Thanks to Paul Lewis, Joe Felsenstein, Peter Beerli, and Tandy Warnow for slides.

## Multiple Sequence Alignment (MSA) - 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
- Most MSA programs use progressive alignment:
  - this reduces MSA to a series of pairwise operations.
  - these algorithms are heuristic.
  - the criteria used are not ideal from an evolutionary standpoint (and this has implications for tree inference).

- Simultaneous inference of MSA and tree is the most appropriate choice, but is computationally demanding (see Bali-Phy, Handel, and AliFritz software)
- Iterative "meta-algorithms" (e.g. SATé ) consider multiple alignments during a tree inference and can handle thousands of sequences.
- GUIDANCE (and similar tools) cull ambiguously aligned regions to lower the chance that misalignment leads to errors in downstream analyses.

human KRSV
chimp KRV
gorilla KSV
orang KPRV

How should we align these sequences?

| | | |
|---|---|---|
| human KRSV | human KRSV | human KRSV |
| chimp KR-V | chimp K-RV | chimp KR-V |
| gorilla KS-V | gorilla K-SV | gorilla K-SV |
| orang KPRV | orang KPRV | 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 substitutions between similar residues (for example Leucine→Isoleucine).
- infrequent types of substitutions receive negative scores
- indels are rare, so gaps are heavily penalized (negative scores).

# 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*. $(k)$ is just an index for the column.

$$D_{ij} = 12$$

## Scoring an alignment with gaps

If we were to use a gap penalty of -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-DLGRLFVVPTQ
```

*Affine gap penalties* are often used to accommodate multi-site indels:

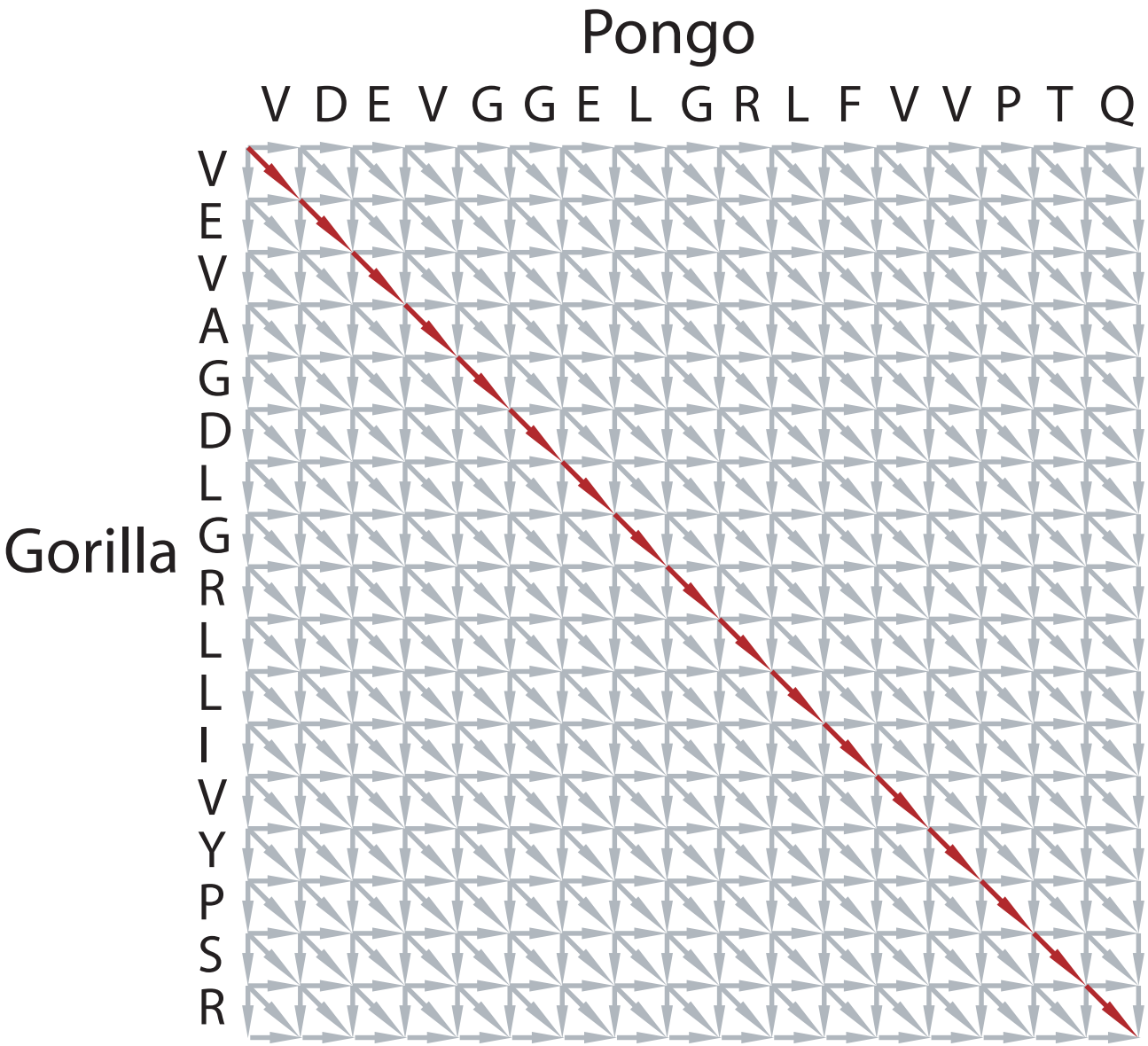
$$\mathsf{GP} = \mathsf{GO} + (l)\mathsf{GE}$$

where:

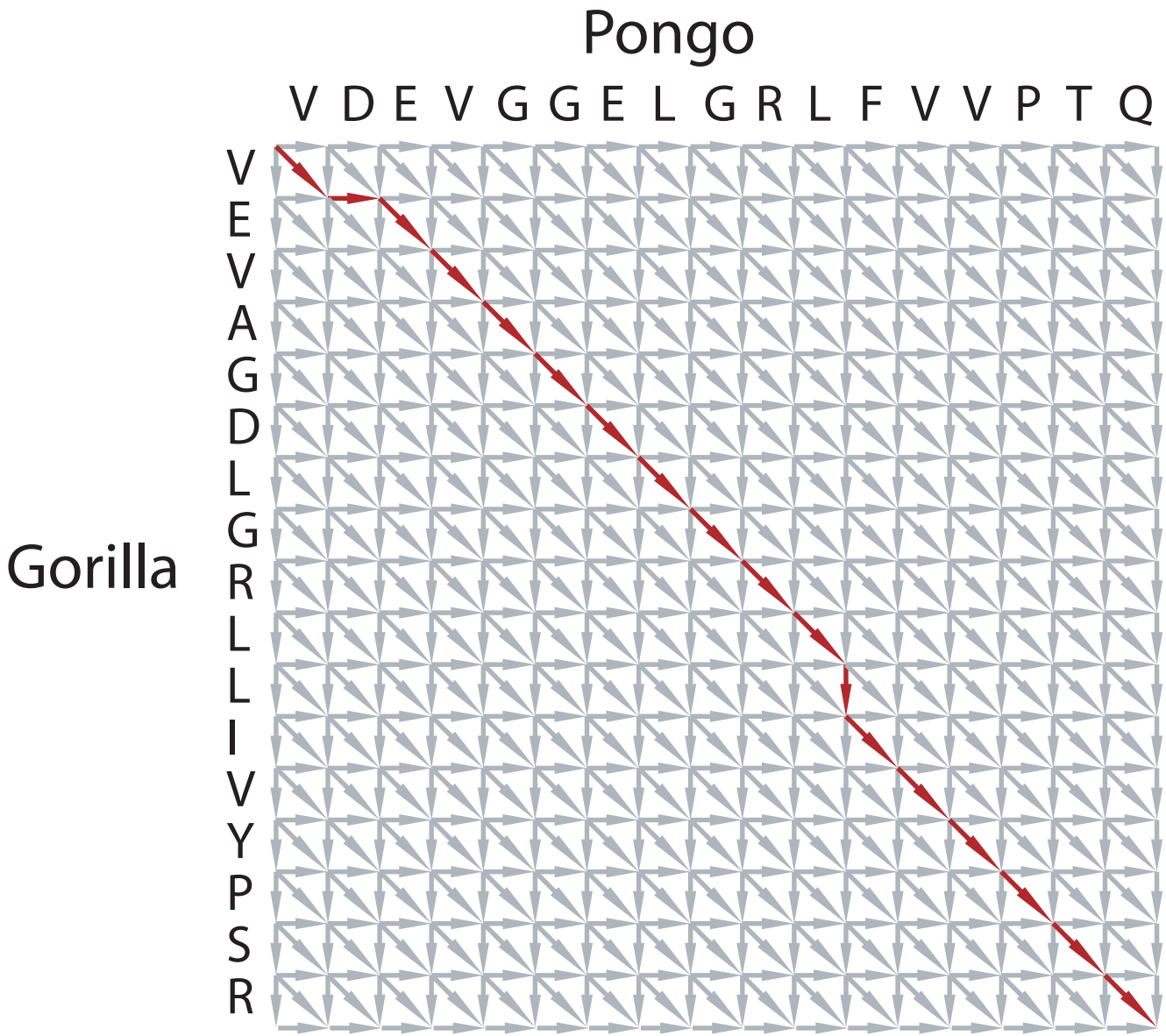
- GP is the gap penalty.
- GO is the "gap-opening penalty"
- GE is the "gap-extension penalty"
- $l$ is the length of the gap

# Finding an optimal alignment

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *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 |
| **Score** | 4 | -8 | 5 | 5 | 0 | 6 | 2 | 4 | 6 | 5 | 4 | -8 | 0 | 4 | -1 | 7 | 4 | 1 |



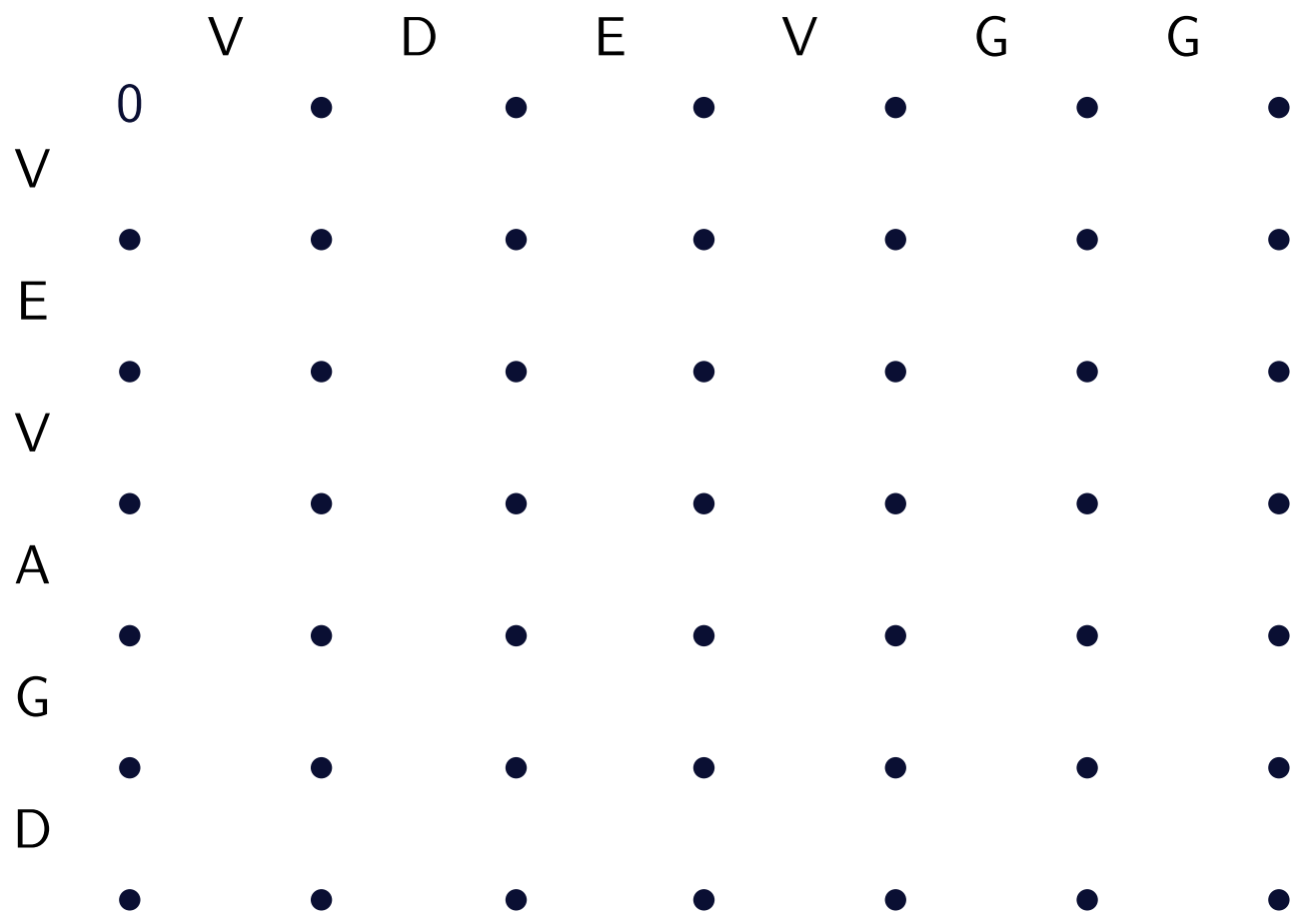
## Needleman-Wunsch algorithm (paraphrased)

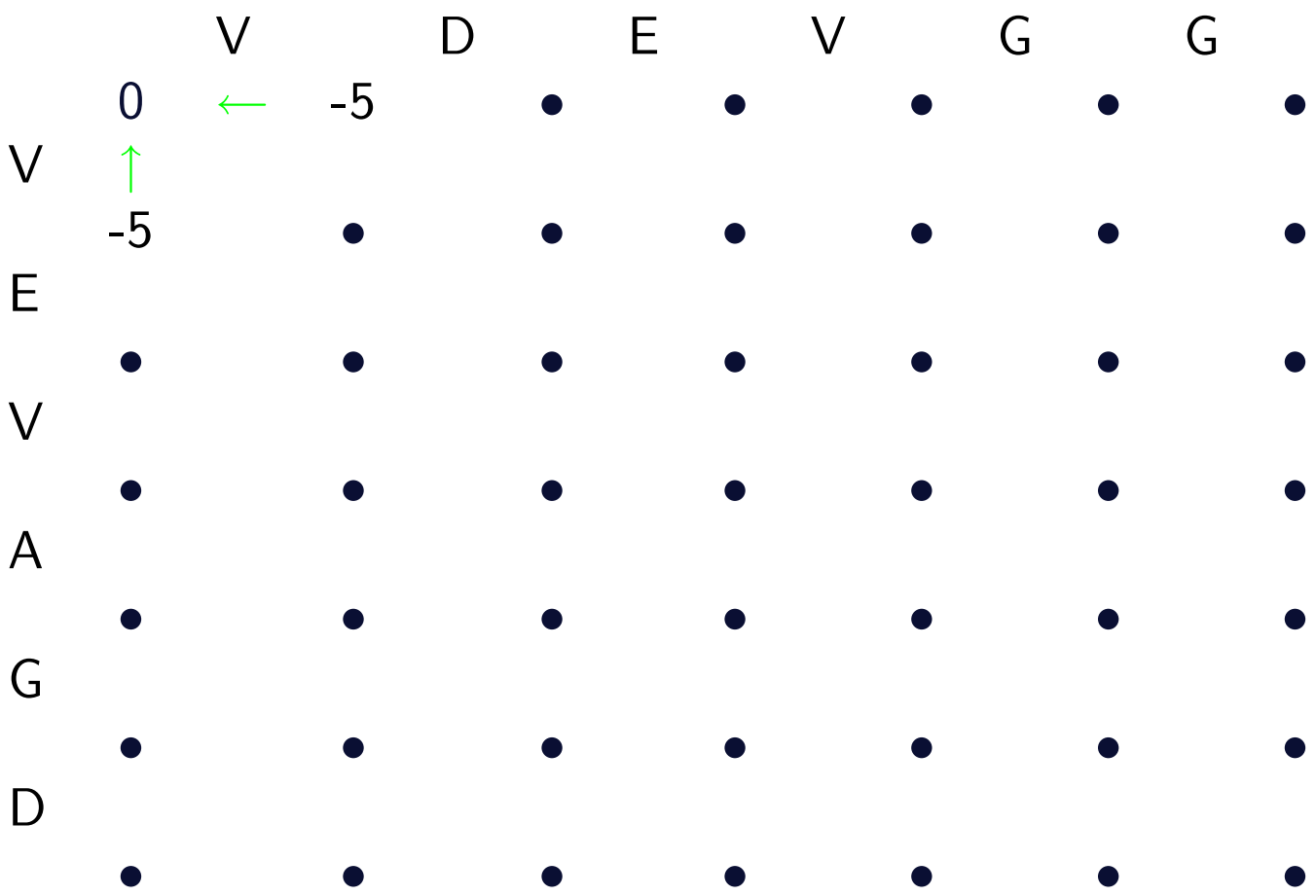
- Work from the top left (beginning of both sequences)
- For each cell store the highest score possible for that cell and a "back" pointer to tell point to the previous step in the best path
- When you reach the lower right corner, you know the optimal score and the back pointers tell you the alignment.

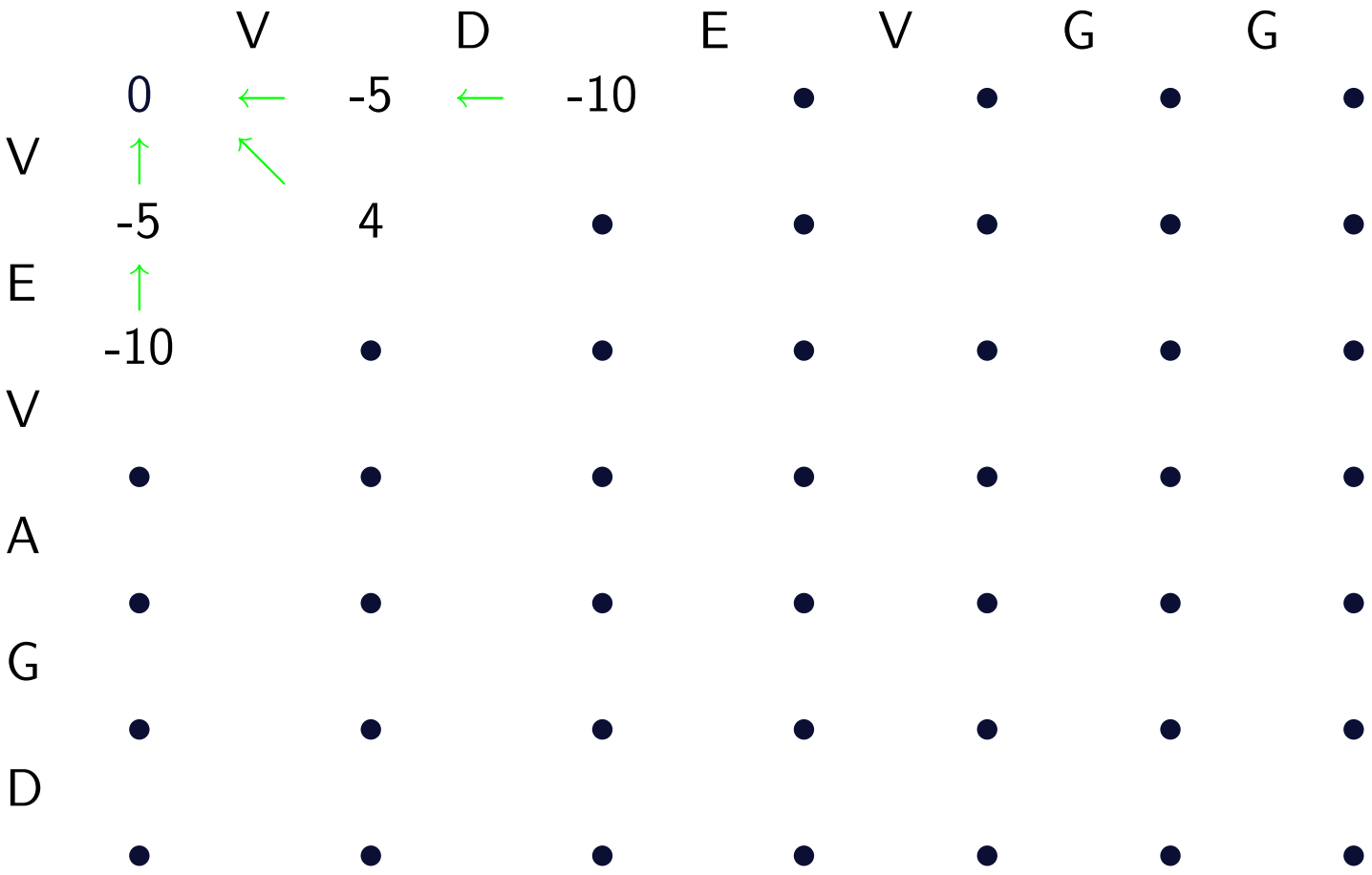
The highest-score calculation at each cell only depends on its the cell's three possible previous neighbors.

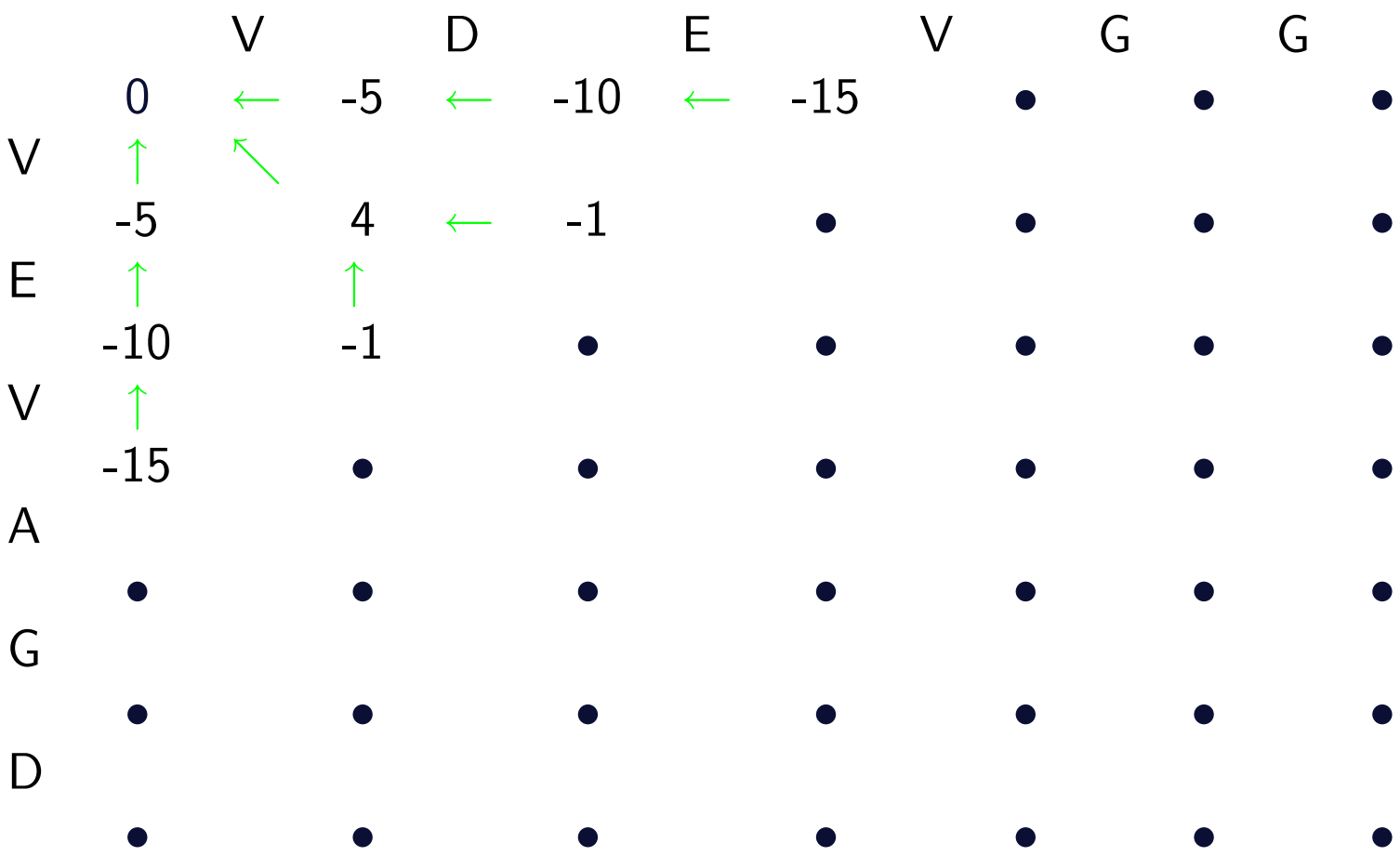
If one sequence is length $M_1$, and the other is length $M_2$, then Needleman-Wunsch only takes $\mathcal{O}(M_1 M_2)$ calculations.

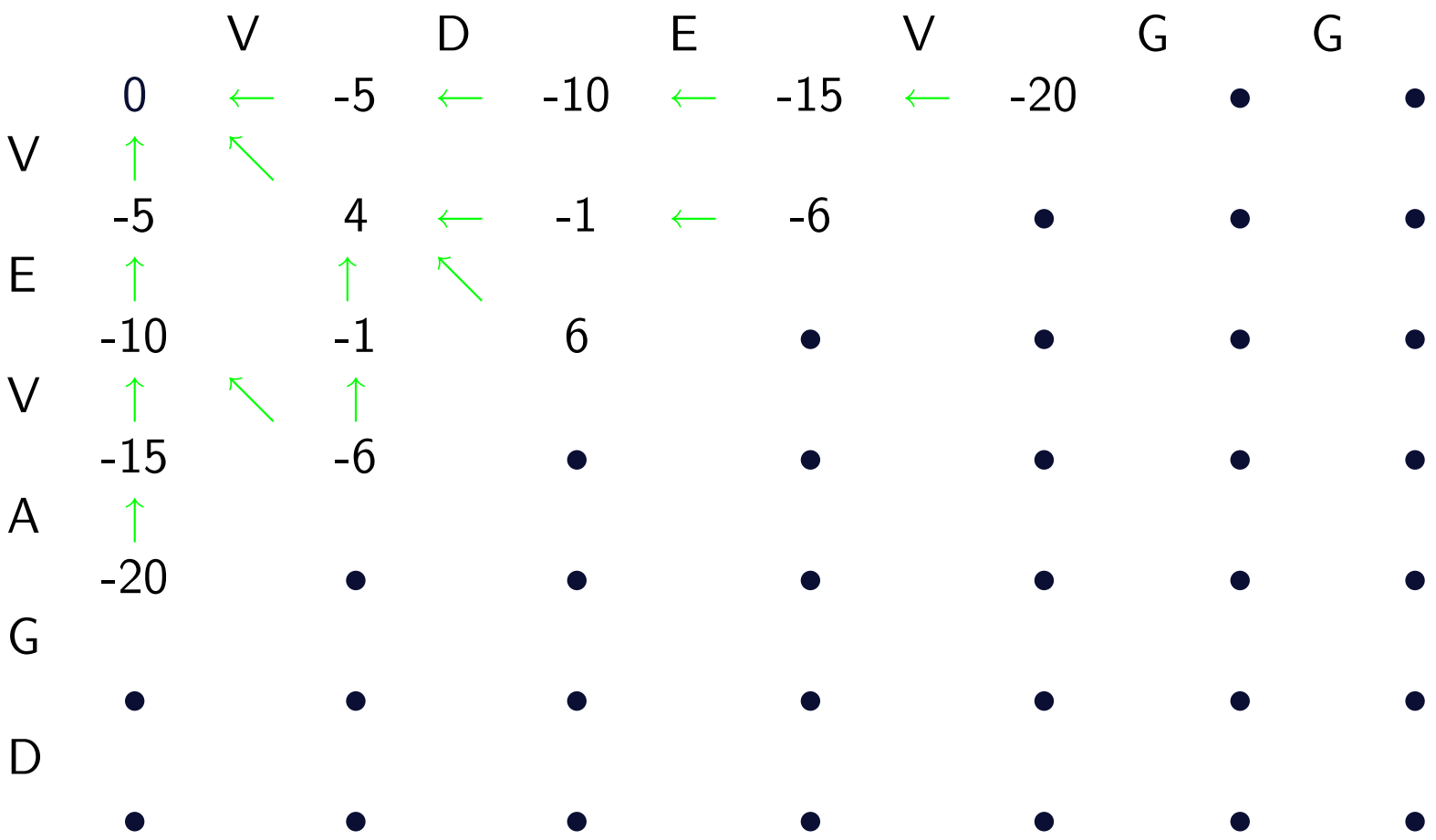
But there are a **much** larger number of possible alignments (# alignments grows exponentially).

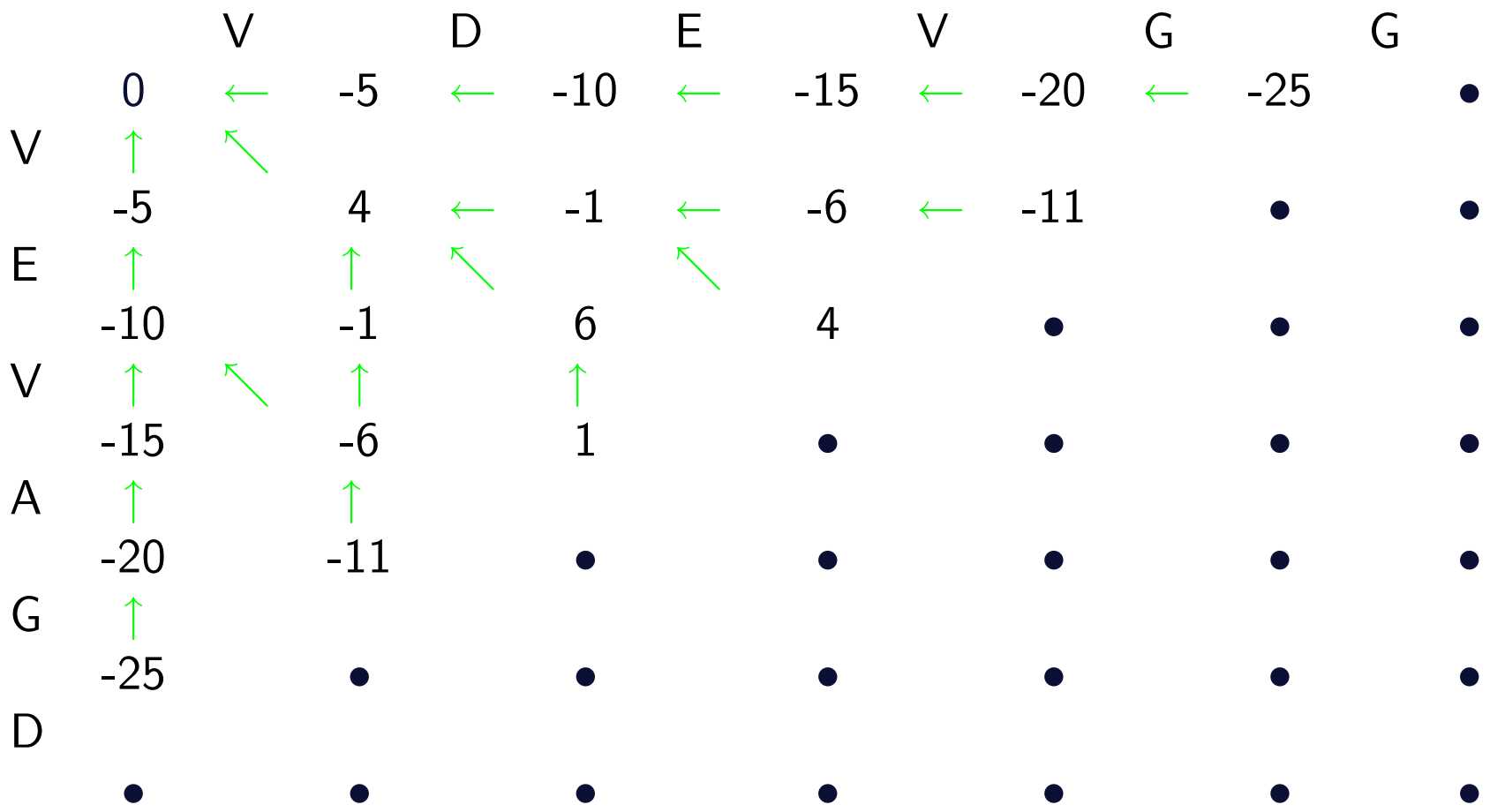

V D E V G G
0
V
E
V
A
G
D

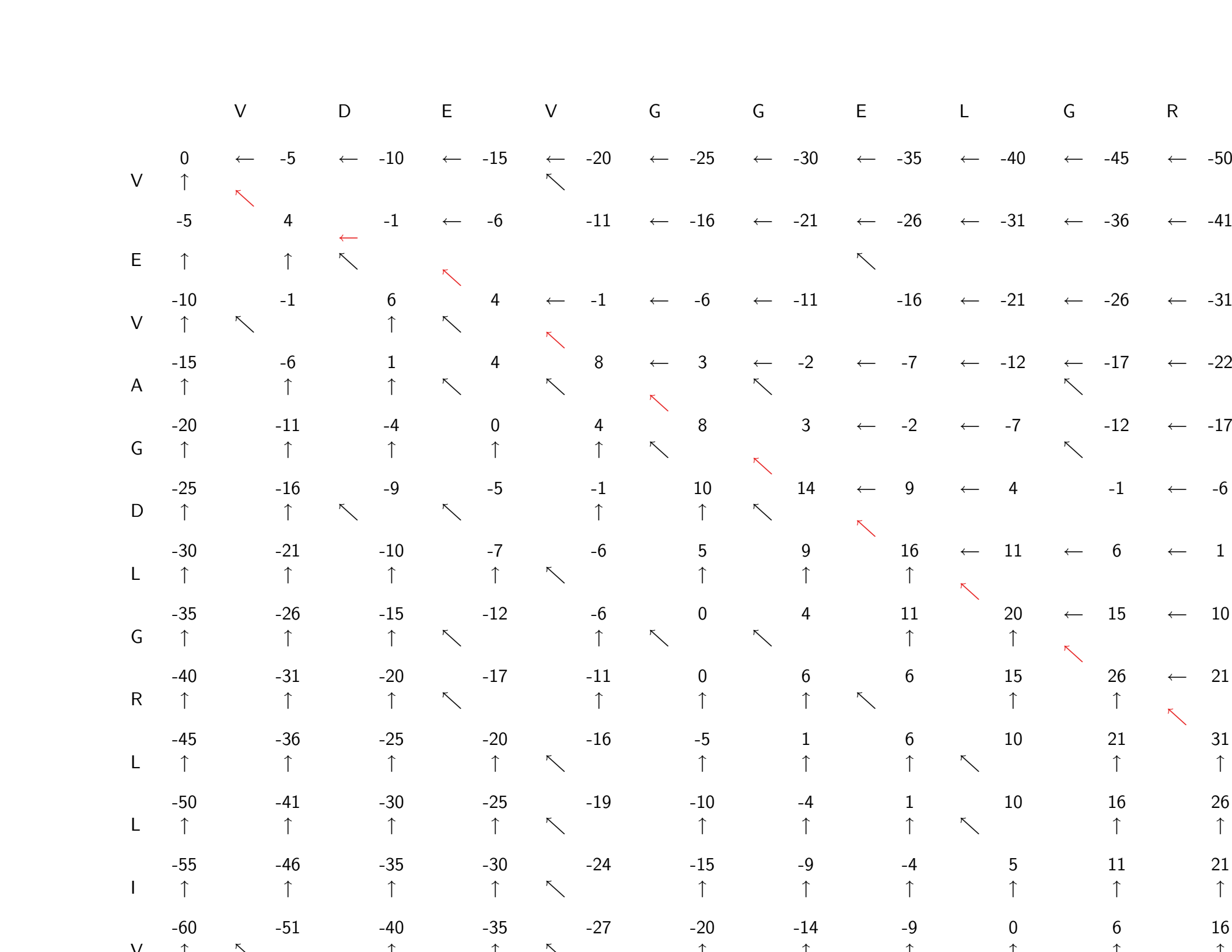

V D E V G G
0 ← -5
V ↑
-5
E
V
A
G
D


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 ↑ ↖ ↑ ↑ ↖ ↑ ↑ ↑ ↑ ↑ ↑

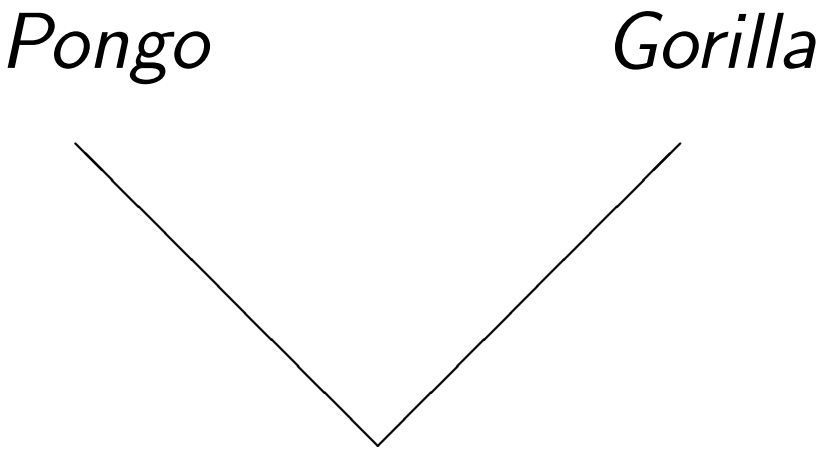
## Pairwise alignment is a beautiful topic in bioinformatics

- Clever programming tricks let us find the best-scoring alignment quickly (in $\mathcal{O}(M_1 M_2)$ number of computations, despite the fact that the number of alignments increases exponentially with $M$)
- The additive scoring system:
  - can incorporate biological knowledge (via empirically-based substitution matrices)
  - can be justified in terms of powerful statistical methodology (maximum likelihood).

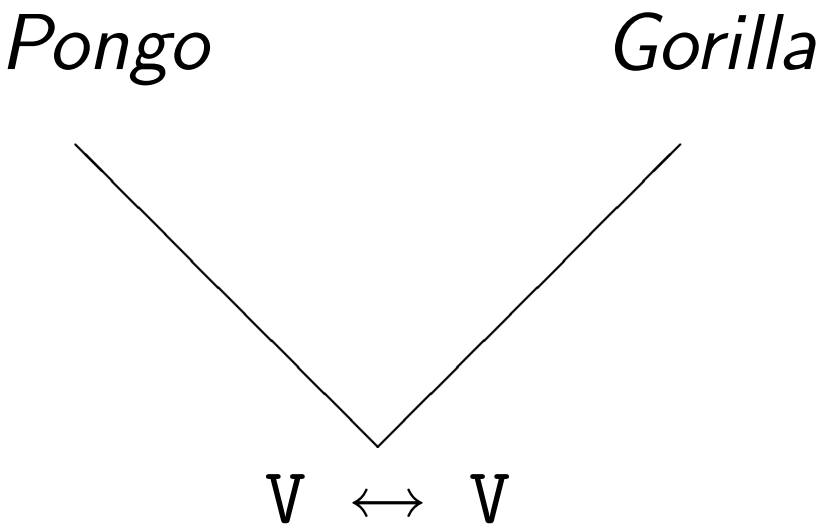
## Pairwise alignment costs

- Paul Lewis will explain likelihood tomorrow,
- Additive costs can be justified as approximations to the log of likelihoods if:
  - we can identify the events that must have occurred in generate the data, and
  - we can assign (relative) probabilities based on whether these events are rare or common.

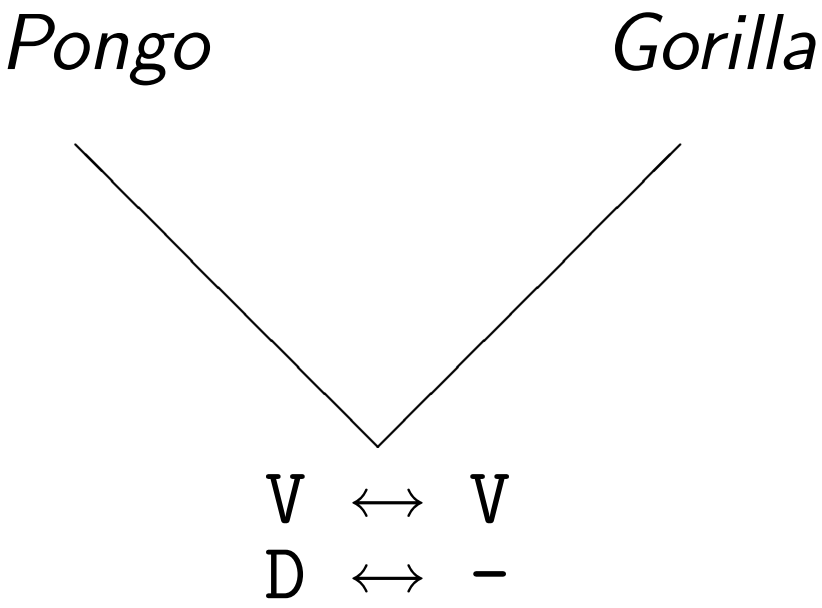
| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Pongo* | V | D | E | V | G | G | E | L | G | R | L | - | F | V | V | P | T | Q |
| *Gorilla* | V | - | E | V | A | G | D | L | G | R | L | L | I | V | Y | P | S | R |
| **Score** | 4 | -8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



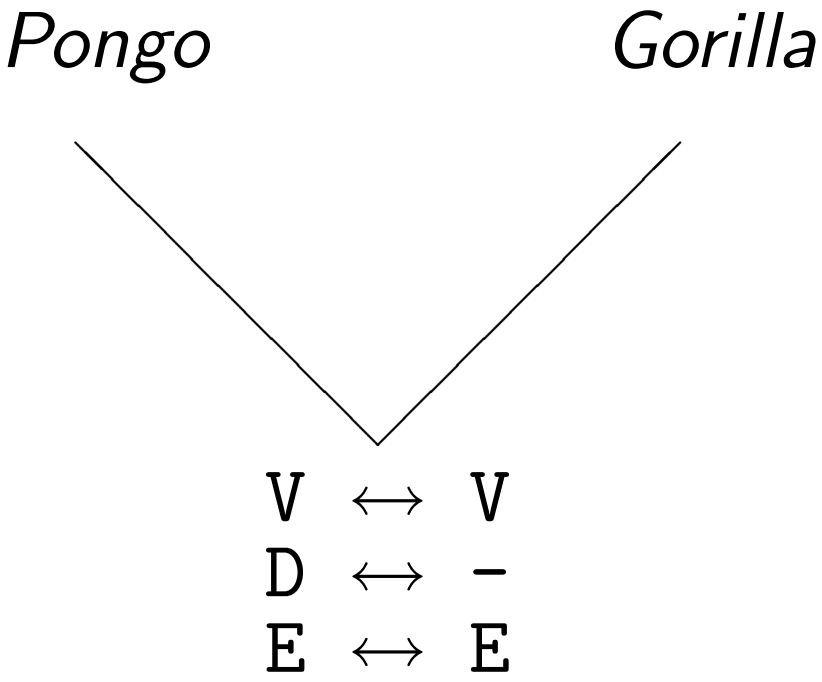
| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Pongo* | **V** | D | E | V | G | G | E | L | G | R | L | - | F | V | V | P | T | Q |
| *Gorilla* | **V** | - | E | V | A | G | D | L | G | R | L | L | I | V | Y | P | S | R |
| **Score**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



| *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 |



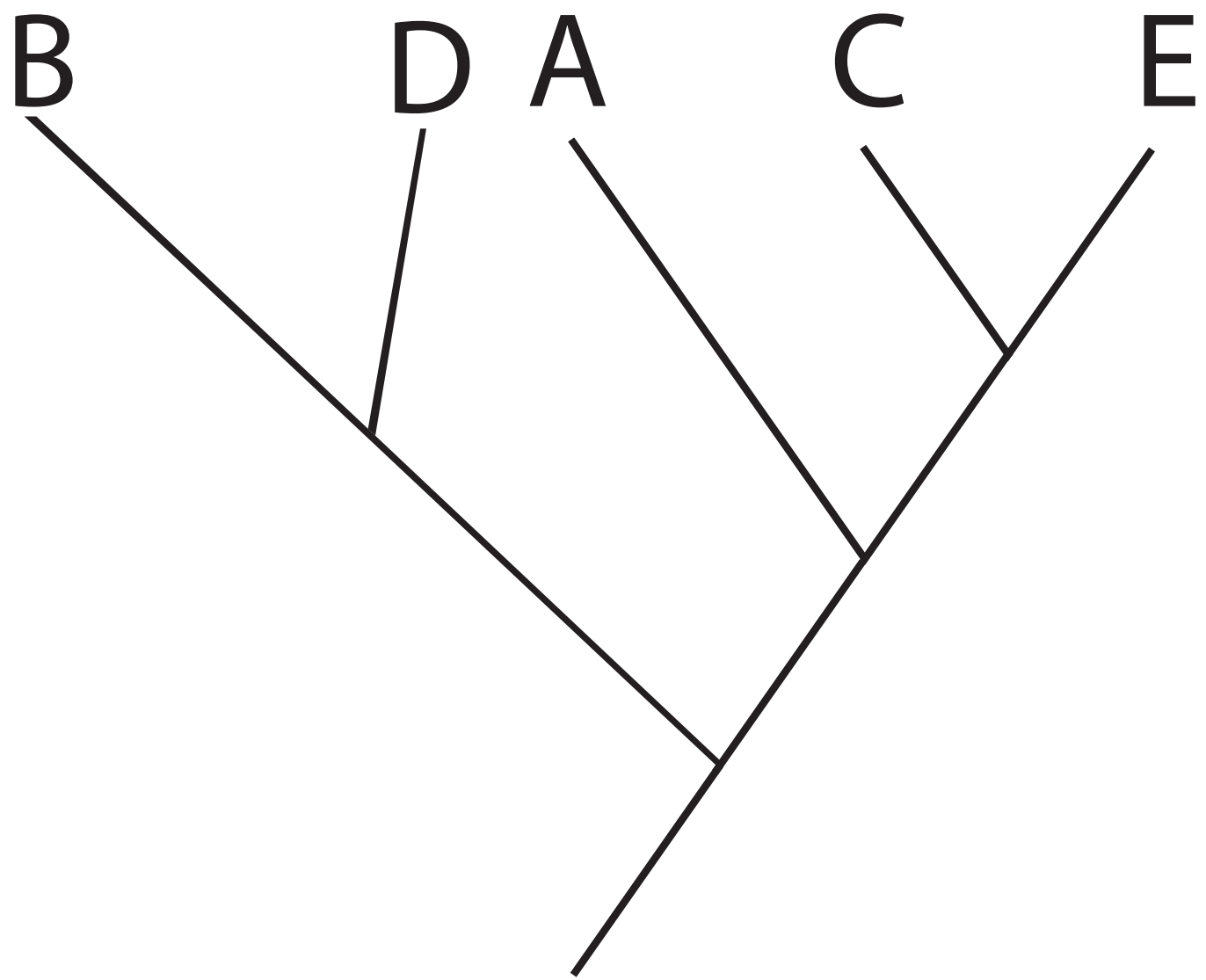
| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Pongo* | **V** | **D** | **E** | V | G | G | E | L | G | R | L | - | F | V | V | P | T | Q |
| *Gorilla* | **V** | - | **E** | V | A | G | D | L | G | R | L | L | I | V | Y | P | S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



## Multiple sequence alignment is an ugly topic in bioinformatics

- Clever programming tricks help, but we still have to rely on *heuristics* – approaches that provide good solutions, but are not guaranteed to find the best solution.
- The additive scoring system suffers from the fact that we do not observe ancestral sequences.

## Aligning more than two sequences

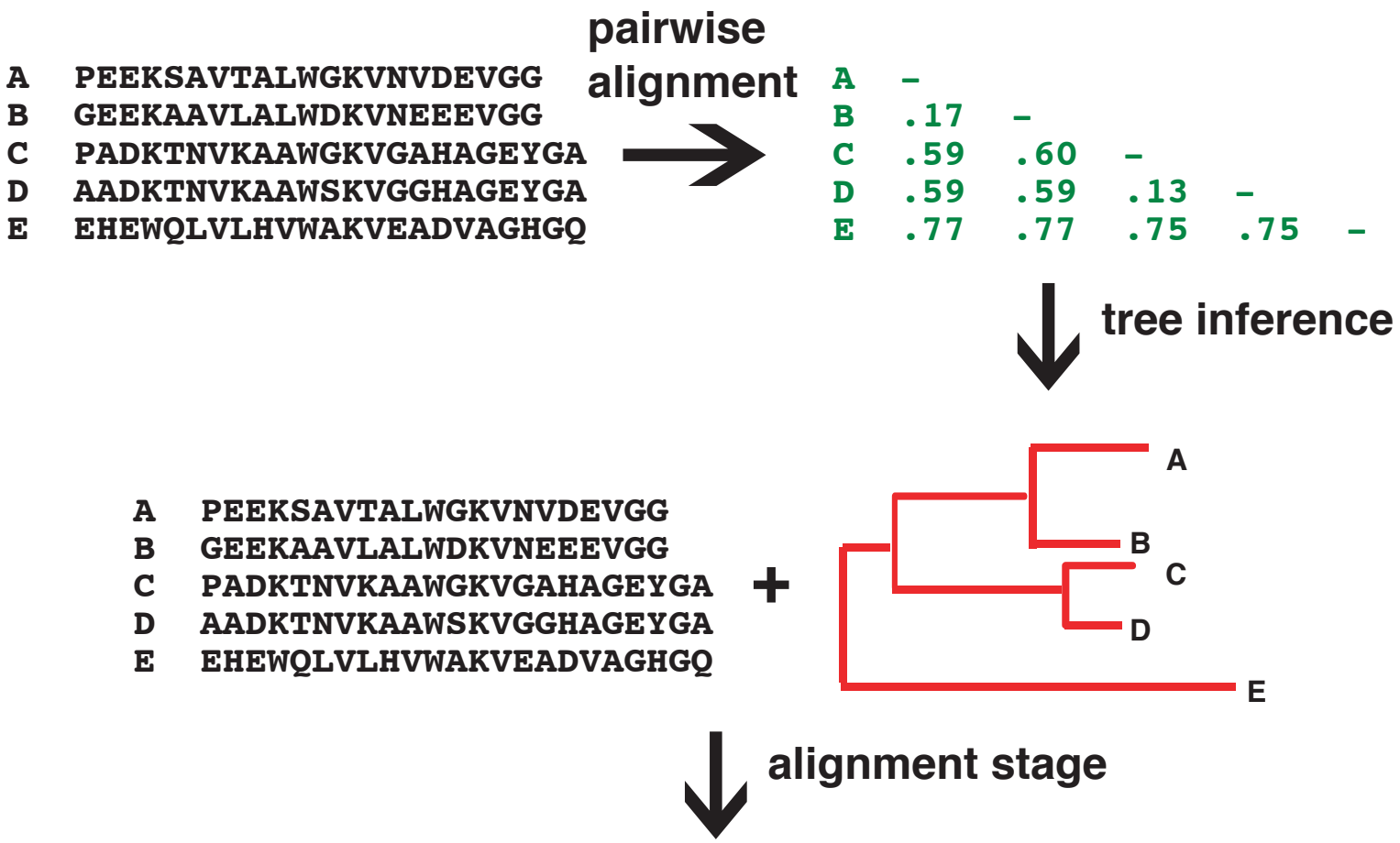


## Progressive alignment

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.

Feng and Doolittle 1987 and Higgins and Sharp, 1988

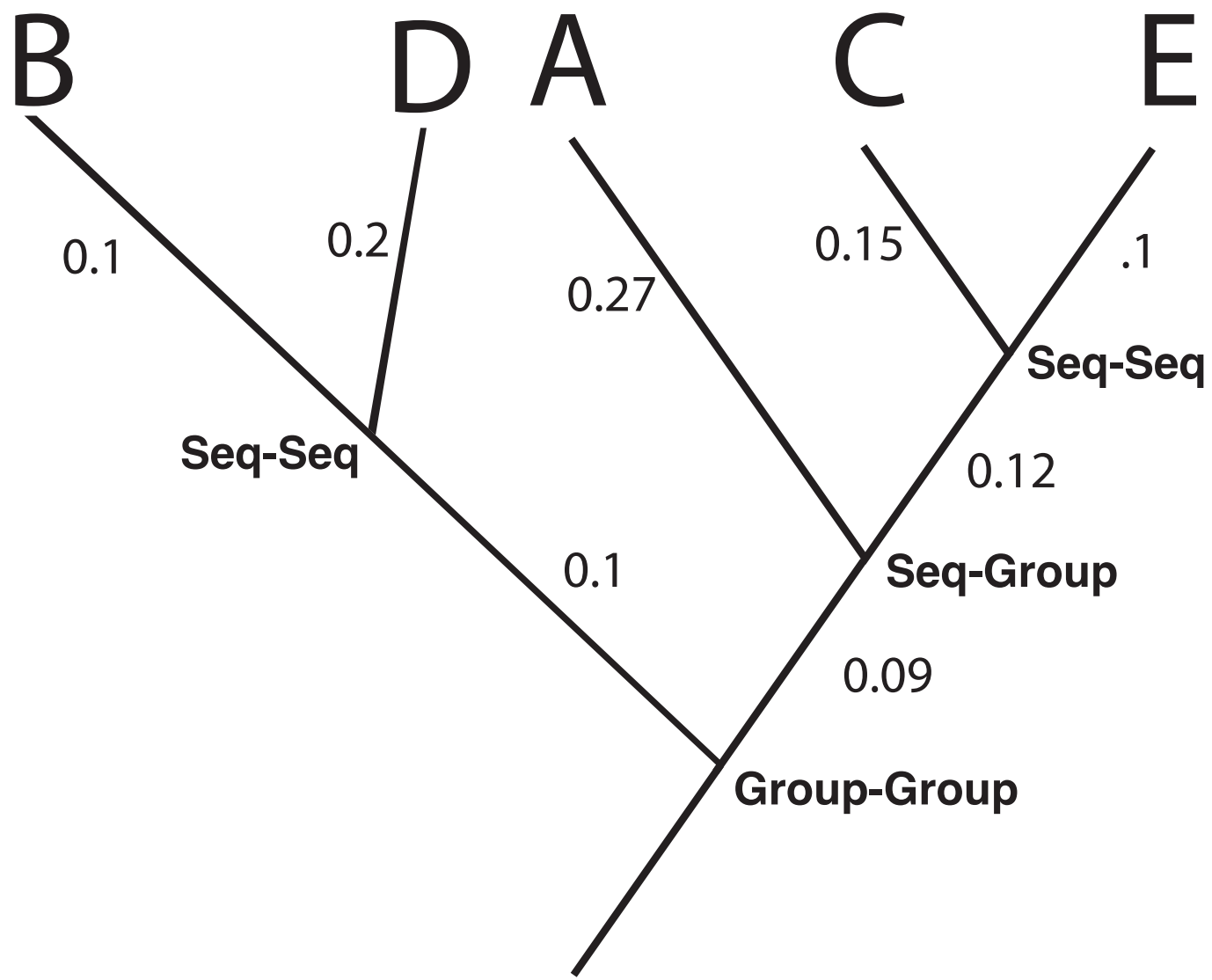

pairwise
alignment
A PEEKSAVTALWGKVNVDEVGG
B GEEKAAVLALWDKVNEEEVGG
C PADKTNVKAAWGKVGAHAGEYGA
D AADKTNVKAAWSKVGGHAGEYGA
E EHEWQLVLHVWAKVEADVAGHGQ
A -
B .17 -
C .59 .60 -
D .59 .59 .13 -
E .77 .77 .75 .75 -
tree inference
A PEEKSAVTALWGKVNVDEVGG
B GEEKAAVLALWDKVNEEEVGG
C PADKTNVKAAWGKVGAHAGEYGA
D AADKTNVKAAWSKVGGHAGEYGA
E EHEWQLVLHVWAKVEADVAGHGQ
+
A
B
C
D
E
alignment stage
A PEEKSAVTALWGKVN--VDEVGG
B GEEKAAVLALWDKVN--EEEVGG
C PADKTNVKAAWGKVGAHAGEYGA
D AADKTNVKAAWSKVGGHAGEYGA
E EHEWQLVLHVWAKVEADVAGHGQ

## Alignment stage of progressive alignments

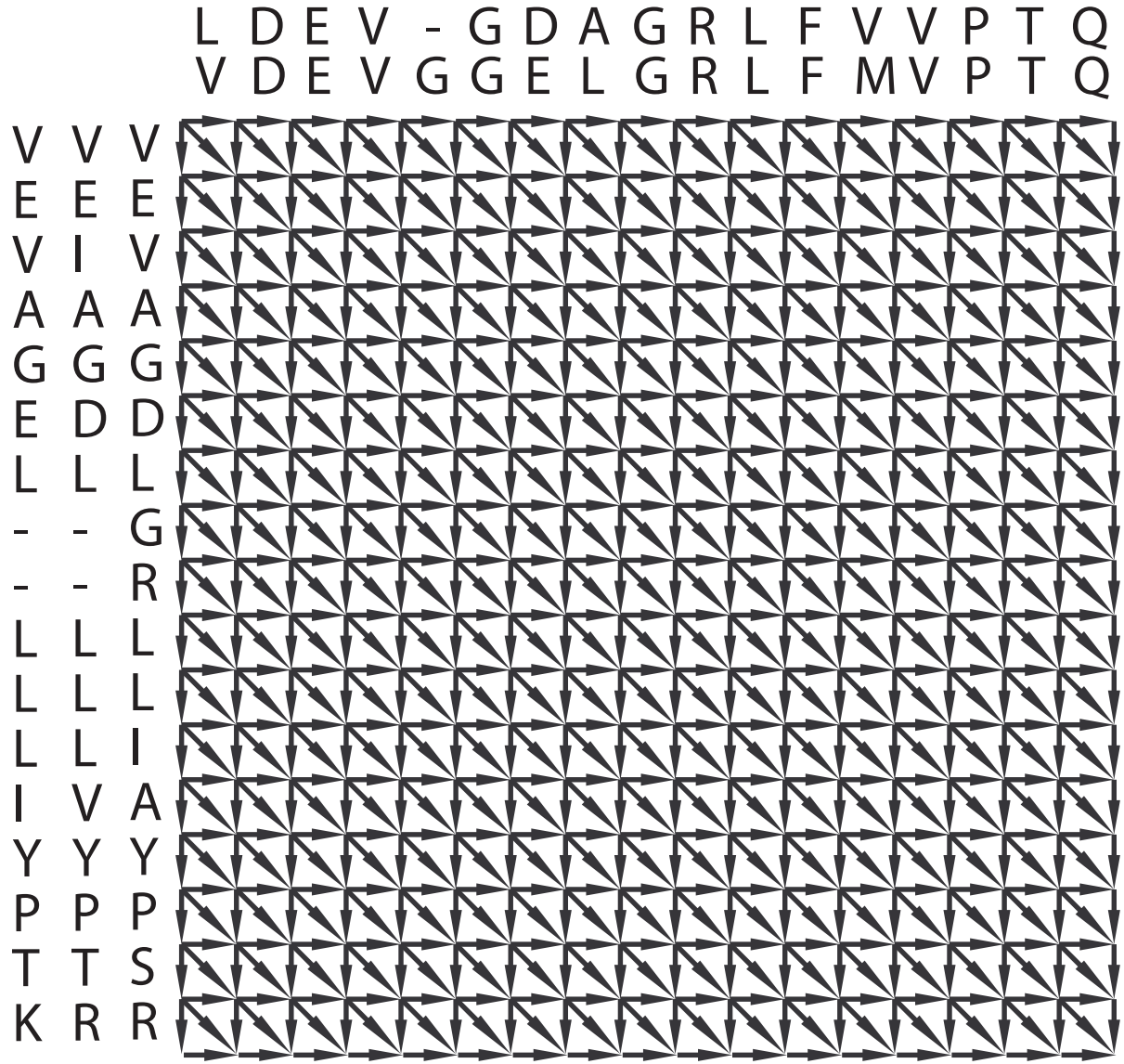
Sequences of clades become grouped as the algorithm descends the tree. Alignment at each step involves

- Sequence-Sequence,
- Sequence-Group, or
- Group-Group

# Aligning multiple sequences



# Group-to-Group alignment



## Group-to-group alignments

Adding a gap to a group means that every member of that group gets a gap at that position.

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

Many tools[1] uses weights assigned to each sequence in a group to down-weight closely related sequences so that they are not overrepresented - this is a weighted sum-of-pair scoring system.

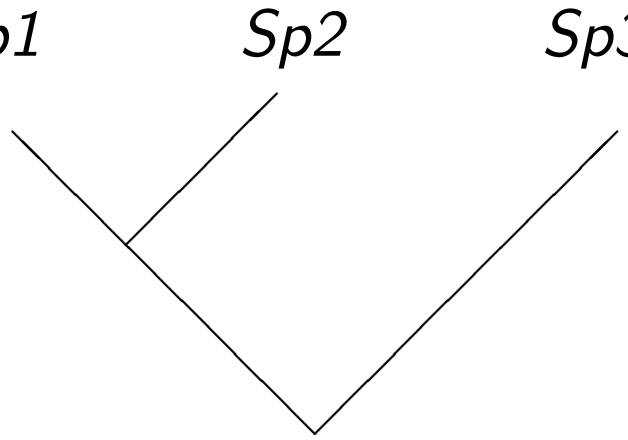
[1] e.g. Clustal and MAFFT

## Greedy choices leading to failure to find the best alignment

Consider the scoring scheme:

match = 0  mismatch = -3  gap = -7

Guide Tree:




Sequences:

| | |
|---|---|
| *Sp1* | GACCGTG |
| *Sp2* | GCCGTAG |
| *Sp3* | GACCGTAG |

## Greedy choices leading to failure to find the best alignment

match = 0  mismatch = -3  gap = -7

ungapped1*vs*2

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Sp1* | G | A | C | C | G | T | G | | |
| *Sp2* | G | C | C | G | T | A | G | | |
| **Score** | 0 | -3 | 0 | -3 | -3 | -3 | 0 | **Total=** | -12 |

would be preferred over gapped1*vs*2:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Sp1* | G | A | C | C | G | T | - | G | |
| *Sp2* | G | - | C | C | G | T | A | G | |
| **Score** | 0 | -7 | 0 | 0 | 0 | 0 | -7 | 0 | **Total=** -14 |

Adding a *Sp3* to ungapped1*vs*2:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Sp1* | G | - | A | C | C | G | T | G |
| *Sp2* | G | - | C | C | G | T | A | G |
| *Sp3* | G | A | C | C | G | T | A | G |

This implies 1 indel, and 4 substitutions. Score = -19

**If** we had been able to use gapped1*vs*2 then we could have:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Sp1* | G | A | C | C | G | T | - | G |
| *Sp2* | G | - | C | C | G | T | A | G |
| *Sp3* | G | A | C | C | G | T | A | G |

score = -14.

## Weighted sum of pairs scoring system

Assigns a score for a group-to-group by averaging all of the implied pairwise alignments.

| Group 1 | | |
|---|---|---|
| Seq | weight | AA |
| taxon A | 0.3 | V |
| taxon C | 0.24 | A |
| taxon E | 0.19 | I |

| Group 2 | | |
|---|---|---|
| Seq | weight | AA |
| taxon B | 0.15 | V |
| taxon D | 0.25 | M |

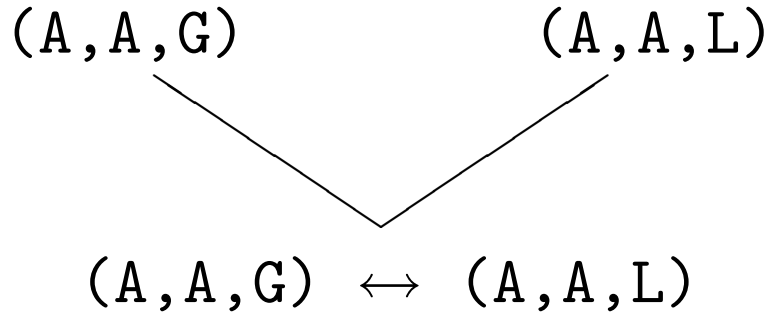
$$D_{G1,G2} = \frac{\sum_i \sum_j w_i w_j d_{ij}}{n_i n_j}$$

| Group 1 | | |
|---|---|---|
| Seq | weight | AA |
| taxon A | 0.3 | V |
| taxon C | 0.24 | A |
| taxon E | 0.19 | I |

| Group 2 | | |
|---|---|---|
| Seq | weight | AA |
| taxon B | 0.15 | V |
| taxon D | 0.25 | M |

$$
\begin{aligned}
D_{G1,G2} &= \frac{\sum_i \sum_j w_i w_j d_{ij}}{n_i n_j} \\
&= \frac{1}{6}\left[d(V,V)w_A w_B + d(V,M)w_A w_D + d(A,V)w_C w_B \ldots\right. \\
&\quad \left.\ldots d(A,M)w_C w_D + d(I,V)w_E w_B + d(I,M)w_E w_D\right] \\
&= \frac{1}{6}(\mathbf{4} \times 0.3 \times 0.15 + \mathbf{1} \times 0.3 \times 0.25 + \mathbf{0} \times 0.24 \times 0.15 \ldots \\
&\quad \ldots -1 \times 0.24 \times 0.25 + \mathbf{3} \times 0.19 \times 0.15 + \mathbf{1} \times 0.19 \times 0.15) \\
&= 1.46225
\end{aligned}
$$

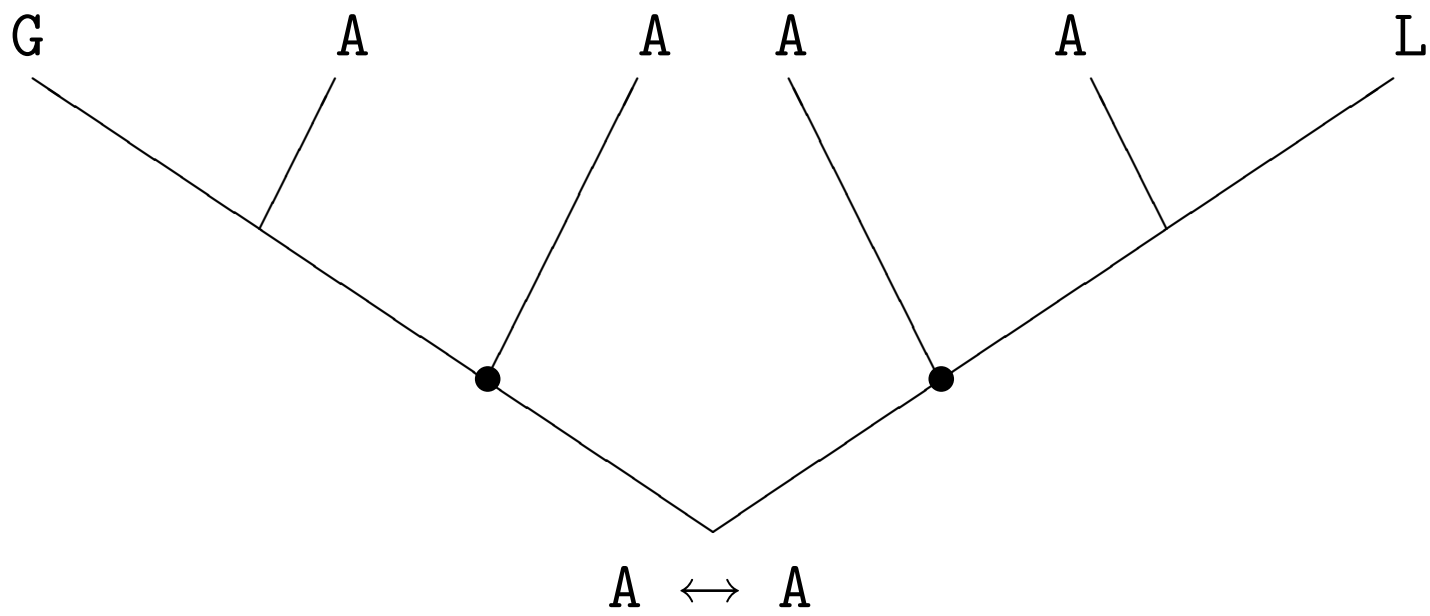
**Imperfect scoring system. Consider one position in a group-to-group alignment:**




The sum-of-pairs score for aligning would be:

$$\frac{4}{9}(A \leftrightarrow A) + \frac{2}{9}(A \leftrightarrow L) + \frac{2}{9}(G \leftrightarrow A) + \frac{1}{9}(G \leftrightarrow L)$$

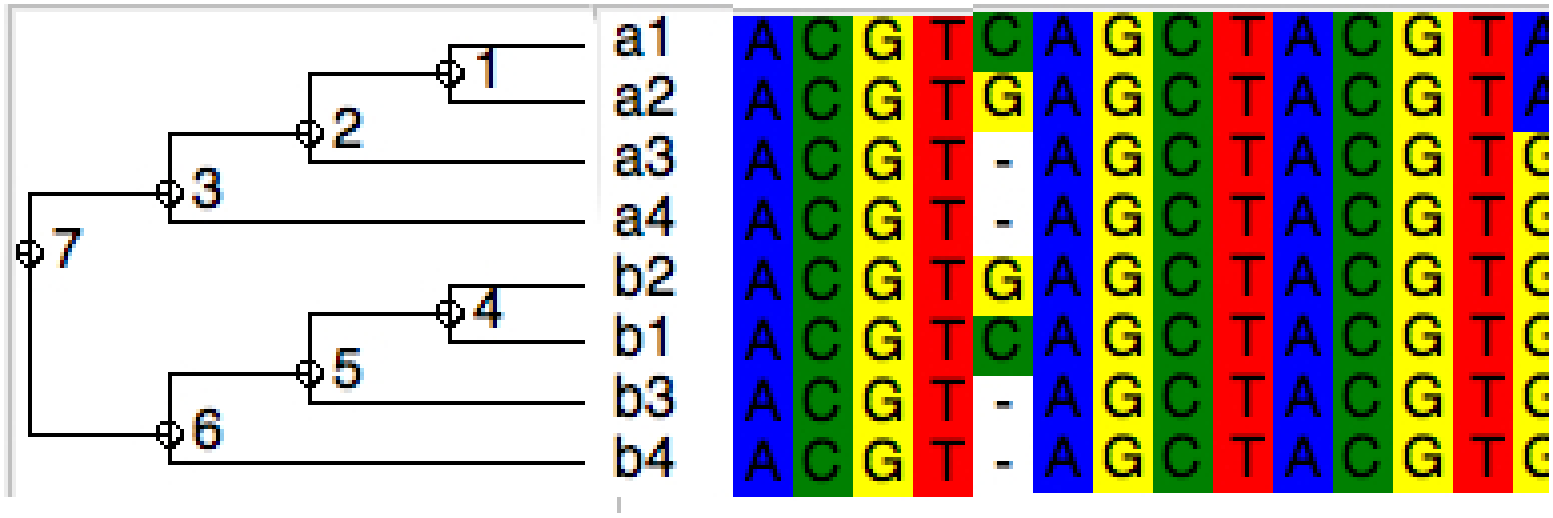
## But in the context of the tree we might be pretty certain of an A↔A event




Note: weighted sum-of-pairs would help reflect the effect of ancestry better (but still not perfectly; sum-of-pairs techniques are simply not very sophisticated forms of ancestral sequence reconstruction).

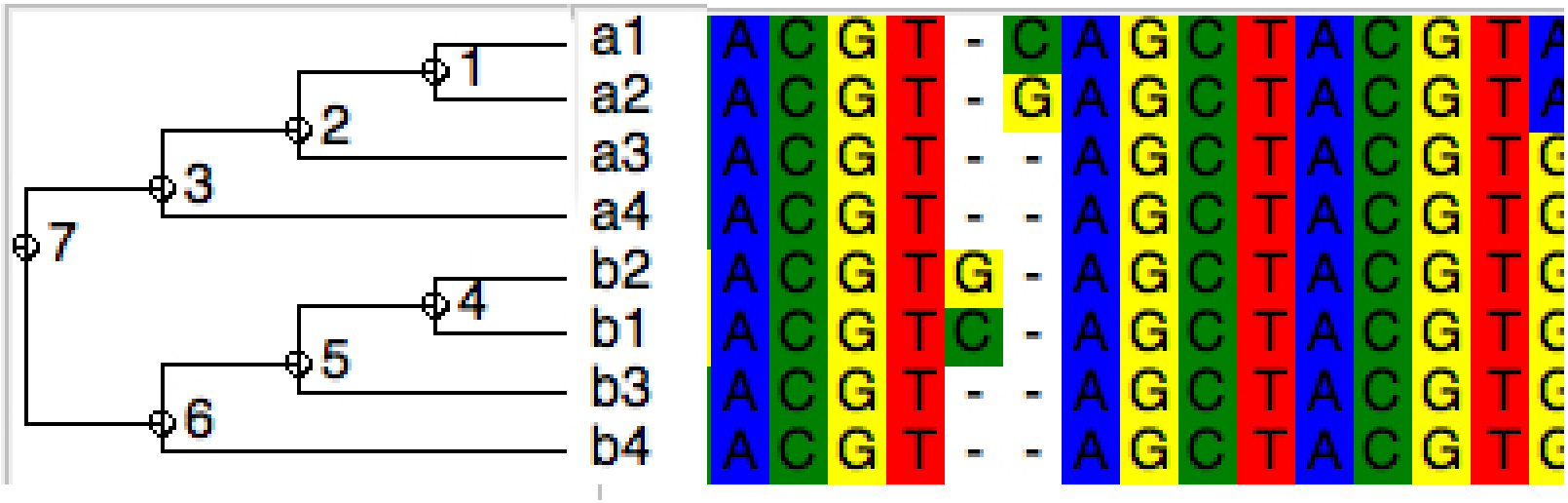
# PRANK

Löytynoja and Goldman (2005) showed most progressive alignment techniques were particularly prone to compression because of poor ancestral reconstruction:



# PRANK

Flagging inserted residues allows PRANK to effectively skip over these positions in the ancestor, producing more phylogenetically-sensible alignments:



## Progressive alignment

- Uses a guide tree to change the MSA problem into a series of pairwise alignment problems;

- May not return the alignment with the best weighted sum-of-pairs scores. Early alignment decisions get "locked in" Most aligners try to polish the alignment, but we cannot guarantee that we have found the optimal alignment;

- Reconstruction of ancestral sequences is usually done in a quick-and-dirty, implicit fashion;

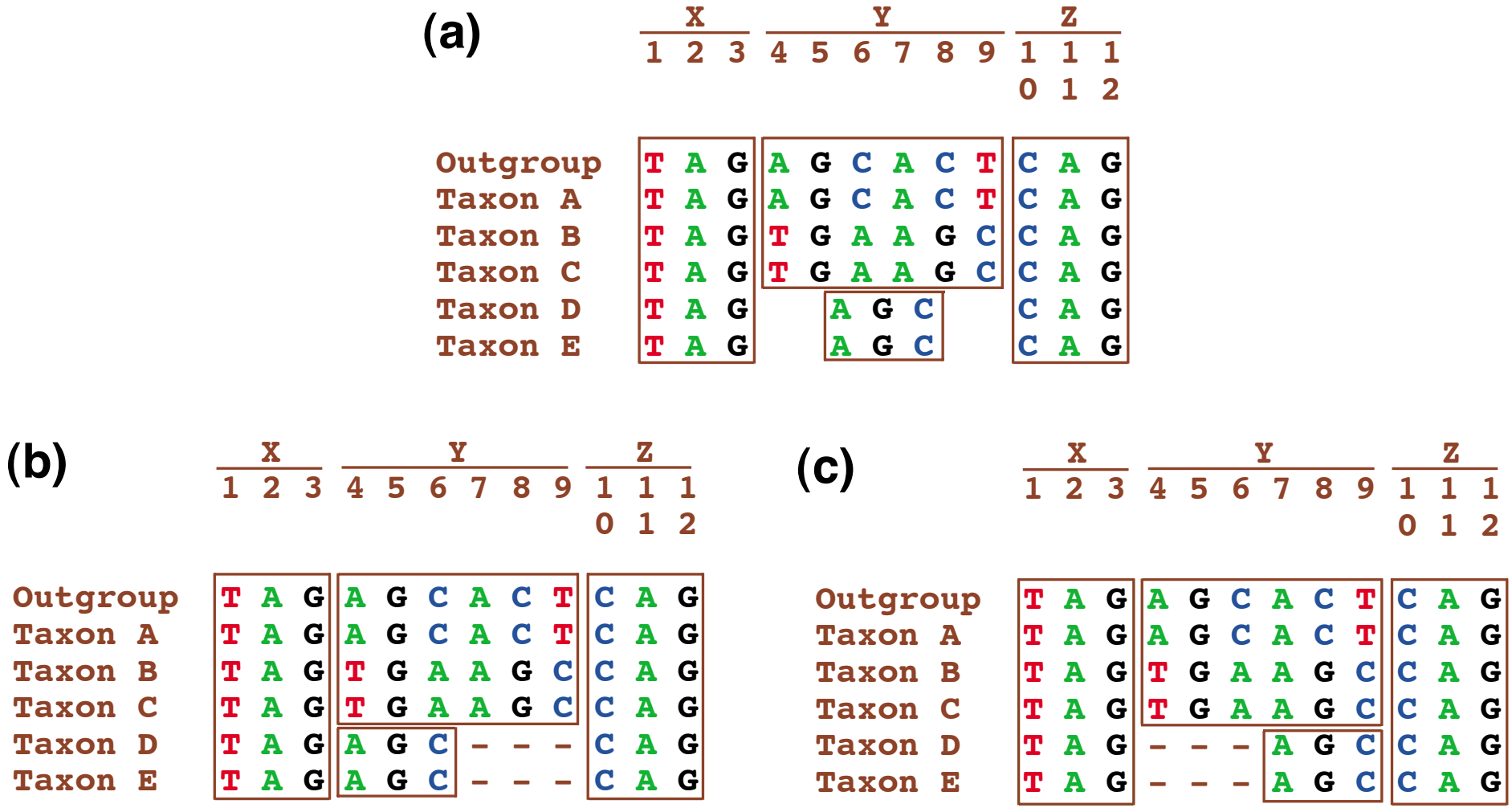
## Impact of the guide tree

Using a guide tree can bias subsequent tree inference toward the guide tree.

This can also cause inflated support.

Ironically, this effect may be more of a problem for a more evolutionarily-sensible aligner such as PRANK!

# Dealing with alignment ambiguity




from M. S. Y. Lee, *TREE*, 2001

# Dealing with alignment ambiguity - deletion

**(a)**

| | X | | | Y | | | | | | Z | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 1<br>0 | 1<br>1 | 1<br>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 | | | Z | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 1<br>0 | 1<br>1 | 1<br>2 |
| 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 | 1<br>0 | 1<br>1 | 1<br>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 | ? | ? | ? | - | - | - | C | A | G |
| Taxon E | T | A | G | ? | ? | ? | - | - | - | C | A | G |

from M. S. Y. Lee, *TREE*, 2001

## Filtering

GUIDANCE web-server (Penn et al., 2010a) provides filtering of columns based on:

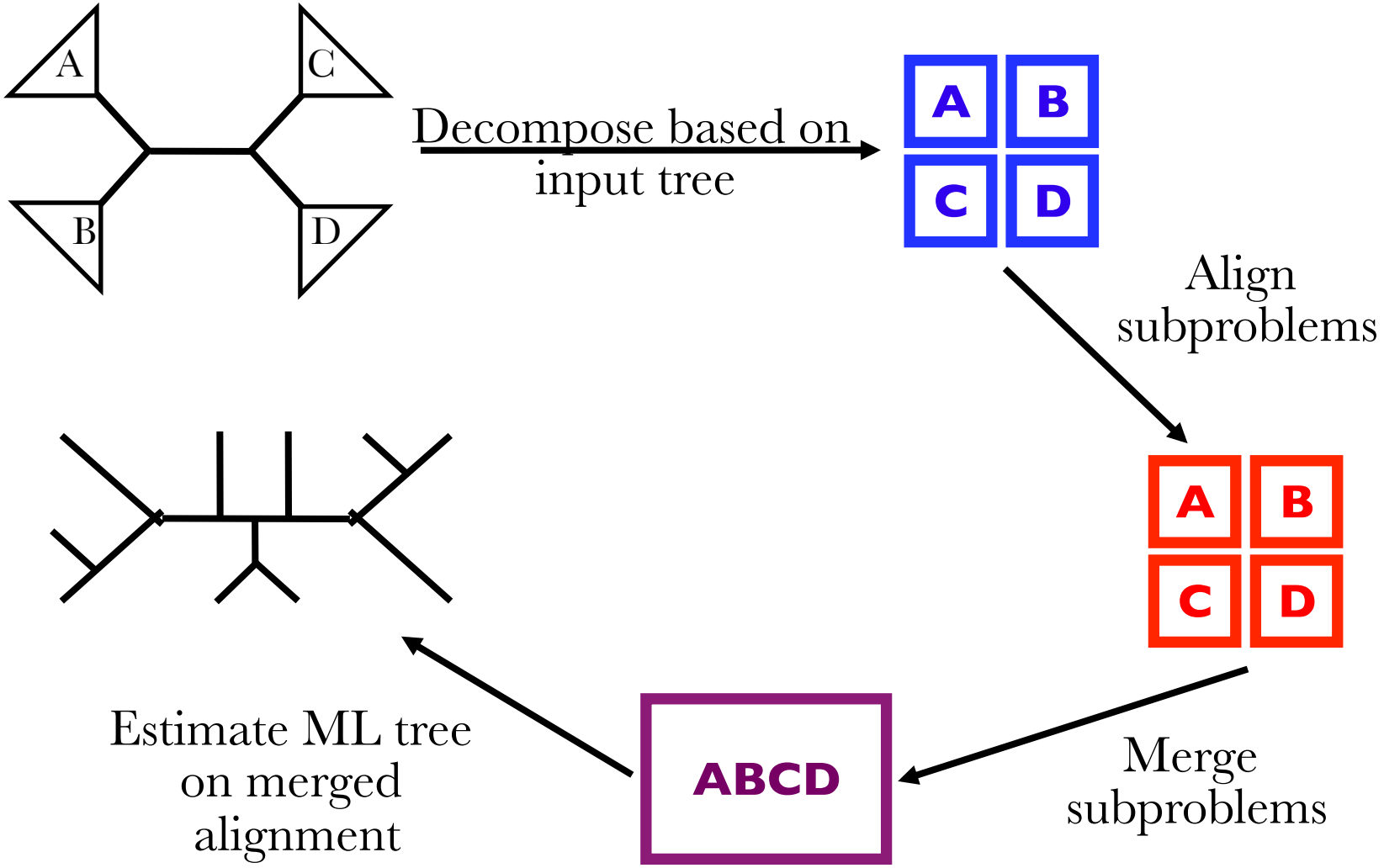
- GUIDANCE score (Penn et al., 2010b) which reflects sensitivity of the alignment to poorly supported parts of the guide tree; and
- Head-or-Tails algorithm (Landan and Graur, 2007) which compares the alignment of the sequences to the alignment that you would get from reversed forms of the sequences.

Other tools, such as Gblocks, examine properties of the columns in the matrix.

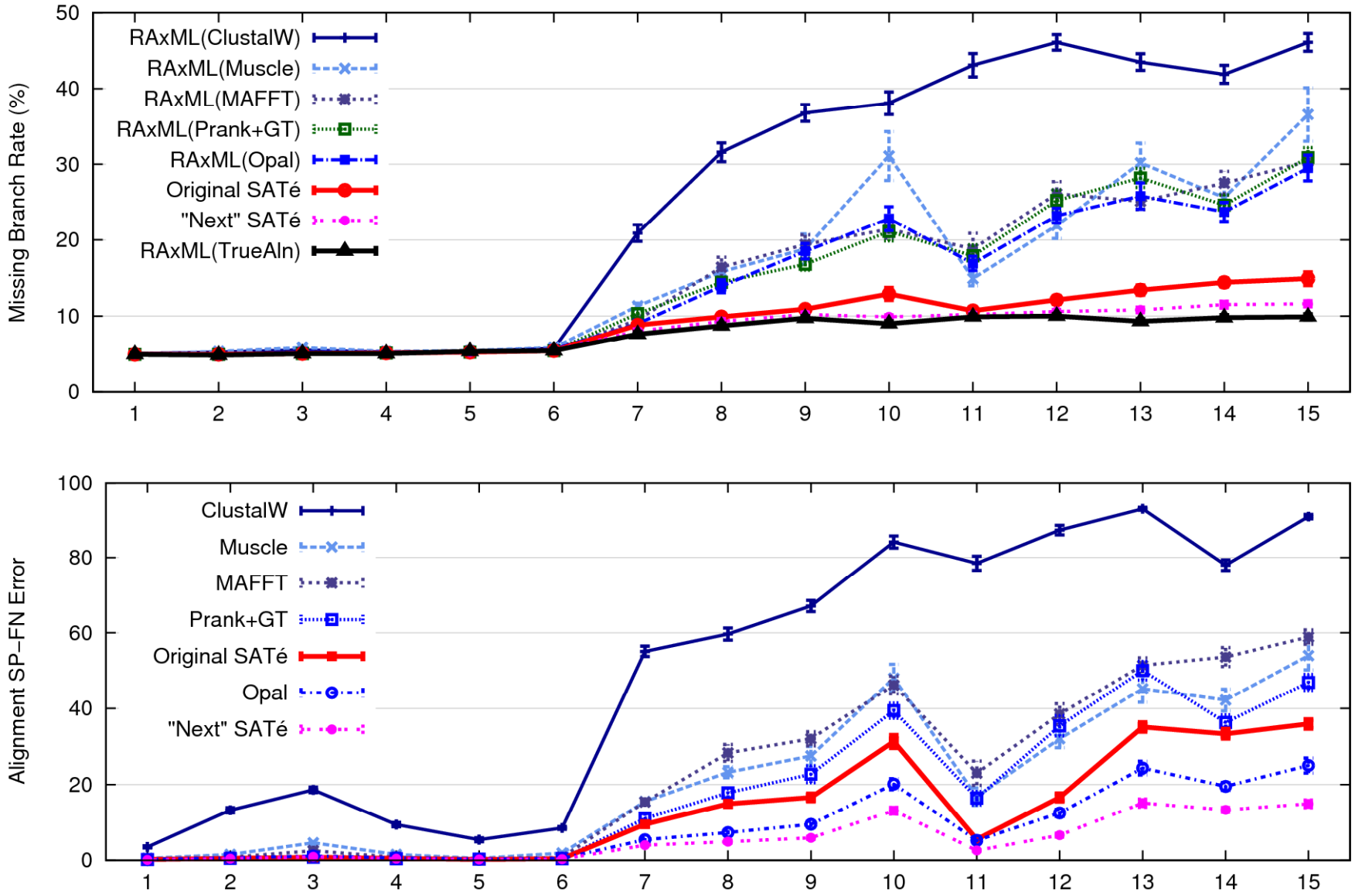
## Simultaneous tree inference and alignment

- 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: `BAliPhy` (Redelings and Suchard, 2005)
  See also: `POY` (Wheeler, Gladstein, Laet, 2002), `Handel` (Holmes and Bruno, 2001) , and `BEAST`(Lunter et al., 2005; Drummond and Rambaut, 2007). SATé (Liu et al., 2009); (Yu, Holder, Sukumaran, and Mirarab software).

## SATé repeats the following steps until termination



# SATé simulation results



## Optional alignment lab

http://molevol.mbl.edu/wiki/index.php/MSA_lab
Covers basic usage of:

- MAFFT,
- SeaView (alignment editor and viewer),

Pointers to:

- GUIDANCE web-server
- PRANK
- SATé

## References

Drummond, A. J. and Rambaut, A. (2007). Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.*, 7:214.

Holmes, I. and Bruno, W. J. (2001). Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics (Oxford, England)*, 17(9):803–820.

Landan, G. and Graur, D. (2007). Heads or tails: a simple reliability check for multiple sequence alignments. *Molecular Biology and Evolution*, 24(6):1380–1383.

Liu, K., Raghavan, S., Nelesen, S., Linder, C., and Warnow, T. (2009). Rapid and accurate large-scale coestimation

of sequence alignments and phylogenetic trees. *Science*, 324(5934):1561–1564.

Löytynoja, A. and Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Science USA*, 102(30):10557–62.

Lunter, G., Miklós, I., Drummond, A., Jensen, J. L., and Hein, J. (2005). Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics*, 6:83.

Penn, O., Privman, E., Ashkenazy, H., Landan, G., Graur, D., and Pupko, T. (2010a). GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic Acids Research*, 38:W23–W28.

Penn, O., Privman, E., Landan, G., Graur, D., and Pupko, T. (2010b). An alignment confidence score capturing robustness to guide tree uncertainty. *Molecular Biology and Evolution*, 27(8):1759–1767.

Redelings, B. and Suchard, M. (2005). Joint bayesian estimation of alignment and phylogeny. *Systematic Biology*, 54(3):401–418.

## Some of the tricks clustal uses to produce better alignments

- Chooses substitution matrix (PAM or BLOSUM series for amino acids) based on sequence similarity
- AA residues in a neighborhood affect gap opening penalty (easier to have gaps in hydrophilic loops)
- Gap penalties are raised if a column has no gaps, but there are gaps nearby.
- Low scoring alignments may be postponed until a later stage.

## Terminal gaps

Using normal gap costs causes problems if one sequence is missing the starting or ending residues:

```
Pongo     VDEFKLIVEGELGRLFVVPTQ
Gorilla   VD-------GELGRLFVVPTQ
```

instead of :

```
Pongo     VDEFKLIVEGELGRLFVVPTQ
Gorilla   -------VDGELGRLFVVPTQ
```

(Methods that utilize local alignment information are more appropriate in these cases).

Using free terminal gaps to avoid this problem, but you have to watch out for:

```
Pongo     VDEPFRFKLTNRGTSHIILVAPR
Gorilla   ------VDEPNRGTSHIILVAPR
```