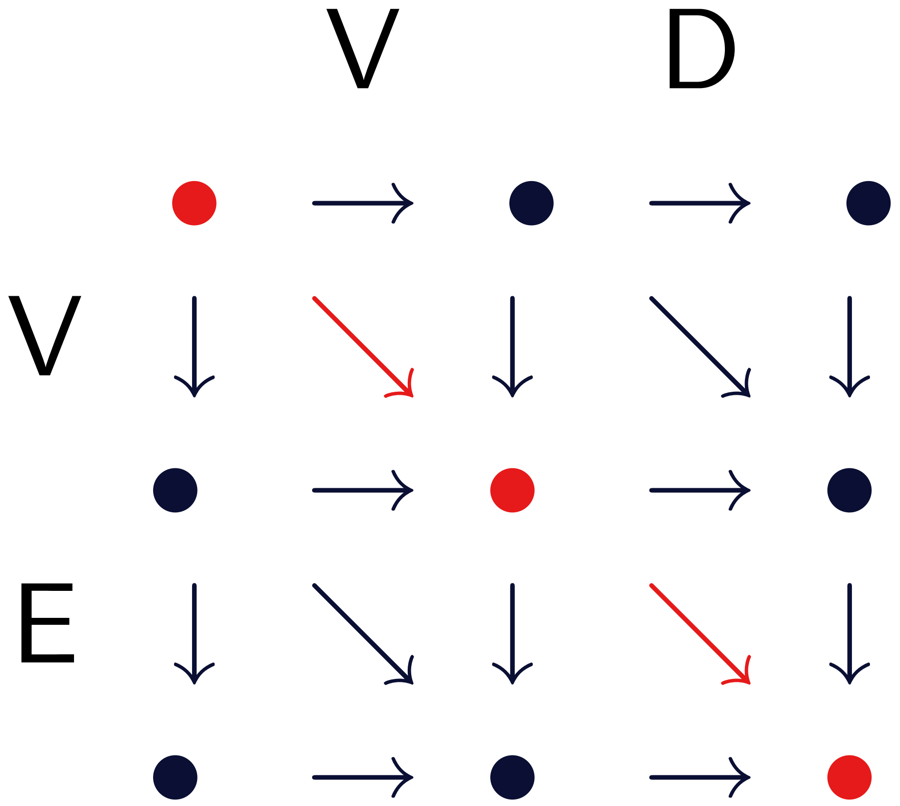
```
V--D
-VE-
```

Alignment 6



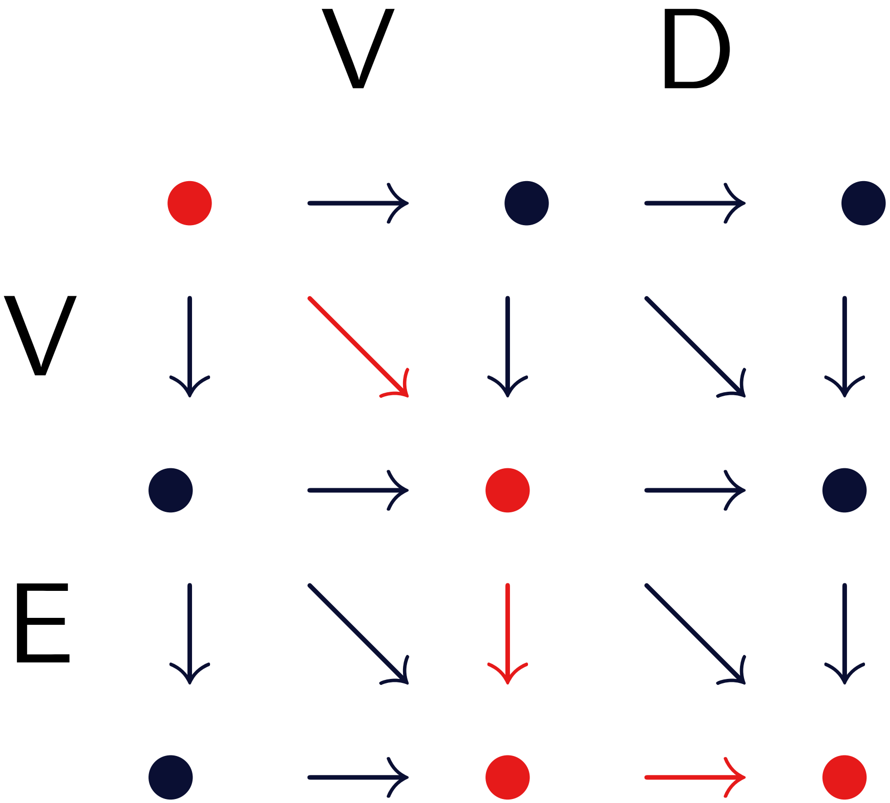
```
VD-
V-E
```

Alignment 7



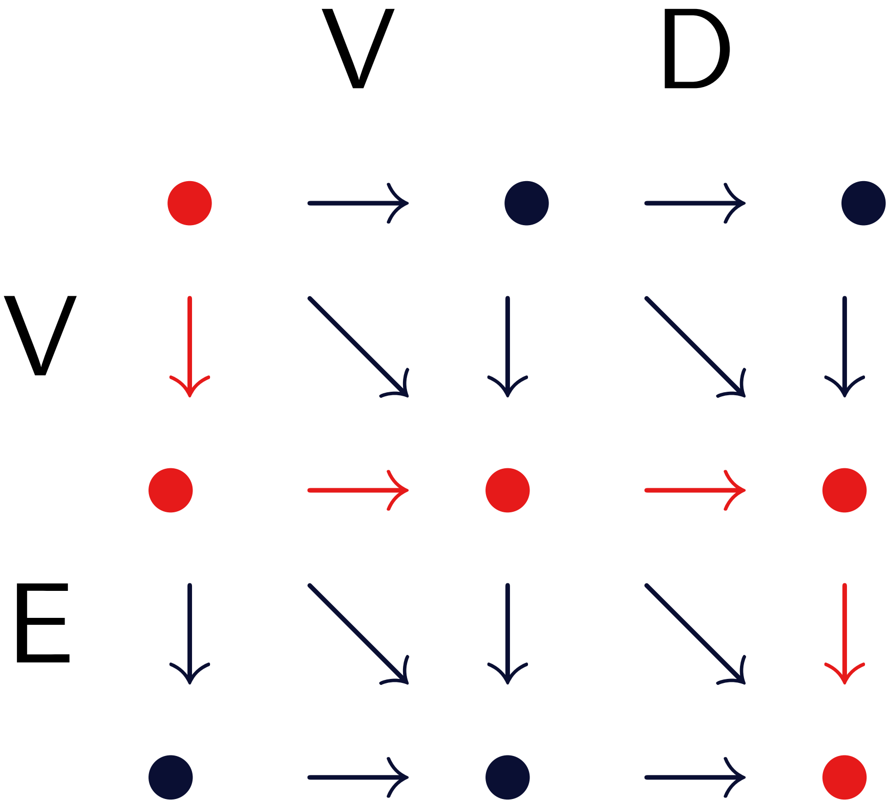
```
VD
VE
```

Alignment 8



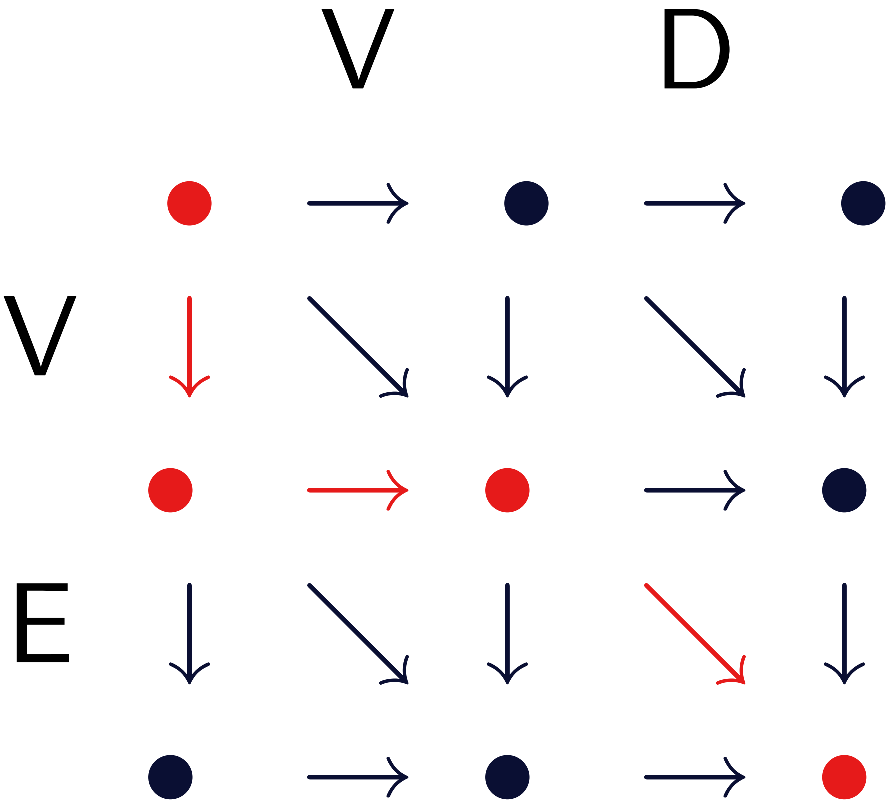
```
V-D
VE-
```

Alignment 9



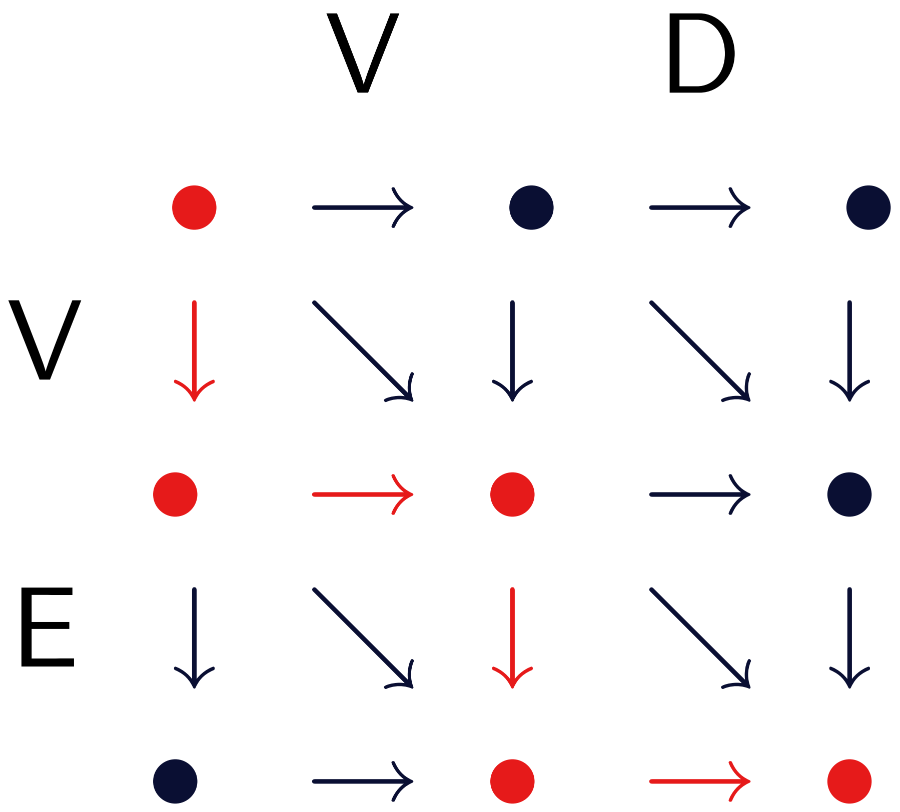
```
-VD-
V--E
```

Alignment 10



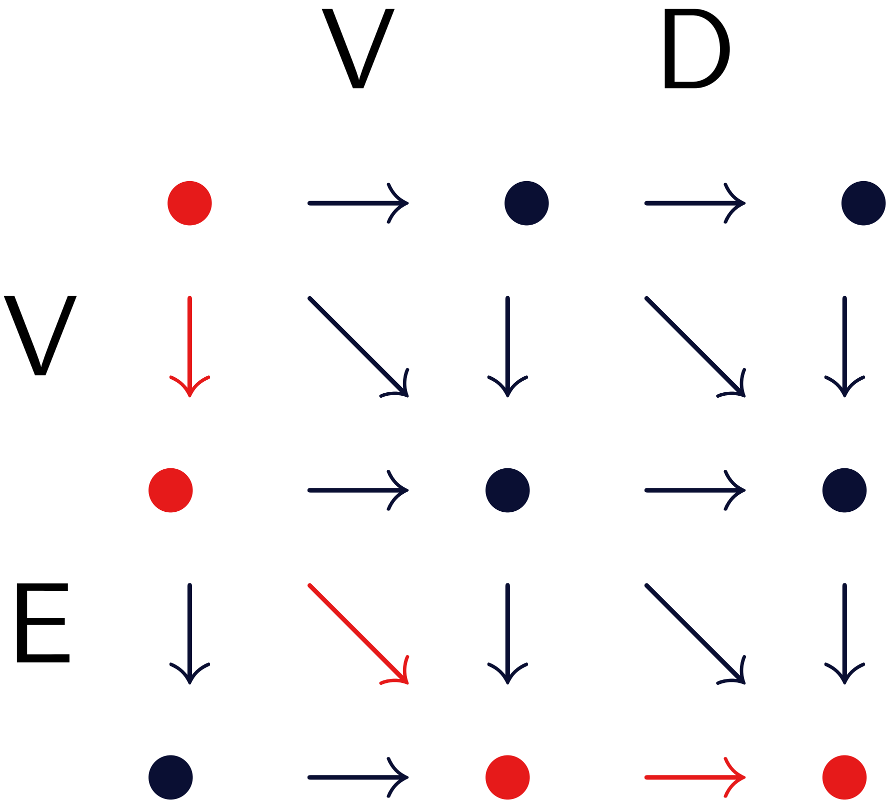
```
-VD
V-E
```

Alignment 11



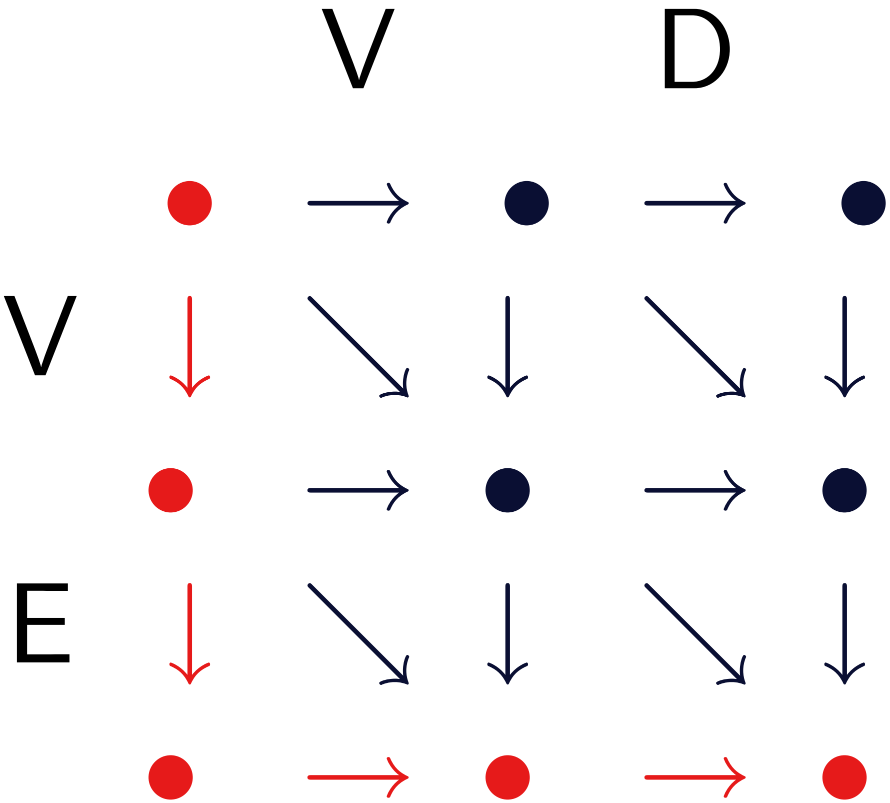
```
-V-D
V-E-
```

Alignment 12



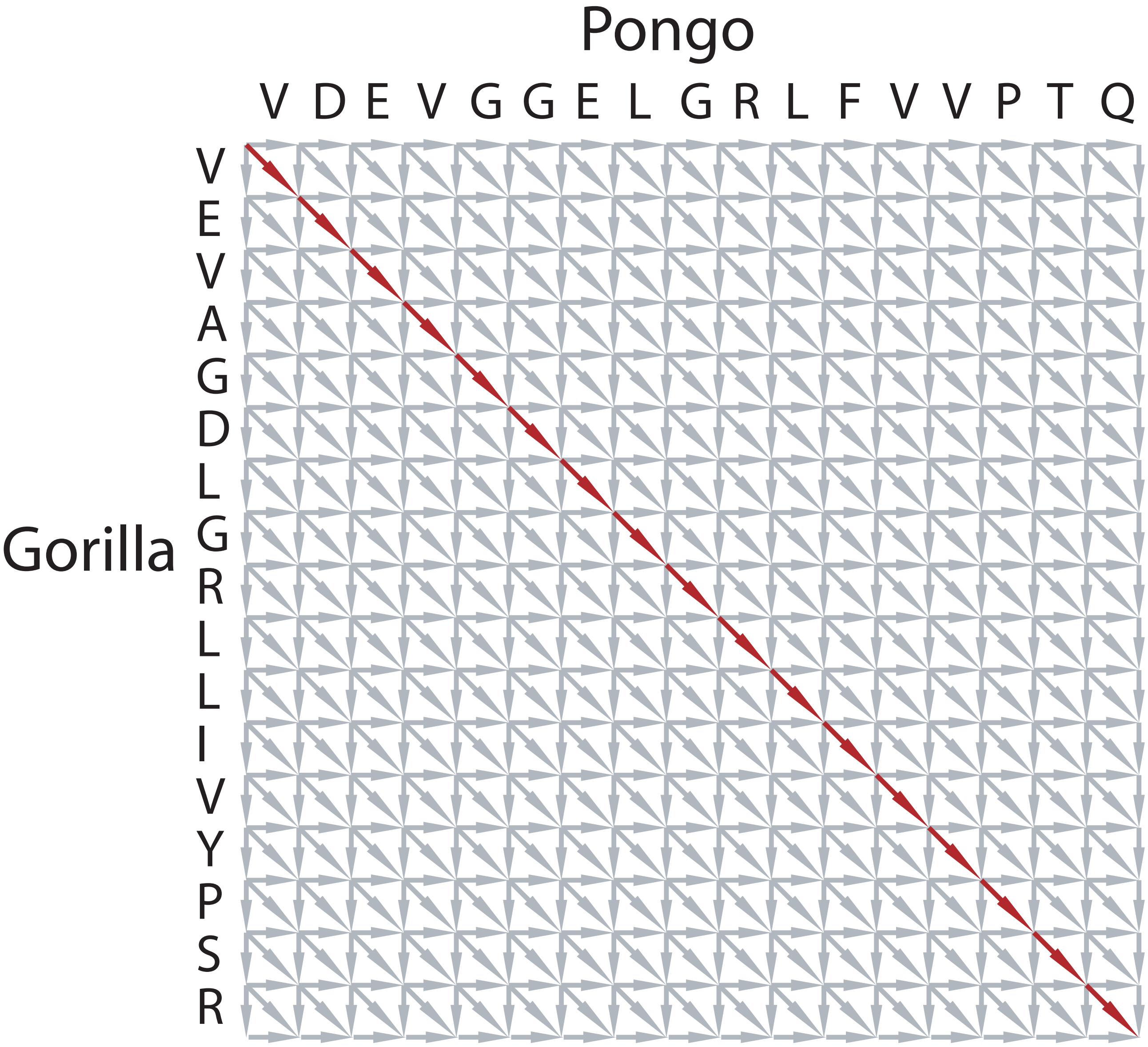
```
-VD
VE-
```

Alignment 13

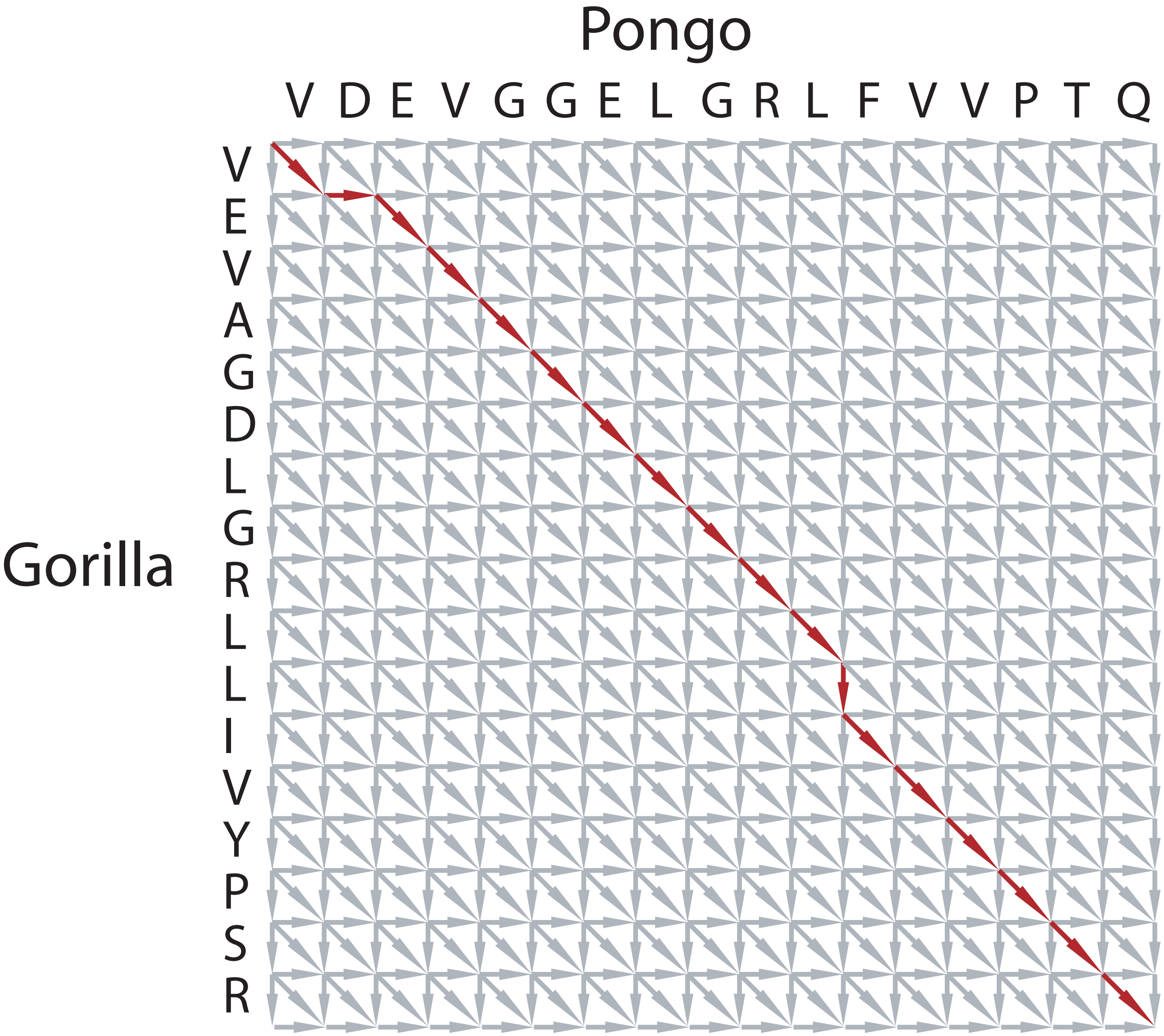


```
--VD
VE--
```

Pongo	V	D	E	V	G	G	E	L	G	R	L	F	V	V	P	T	Q
Gorilla	V	E	V	A	G	D	L	G	R	L	L	I	V	Y	P	S	R
Score	4	2	-2	0	6	-6	-3	-4	-2	-2	4	0	4	-1	7	4	1


Pongo
V D E V G G E L G R L F V V P T Q
Gorilla
V E V A G D L G R L L I V Y P S R

Pongo	V	D	E	V	G	G	E	L	G	R	L	-	F	V	V	P	T	Q
Gorilla	V	-	E	V	A	G	D	L	G	R	L	L	I	V	Y	P	S	R
Score	4	-8	5	5	0	6	2	4	6	5	4	-8	0	4	-1	7	4	1



length Seq # 1	length Seq # 2	# alignments
1	1	3
2	2	13
3	3	63
4	4	321
5	5	1,683
6	6	8,989
7	7	48,639
8	8	265,729
9	9	1,462,563
⋮	⋮	⋮
17	17	1,425,834,724,419

Needleman-Wunsch algorithm (paraphrased)

- Work from the top left (beginning of both sequences)
- For each cell store the highest score possible for that cell and a “back” pointer to tell point to the previous step in the best path
- When you reach the lower right corner, you know the optimal score and the back pointers tell you the alignment.

The highest-score calculation at each cell only depends on its the cell’s three possible previous neighbors.

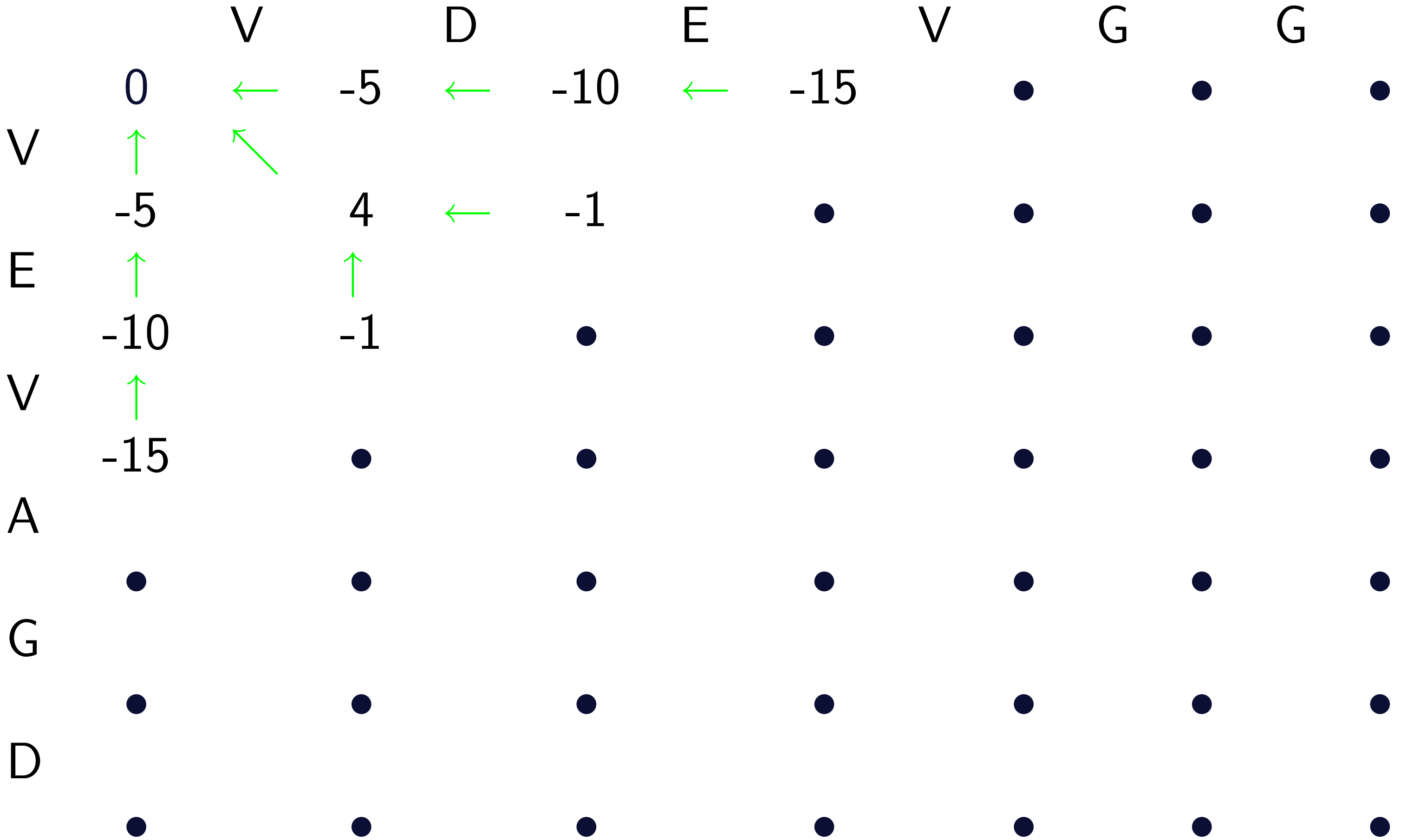
If one sequence is length N, and the other is length M, then Needleman-Wunsch only takes $\approx 6NM$ calculations.

But there are a **much** larger number of possible alignments.

		V		D		E		V		G		G	
	0		●		●		●		●		●		●
V													
	●		●		●		●		●		●		●
E													
	●		●		●		●		●		●		●
V													
	●		●		●		●		●		●		●
A													
	●		●		●		●		●		●		●
G													
	●		●		●		●		●		●		●
D													
	●		●		●		●		●		●		●

		V	D	E	V	G	G
	0	← -5	●	●	●	●	●
V	↑ -5	●	●	●	●	●	●
E	●	●	●	●	●	●	●
V	●	●	●	●	●	●	●
A	●	●	●	●	●	●	●
G	●	●	●	●	●	●	●
D	●	●	●	●	●	●	●

		V	D	E	V	G	G
	0	← -5	← -10	●	●	●	●
V	↑ -5	↖ 4	●	●	●	●	●
E	↑ -10	●	●	●	●	●	●
V	●	●	●	●	●	●	●
A	●	●	●	●	●	●	●
G	●	●	●	●	●	●	●
D	●	●	●	●	●	●	●



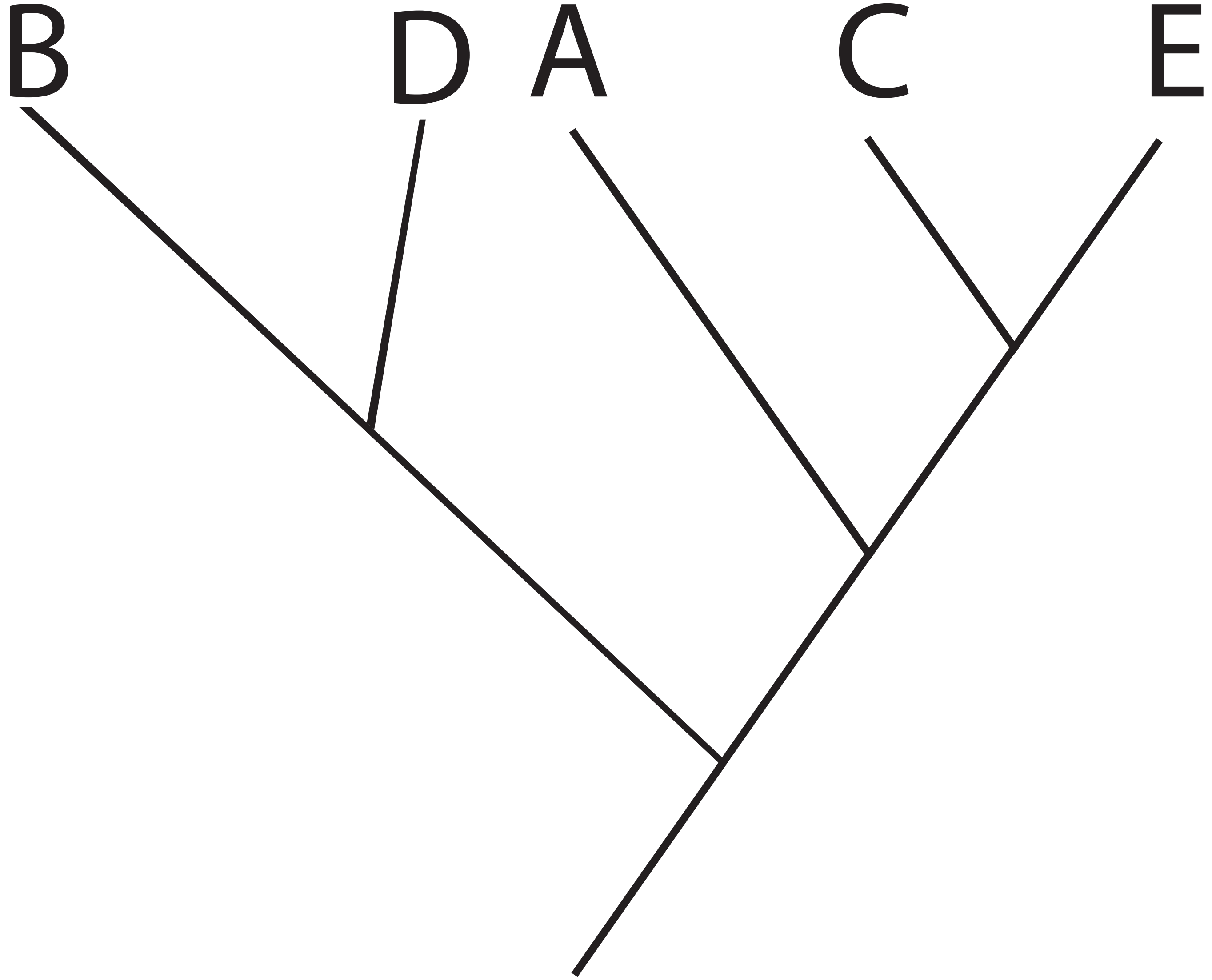
		V	D	E	V	G	G
	0	-5	-10	-15	●	●	●
V	-5	4	-1	●	●	●	●
E	-10	-1	●	●	●	●	●
V	-15	●	●	●	●	●	●
A	●	●	●	●	●	●	●
G	●	●	●	●	●	●	●
D	●	●	●	●	●	●	●

		V		D		E		V		G		G	
	0	←	-5	←	-10	←	-15	←	-20		●		●
V	↑	↖											
	-5		4	←	-1	←	-6		●		●		●
E	↑		↑	↖									
	-10		-1		6		●		●		●		●
V	↑	↖	↑										
	-15		-6		●		●		●		●		●
A	↑												
	-20		●		●		●		●		●		●
G													
	●		●		●		●		●		●		●
D													
	●		●		●		●		●		●		●

		V		D		E		V		G		G		
	0	←	-5	←	-10	←	-15	←	-20	←	-25		●	
V	↑	↖												
	-5		4	←	-1	←	-6	←	-11		●		●	
E	↑		↑	↖		↖								
	-10		-1		6		4		●		●		●	
V	↑	↖	↑		↑									
	-15		-6		1		●		●		●		●	
A	↑		↑											
	-20		-11		●		●		●		●		●	
G	↑													
	-25		●		●		●		●		●		●	
D														
	●		●		●		●		●		●		●	

		V	D	E	V	G	G	E	L	G	R
	0	← -5	← -10	← -15	← -20	← -25	← -30	← -35	← -40	← -45	← -50
V	↑ -5	↖ 4	← -1	← -6	↖ -11	← -16	← -21	← -26	← -31	← -36	← -41
E	↑ -10	↑ -1	↖ 6	↖ 4	← -1	← -6	← -11	↖ -16	← -21	← -26	← -31
V	↑ -15	↖ -6	↑ 1	↖ 4	↖ 8	← 3	← -2	← -7	← -12	← -17	← -22
A	↑ -20	↑ -11	↑ -4	↖ 0	↖ 4	↖ 8	↖ 3	← -2	← -7	↖ -12	← -17
G	↑ -25	↑ -16	↑ -9	↑ -5	↑ -1	↖ 10	↖ 14	← 9	← 4	↖ -1	← -6
D	↑ -30	↑ -21	↖ -10	↖ -7	↑ -6	↑ 5	↖ 9	↖ 16	← 11	← 6	← 1
L	↑ -35	↑ -26	↑ -15	↑ -12	↖ -6	↑ 0	↑ 4	↑ 11	↖ 20	← 15	← 10
G	↑ -40	↑ -31	↑ -20	↖ -17	↑ -11	↖ 0	↖ 6	↑ 6	↑ 15	↖ 26	← 21
R	↑ -45	↑ -36	↑ -25	↖ -20	↑ -16	↑ -5	↑ 1	↖ 6	↑ 10	↑ 21	↖ 31
L	↑ -50	↑ -41	↑ -30	↑ -25	↖ -19	↑ -10	↑ -4	↑ 1	↖ 10	↑ 16	↑ 26
L	↑ -55	↑ -46	↑ -35	↑ -30	↖ -24	↑ -15	↑ -9	↑ -4	↖ 5	↑ 11	↑ 21
I	↑ -60	↑ -51	↑ -40	↑ -35	↖ -27	↑ -20	↑ -14	↑ -9	↑ 0	↑ 6	↑ 16
V	↑	↖	↑	↑	↖	↑	↑	↑	↑	↑	↑

Aligning multiple sequences



Progressive alignment

Devised by Feng and Doolittle 1987 and Higgins and Sharp, 1988. An approximate method for producing multiple sequence alignments using a guide tree.

- Perform pairwise alignments to produce a distance matrix
- Produce a guide tree from the distances
- Use the guide tree to specify the ordering used for aligning sequences, closest to furthest.

```
A   PEEKSAVTALWGKVNVDEVGG
B   GEEKAAVLALWDKVNEEEVGG
C   PADKTNVKAAWGKVGAHAGEYGA
D   AADKTNVKAAWSKVGGHAGEYGA
E   EHEWQLVLHVWAKVEADVAGHGQ
```

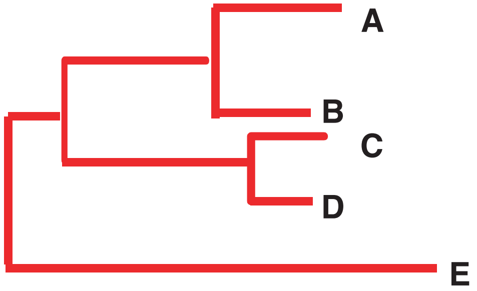
pairwise alignment →

A	–				
B	.17	–			
C	.59	.60	–		
D	.59	.59	.13	–	
E	.77	.77	.75	.75	–

↓ tree inference

```
A   PEEKSAVTALWGKVNVDEVGG
B   GEEKAAVLALWDKVNEEEVGG
C   PADKTNVKAAWGKVGAHAGEYGA
D   AADKTNVKAAWSKVGGHAGEYGA
E   EHEWQLVLHVWAKVEADVAGHGQ
```

+



↓ alignment stage

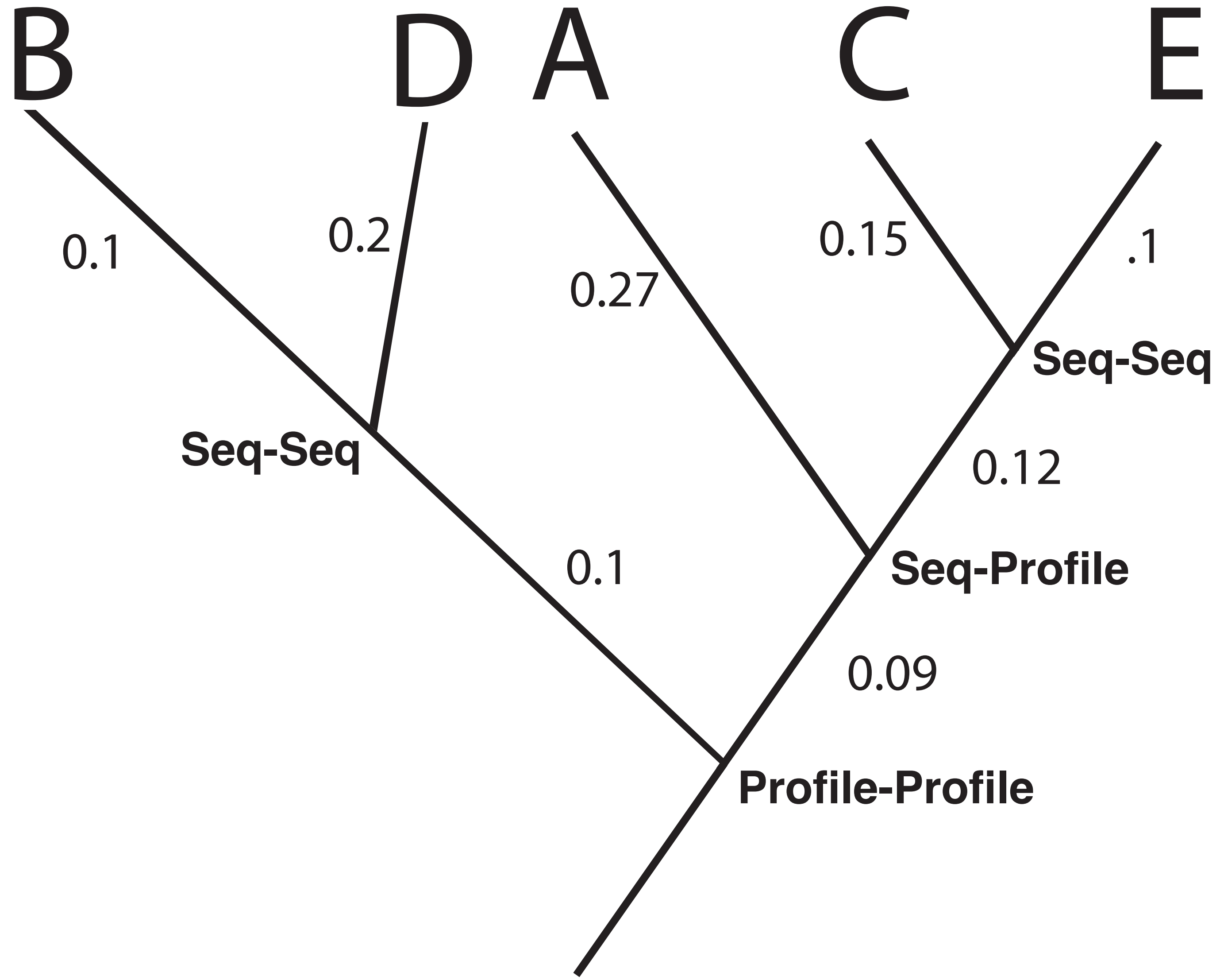
```
A   PEEKSAVTALWGKVN--VDEVGG
B   GEEKAAVLALWDKVN--EEEVGG
C   PADKTNVKAAWGKVGAHAGEYGA
D   AADKTNVKAAWSKVGGHAGEYGA
E   EHEWQLVLHVWAKVEADVAGHGQ
```

Alignment stage of progressive alignments

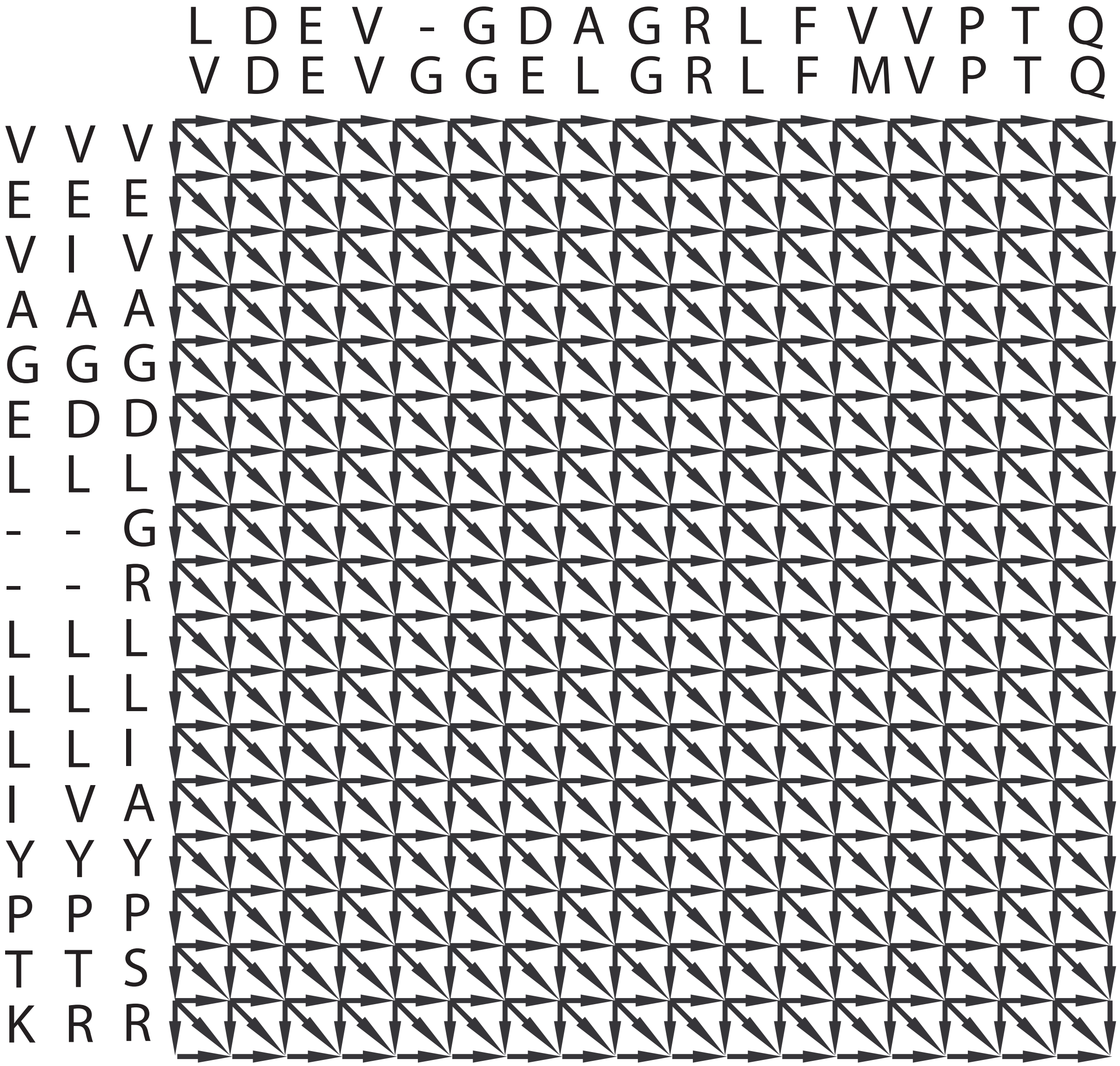
Sequences of clades become grouped into “profiles” as the algorithm descends the tree. The next youngest internal nodes is selected at each step to create a new profile. Alignment at each step involves

- Sequence-Sequence
- Sequence-Profile
- Profile-Profile

Aligning multiple sequences



Profile to Profile alignment


L D E V - G D A G R L F V V P T Q
V D E V G G E L G R L F M V P T Q
V E V A G E L - - L L L I Y P T K
V E I A G D L - - L L L V Y P T R
V E V A G D L G R L L I A Y P S R

Profile to profile alignments

Adding a gap to a profile means that every member of that group of sequences gets a gap at that position of the sequence.

Usually the scores for each edge in the Needleman-Wünsch graph are calculated using a “sum of pairs” scoring system.

`clustal W`[1] uses weights assigned to each sequence in a profile group to downweight closely related sequences so that they are not overrepresented.

[1] Thompson, Higgins, and Gibson. **Nuc. Acids. Res**. 1994

Profile 1

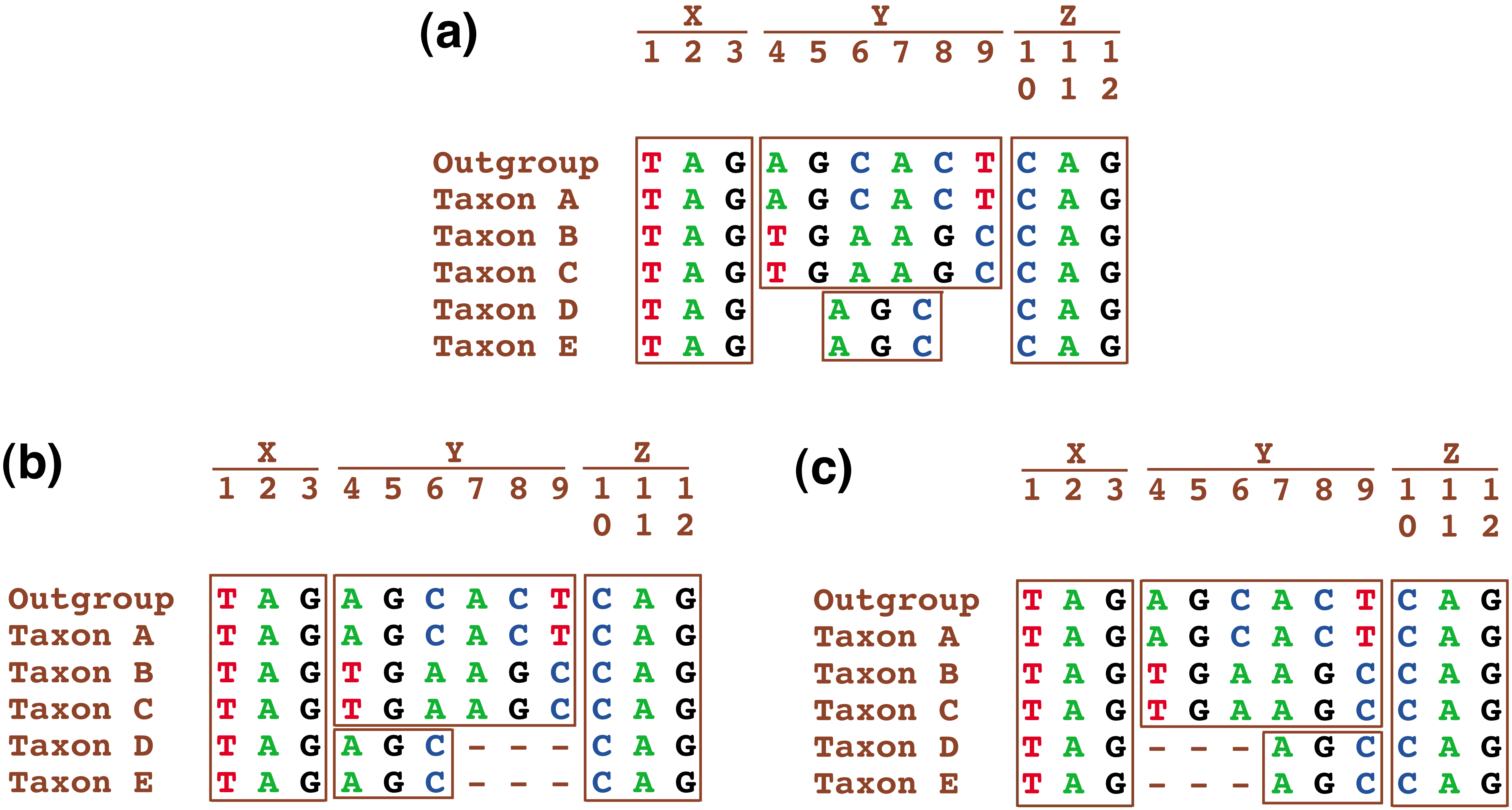
Seq	weight	AA
taxon A	0.3	V
taxon C	0.24	A
taxon E	0.19	I

Profile 2

Seq	weight	AA
taxon B	0.15	V
taxon D	0.25	M

$$\begin{aligned}
D_{P1,P2} &= \frac{\sum_i \sum_j w_i w_j d_{ij}}{n_i n_j} \\
&= \frac{1}{6}\left[d(V,V)w_A w_B + d(V,M)w_A w_D + d(A,V)w_C w_B \ldots\right. \\
&= \left.\ldots d(A,M)w_C w_D + d(I,V)w_E w_B + d(I,M)w_E w_D\right] \\
&= \frac{1}{6}(\mathbf{4} \times 0.3 \times 0.15 + \mathbf{1} \times 0.3 \times 0.25 + \mathbf{0} \times 0.24 \times 0.15 \ldots \\
&= \ldots -1 \times 0.24 \times 0.25 + \mathbf{3} \times 0.19 \times 0.15 + \mathbf{1} \times 0.19 \times 0.1 \\
&= 1.46225
\end{aligned}$$

Dealing with alignment ambiguity[2]



```
               X           Y             Z
            -----  -----------------  -----
            1 2 3  4 5 6 7 8 9        1 1 1
                                      0 1 2

Outgroup    T A G  A G C A C T        C A G
Taxon A     T A G  A G C A C T        C A G
Taxon B     T A G  T G A A G C        C A G
Taxon C     T A G  T G A A G C        C A G
Taxon D     T A G      A G C          C A G
Taxon E     T A G      A G C          C A G
```

```
               X           Y             Z
            -----  -----------------  -----
            1 2 3  4 5 6 7 8 9        1 1 1
                                      0 1 2

Outgroup    T A G  A G C A C T        C A G
Taxon A     T A G  A G C A C T        C A G
Taxon B     T A G  T G A A G C        C A G
Taxon C     T A G  T G A A G C        C A G
Taxon D     T A G  A G C – – –        C A G
Taxon E     T A G  A G C – – –        C A G
```

```
               X           Y             Z
            -----  -----------------  -----
            1 2 3  4 5 6 7 8 9        1 1 1
                                      0 1 2

Outgroup    T A G  A G C A C T        C A G
Taxon A     T A G  A G C A C T        C A G
Taxon B     T A G  T G A A G C        C A G
Taxon C     T A G  T G A A G C        C A G
Taxon D     T A G  – – – A G C        C A G
Taxon E     T A G  – – – A G C        C A G
```

[2] from M. S. Y. Lee, *TREE*, 2001

Dealing with alignment ambiguity[3] - deletion

(a)

	X			Y						Z		
	1	2	3	4	5	6	7	8	9	10	11	12
Outgroup	T	A	G	A	G	C	A	C	T	C	A	G
Taxon A	T	A	G	A	G	C	A	C	T	C	A	G
Taxon B	T	A	G	T	G	A	A	G	C	C	A	G
Taxon C	T	A	G	T	G	A	A	G	C	C	A	G
Taxon D	T	A	G		A	G	C			C	A	G
Taxon E	T	A	G		A	G	C			C	A	G

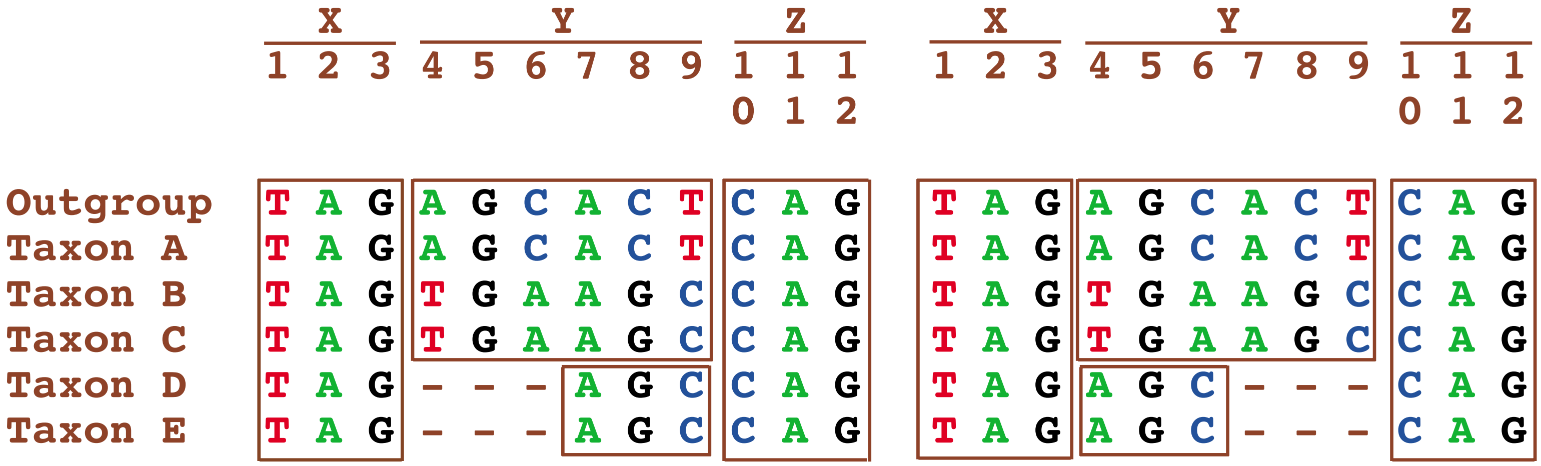
	X			Z		
	1	2	3	10	11	12
Outgroup	T	A	G	C	A	G
Taxon A	T	A	G	C	A	G
Taxon B	T	A	G	C	A	G
Taxon C	T	A	G	C	A	G
Taxon D	T	A	G	C	A	G
Taxon E	T	A	G	C	A	G

	X			Y						Z		
	1	2	3	4	5	6	7	8	9	10	11	12
Outgroup	T	A	G	A	G	C	A	C	T	C	A	G
Taxon A	T	A	G	A	G	C	A	C	T	C	A	G
Taxon B	T	A	G	T	G	A	A	G	C	C	A	G
Taxon C	T	A	G	T	G	A	A	G	C	C	A	G
Taxon D	T	A	G	?	?	?	-	-	-	C	A	G
Taxon E	T	A	G	?	?	?	-	-	-	C	A	G

[3] from M. S. Y. Lee, *TREE*, 2001

Dealing with alignment ambiguity[4]

Elision method (Wheeler, 1995) involves simply concatenating matrices.



	X			Y						Z		
	1	2	3	4	5	6	7	8	9	10	11	12
Outgroup	T	A	G	A	G	C	A	C	T	C	A	G
Taxon A	T	A	G	A	G	C	A	C	T	C	A	G
Taxon B	T	A	G	T	G	A	A	G	C	C	A	G
Taxon C	T	A	G	T	G	A	A	G	C	C	A	G
Taxon D	T	A	G	–	–	–	A	G	C	C	A	G
Taxon E	T	A	G	–	–	–	A	G	C	C	A	G

	X			Y						Z		
	1	2	3	4	5	6	7	8	9	10	11	12
Outgroup	T	A	G	A	G	C	A	C	T	C	A	G
Taxon A	T	A	G	A	G	C	A	C	T	C	A	G
Taxon B	T	A	G	T	G	A	A	G	C	C	A	G
Taxon C	T	A	G	T	G	A	A	G	C	C	A	G
Taxon D	T	A	G	A	G	C	–	–	–	C	A	G
Taxon E	T	A	G	A	G	C	–	–	–	C	A	G

[4] from M. S. Y. Lee, *TREE*, 2001

Simultaneous tree inference and alignment

- Ideally we would address uncertainty in both types of inference at the same time

- Allows for application of statistical models to improve inference and assessments of reliability

- Just now becoming feasible: `POY` (Wheeler, Gladstein, Laet, 2002), `Handel` (Holmes and Bruno, 2001), `BAliPhy` (Redelings and Suchard, 2005), and `BEAST`(Lunter *et al.*, 2005, Drummond and Rambaut, 2003). SATe (Liu *et al* 2009; Yu and Holder software).

References

Edwards, S. V., Liu, L., and Pearl, D. K. (2007). High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences*, 104(14):5936–5941.

Liu, L. and Pearl, D. K. (2007). Species trees from gene trees: reconstruction Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic Biology*, 56(3):504–514.