CMPS 6630: Introduction to Computational Biology and Bioinformatics

Multiple Sequence Alignment and Phylogenetic Trees

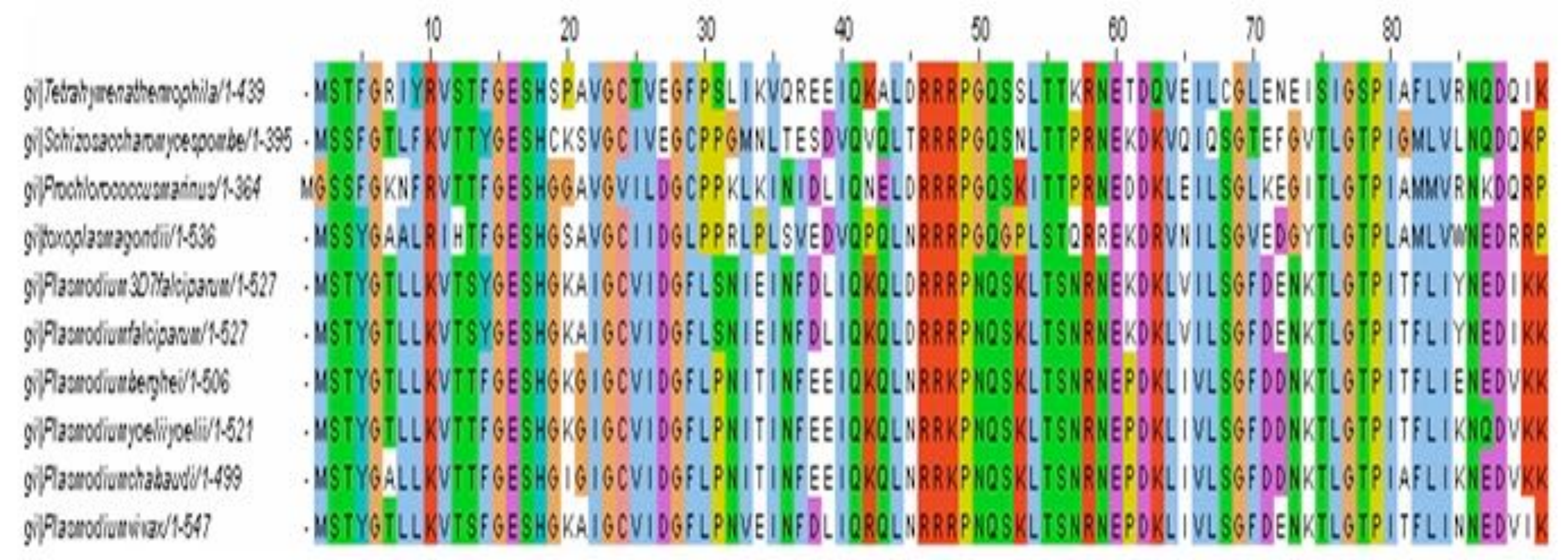
Sequence Alignment

- So Far:
  - Global Alignment
  - Local Alignment
  - Database Matching

- All were for a single pair of sequences

- What if we want to look at more than 2 sequences?

Multiple Sequence Alignment



```
                                           10        20        30        40        50        60        70        80
gi|Tetrahymenathermophila/1-439    .MSTFGRIYRVSTFGESHSPAVGCTVEGFPSLIKVQREEIQKALDRRRPGQSSLTTKRNETDQVEILCGLENEISIGSPIAFLVRNQDQIK
gi|Schizosaccharomycespombe/1-395  .MSSFGTLFKVTTYGESHCKSVGCIVEGCPPGMNLTESDVQVQLTRRRPGQSNLTTPRNEKDKVQIQSGTEFGVTLGTPIGMLVLNQDQKP
gi|Prochlorococcusmarinus/1-364    MGSSFGKNFRVTTFGESHGGAVGVILDGCPPKLKINIDLIQNELDRRRPGQSKITTPRNEDDKLEILSGLKEGITLGTPIAMMVRNKDQRP
gi|toxoplasmagondii/1-536          .MSSYGAALRIHTFGESHGSAVGCIIDGLPPRLPLSVEDVQPQLNRRRPGQGPLSTQRREKDRVNILSGVEDGYTLGTPLAMLVWNEDRRP
gi|Plasmodium3D7falciparum/1-527   .MSTYGTLLKVTSYGESHGKAIGCVIDGFLSNIEINFDLIQKQLDRRRPNQSKLTSNRNEKDKLVILSGFDENKTLGTPITFLIYNEDIKK
gi|Plasmodiumfalciparum/1-527      .MSTYGTLLKVTSYGESHGKAIGCVIDGFLSNIEINFDLIQKQLDRRRPNQSKLTSNRNEKDKLVILSGFDENKTLGTPITFLIYNEDIKK
gi|Plasmodiumberghei/1-506         .MSTYGTLLKVTTFGESHGKGIGCVIDGFLPNITINFEEIQKQLNRRKPNQSKLTSNRNEPDKLIVLSGFDDNKTLGTPITFLIENEDVKK
gi|Plasmodiumyoeliiyoelii/1-521    .MSTYGTLLKVTTFGESHGKGIGCVIDGFLPNITINFEEIQKQLNRRKPNQSKLTSNRNEPDKLIVLSGFDDNKTLGTPITFLIKNQDVKK
gi|Plasmodiumchabaudi/1-499        .MSTYGALLKVTTFGESHGIGIGCVIDGFLPNITINFEEIQKQLNRRKPNQSKLTSNRNEPDKLIVLSGFDDNKTLGTPIAFLIKNEDVKK
gi|Plasmodiumvivax/1-547           .MSTYGTLLKVTSFGESHGKAIGCVIDGFLPNVEINFDLIQRQLNRRRPNQSKLTSNRNEPDKLIVLSGFDENKTLGTPITFLINNEDVIK
```

- A weak signal between 2 sequences may be stronger in the context of multiple sequences
- May allow construction of Phylogenetic trees
- May assist in protein structure prediction

Approach 1:

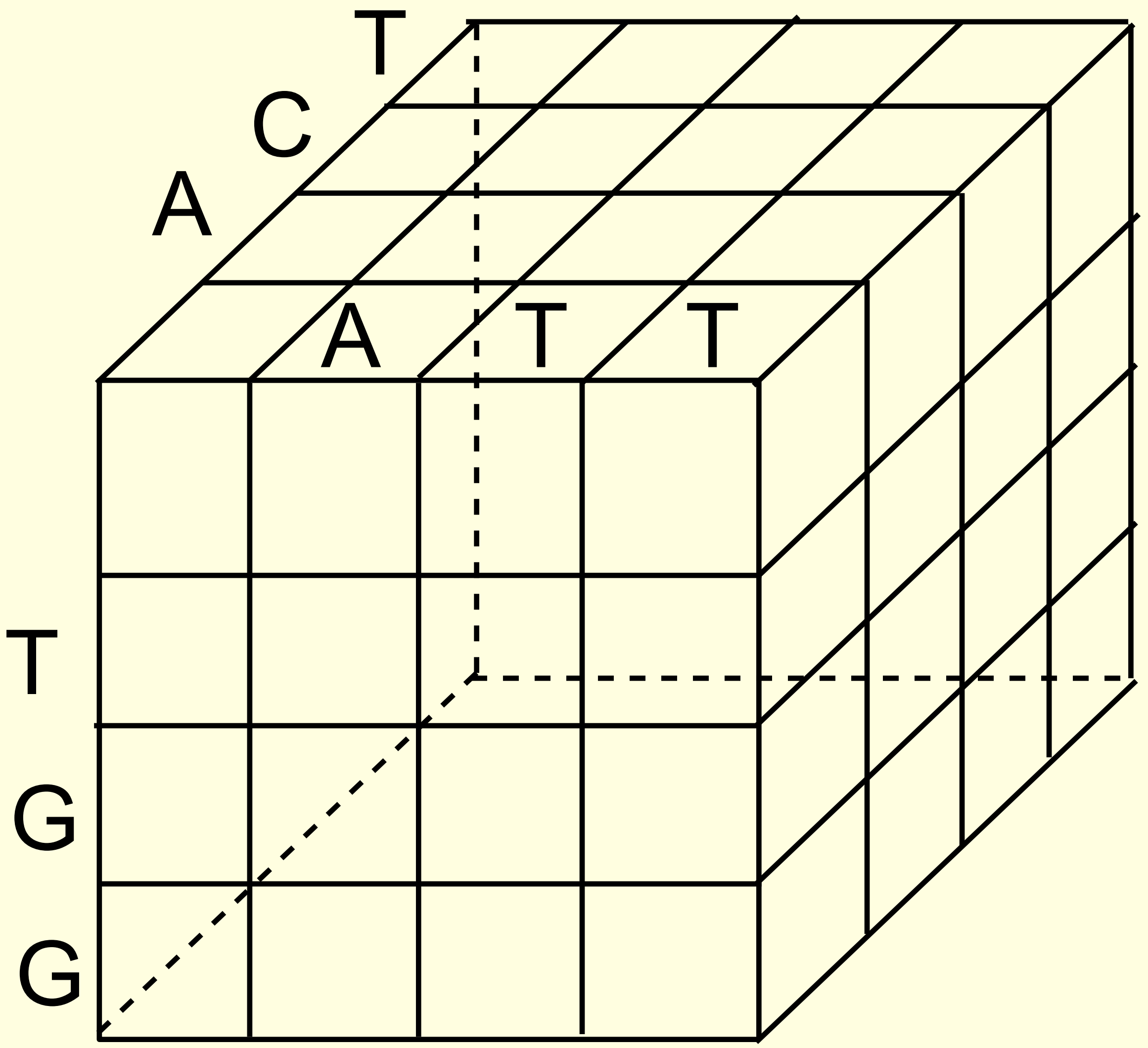
Select one sequence as the pivot. Perform pairwise alignment between each sequence and the pivot.

Locally Optimal

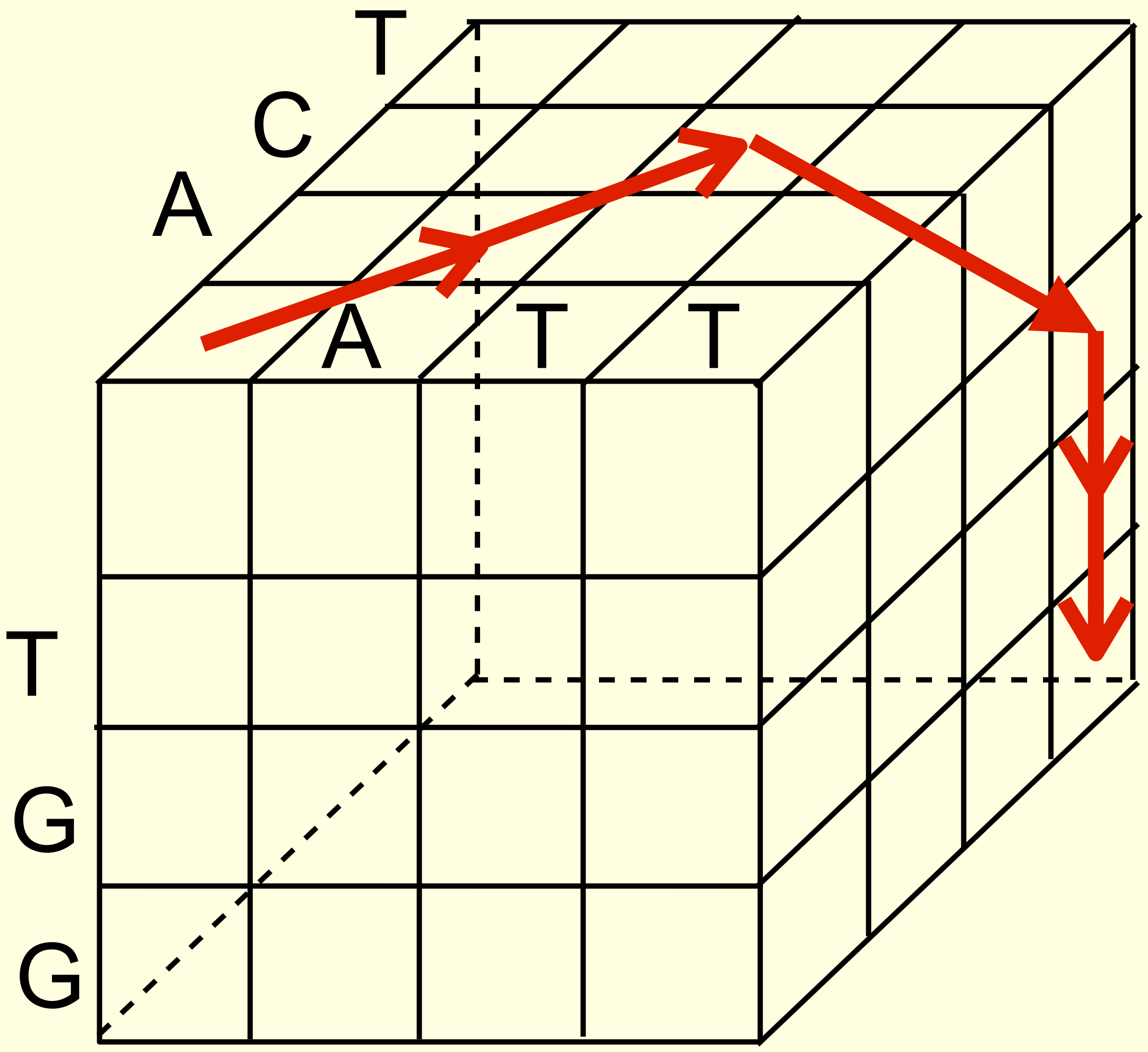
Multiple Sequence Analysis

- Multiple Sequence Alignment (MSA)
  - Extension of Pairwise Alignments
  - Scoring
  - Branch and Bound Pruning
- Phylogenetic Trees
  - Structure and How to Compute
- Combining Phylogenetic Trees and MSA
  - Progressive Alignment
  - Weighting
- CLUSTAL

Multiple Alignment



Multiple Alignment

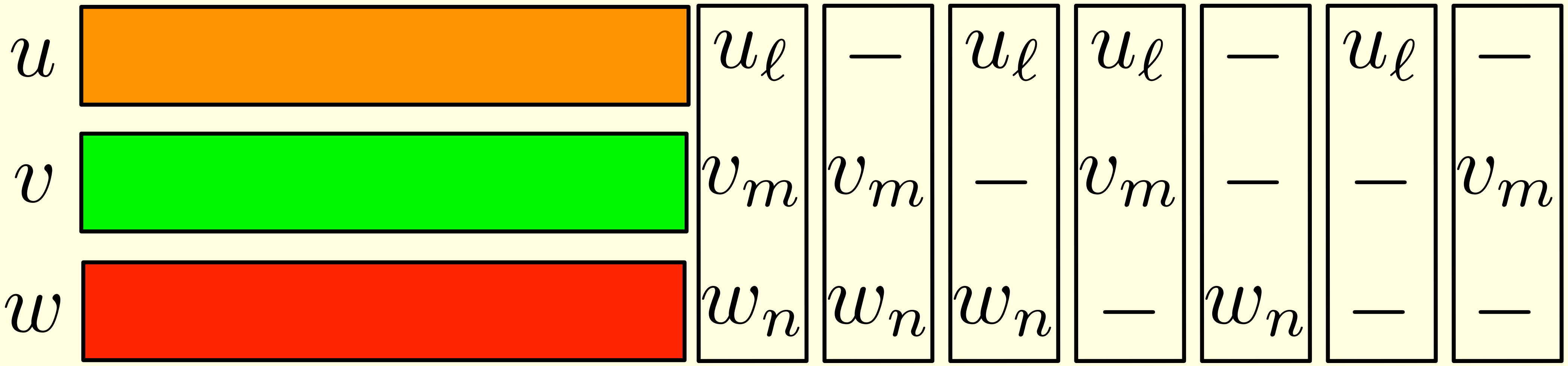


```
ACT--
ATT--
--TGG
```

Multiple Alignment

- What about the simple extension from 2D?

- There are seven possible “endings”:



$$\begin{matrix} u \\ v \\ w \end{matrix} \quad \begin{pmatrix} u_\ell \\ v_m \\ w_n \end{pmatrix} \begin{pmatrix} - \\ v_m \\ w_n \end{pmatrix} \begin{pmatrix} u_\ell \\ - \\ w_n \end{pmatrix} \begin{pmatrix} u_\ell \\ v_m \\ - \end{pmatrix} \begin{pmatrix} - \\ - \\ w_n \end{pmatrix} \begin{pmatrix} u_\ell \\ - \\ - \end{pmatrix} \begin{pmatrix} - \\ v_m \\ - \end{pmatrix}$$

$2^k - 1$ endings for k sequences. Why?

Multiple Alignment

$$s_{i,j,k} = \max \begin{cases} s_{i-1,j-1,k-1} + \delta(u_i, v_j, w_k) \\ s_{i-1,j-1,k} + \delta(u_i, v_j, -) \\ s_{i-1,j,k-1} + \delta(u_i, -, w_k) \\ s_{i,j-1,k-1} + \delta(-, v_j, w_k) \\ s_{i-1,j,k} + \delta(u_i, -, -) \\ s_{i,j,k-1} + \delta(-, -, w_k) \\ s_{i,j-1,k} + \delta(-, v_j, -) \\ s_{i,j,k-1} + \delta(-, -, w_k) \end{cases}$$

$\delta(x, y, z)$ is an entry in the 3D scoring matrix

Time and space grow exponentially with number of sequences.

Scoring

Sum-of-Pairs Scoring (SP):

$$S(\mathcal{A}) = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} S(\bar{s}^i, \bar{s}^j)$$

$\mathcal{A}$	a multiple alignment
$\bar{s}^i$	“projection” of sequence i (i with gaps)
$S(\bar{s}^i, \bar{s}^j)$	score of pairwise alignment

Idea: A good multiple alignment should contain good pairwise alignments

Scoring

Weighted Sum-of-Pairs Scoring (SP):

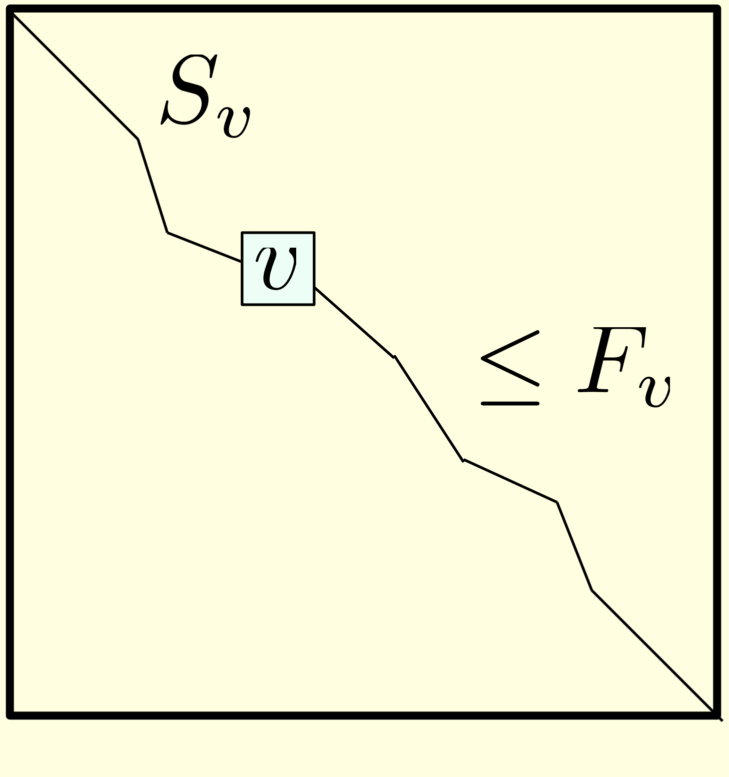
$$S(\mathcal{A}) = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} S(\bar{s}^i, \bar{s}^j) w^i w^j$$

The presence of many similar scores should not dominate the alignment, so we should incorporate some *a priori* knowledge of the sequences (i.e., organism/ family).

Pruning the DP Matrix

The dynamic programming matrix is large, but we only want the best alignment, and most matrix elements are not on that path.

Can we ‘direct’ the search to avoid evaluating cells that are provably not on the best path?



S_v Score of the best path from start to v

F_v Bound on the best path from v to the end

K Score of best known alignment

What if: $S_v + F_v < K$

Pruning the DP Matrix

`ARSTVK, ASVK, ARTR`

Let $v = (3, 2, 2)$

S_v is score of best alignment of: **`ARS, AS, AR`**

F_v is upper bound on score of aligning: **`TVK, VK, TR`**

If $S_v + F_v < K$ then mark v as dead-ending (aka prune v)

Pruning the DP Matrix

```
begin
  $v = h_0$; $P(v) = 0$; push($v$, $Q$)    ;push start cell on the queue
  while $Q \neq \emptyset$
    pop($v$, $Q$); $S(v) = P(v)$
    if $S(v) + \boxed{F(v,h_N)} \geq K$ then
      for all forward neighbors $w$ of $v$ do in the right order
        if $w \notin Q$ then
          push($w$, $Q$); $P(w) = S(v) + D(v,w)$
        else
          $P(w) = \max( P(w), S(v) + D(v,w) )$
        end
      end
    end
  end
end
```

Pruning the DP Matrix

We know the alignment score is:

$$S(\mathcal{A}) \leq \sum_{k=1}^{m-1} \sum_{l=k+1}^{m} S(s^k, s^l)$$

Observation:

$$S(\mathcal{A}) = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} S(\bar{s}^i, \bar{s}^j)$$

$$S(\bar{s}^i, \bar{s}^j) \leq S(s^i, s^j)$$

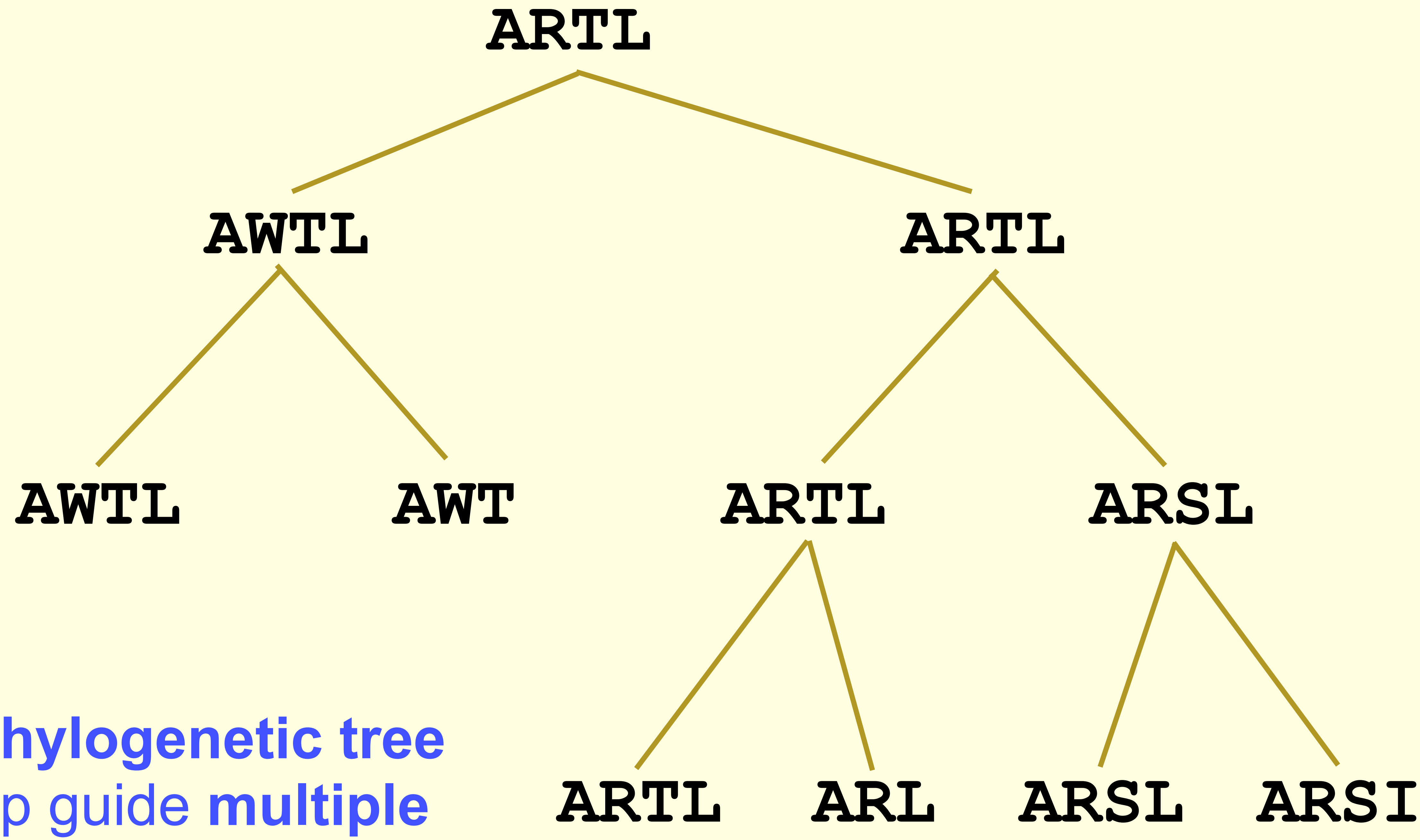
So our bound can be:

$$F_v = \sum_{k=1}^{m-1} \sum_{l=k+1}^{m} S(s^k_{v_k+1 \ldots n_k}, s^l_{v_l+1 \ldots n_l})$$

Runtime for computing F (using dynamic programming):

$$O(n^2 m^2)$$

Phylogenetic Trees



A true **phylogenetic tree** can help guide **multiple alignment** and vice-versa

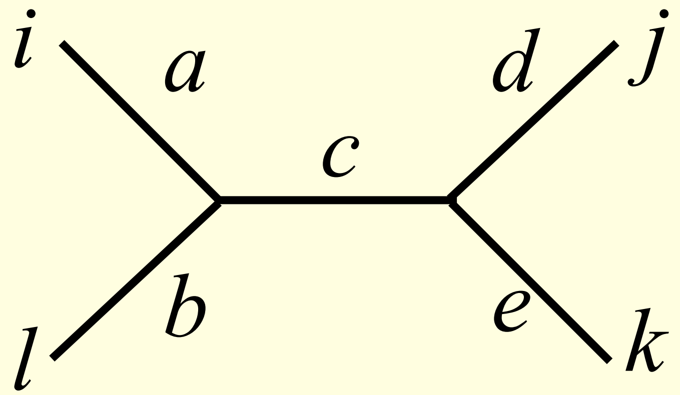
Phylogenetic Trees

- **Goals**
  - Reconstruct genealogical ties (the topology of the tree)
  - Estimate time of divergence (last common ancestor)
- **Given *m* original sequences**
  - Tree should have *m* **leaf** nodes
  - Tree is **rooted** or **unrooted**
  - Set of internal nodes
    - Rooted tree: 2 children, one parent
    - Unrooted tree: 3 connected nodes
  - Internal nodes may have explicit sequence
  - Edges may carry a **weight** (representing *distance*)
  - **Additive** tree
    - Distance between two nodes is sum of edge distances
  - **Ultrametric** tree
    - An Additive tree where distance to common ancestors are equal

Phylogenetic Trees

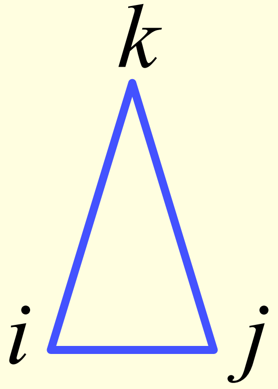
- Additive Trees
  - Distance between two nodes is sum of edge distances
  - It's possible to construct an additive tree iff for any 4 nodes they can be labeled:

$$D_{i,j} + D_{k,l} = D_{i,k} + D_{j,l} \geq D_{i,l} + D_{j,k}$$



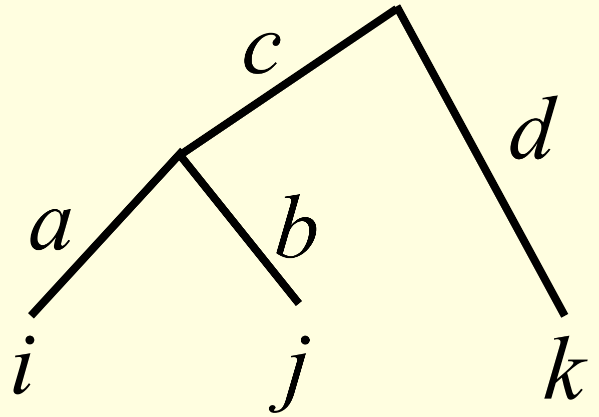
- Ultrametric Trees
  - An Additive tree where distance to common ancestors are equal
  - It's possible to construct an ultrametric tree iff for every triple *i, j, k*

$$D_{i,j} \leq \max(D_{i,k}, D_{k,j})$$



3 points arranged as an isosceles triangle with the odd side shorter than the other two

$$a = b$$
$$a + c = b + c = d$$

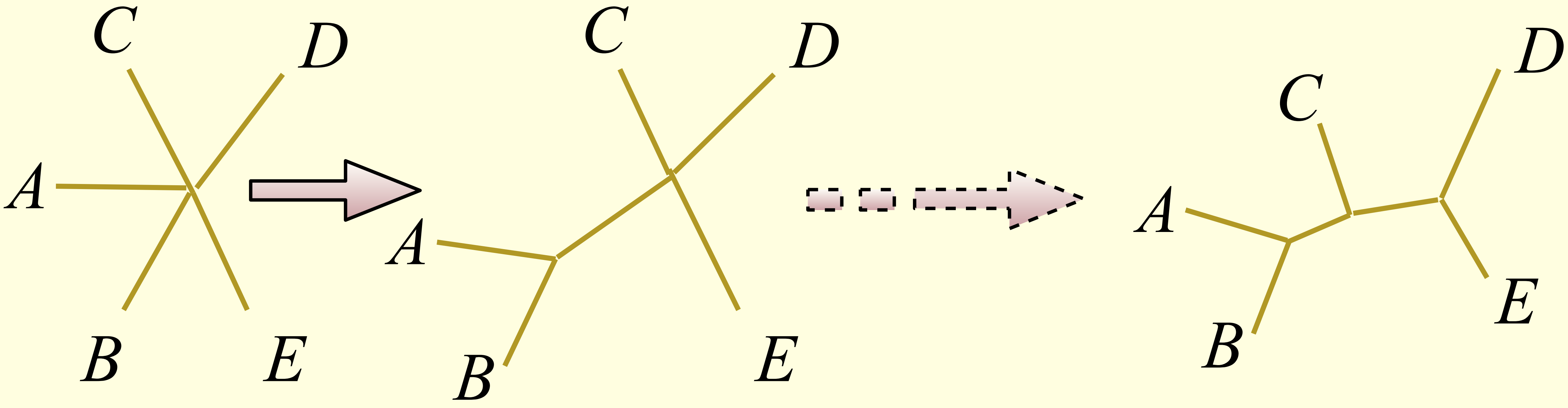


Constructing Phylo Trees

- Three Main Methods
  - Maximum Parsimony
    - Input: Multiple alignment
    - Method: Find tree that minimizes number of evolutionary changes
  - Maximum Likelihood
    - Input: Multiple alignment
    - Method: Identify statistically most likely tree (use substitution probs)
  - Distance-Based Methods
    - Input: Multiple alignment
    - Method: Use pairwise alignment distances, connect similar sequences

Distance-Based Methods

- Neighbor-Joining Method (NJ)



- Start with all nodes arranged in a star topology
  - This topology has minimum number of edges
- Tree is successively changed by adding edges until an unrooted tree is constructed
- The two elements (a neighbor pair) that would create the overall smallest total edge length, are ‘merged’

Heuristic Alignment

Progressive Alignment

- Let’s return to Multiple Alignments
- Consider constructing an alignment ‘pairwise’

```
$C = \emptyset$
for $i = 1$ to $m$ do
    $C = C \bigcup \{\{s^i\}\}$
end
for $i = 1$ to $m - 1$ do
    *choose two alignments $A_p$, $A_q$ from $C$;*
    $C = C - \{A_p, A_q\}$
    $A_r = \boxed{\text{align}(A_p, A_q)}$
    $C = C \bigcup \{A_r\}$
end
```

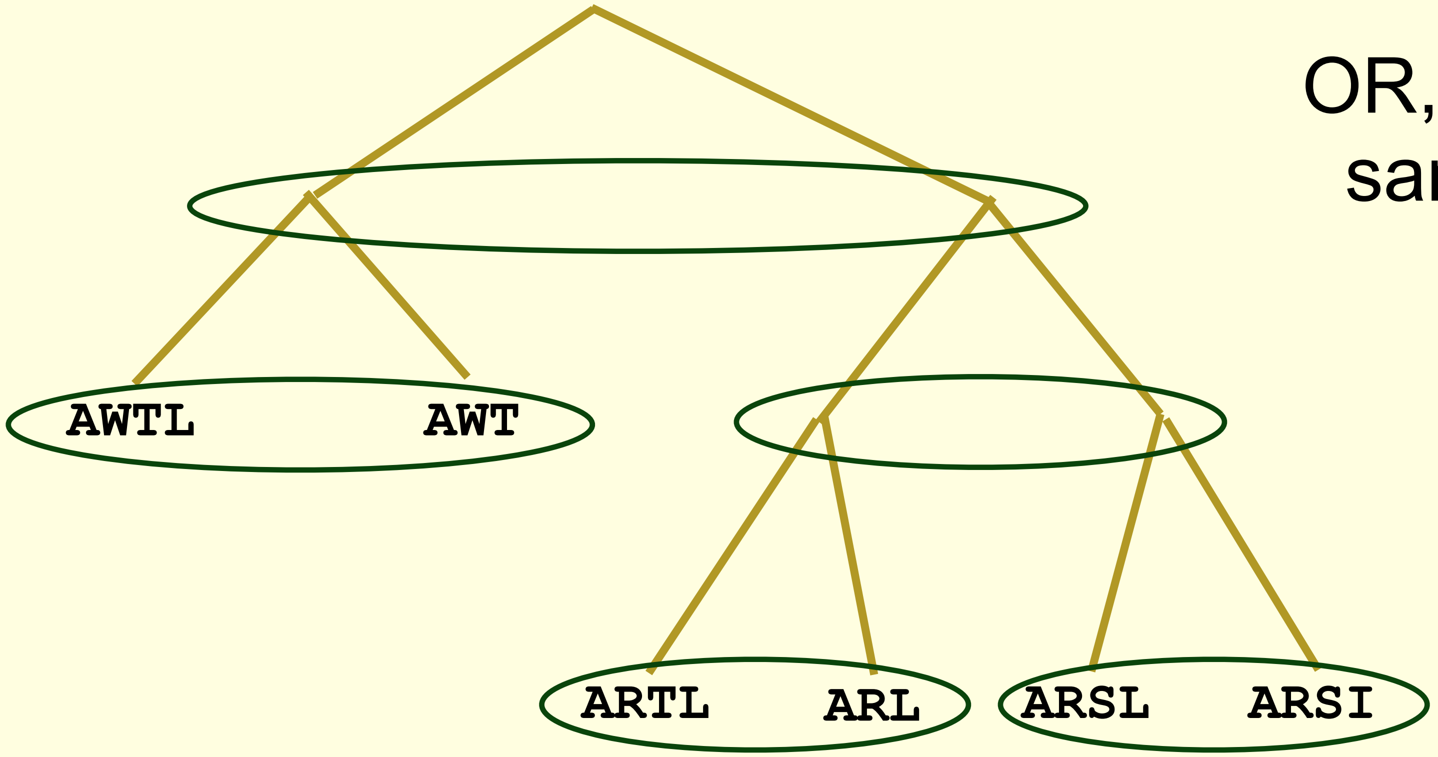
Method of aligning two sequences, an alignment and a sequence, or two alignments

Merge order is important

Guiding Progressive Alignment

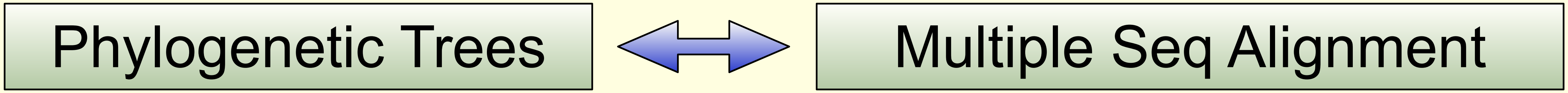
Phylogenetic Trees ↔ Multiple Seq Alignment

- **Observation**: Deterministic Multiple Seq Alignment is difficult
- **Observation**: Result of heuristic alignment very much depends on alignment ordering
- **Observation**: Multiple Sequence Alignment is supposed to reflect evolutionary relationships between sequences
- **Idea**: Use phylogenetic tree to guide alignment



OR, construct both at the same time

Guiding Progressive Alignment


Phylogenetic Trees ↔ Multiple Seq Alignment

- **Observation**: Deterministic Multiple Seq Alignment is difficult
- **Observation**: Result of heuristic alignment very much depends on alignment ordering
- **Observation**: Multiple Sequence Alignment is supposed to reflect evolutionary relationships between sequences
- **Idea**: Use phylogenetic tree to guide alignment

Hierarchical Clustering

Select 2 subset alignments using the **Average Linkage Model**

Score is average score between all pairs of sequences (one from each alignment)

Similar to PGMA

Align 2 subset alignments using **Complete Alignment**

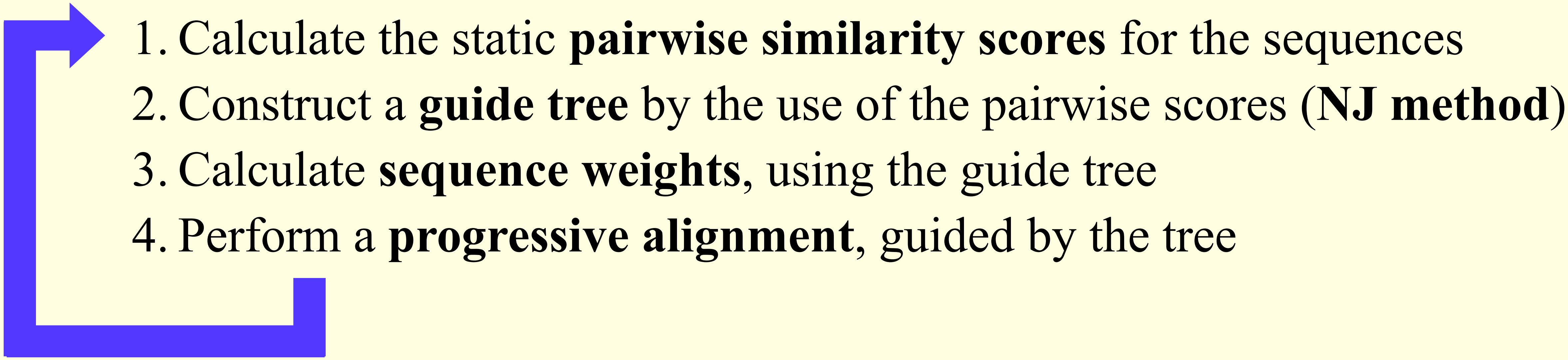
Linear Clustering

Special case where one member being aligned must be a sequence

CLUSTAL

http://www.clustal.org/

- Widely used Multiple Sequence Alignment Program



1. Calculate the static **pairwise similarity scores** for the sequences
2. Construct a **guide tree** by the use of the pairwise scores (**NJ method**)
3. Calculate **sequence weights**, using the guide tree
4. Perform a **progressive alignment**, guided by the tree

Profiles

- Consider you have multiple genes from the same family and you want to find more.
  - Pairwise alignment?
    - Slow, may miss weak signals
  - Multiple sequence alignment?
    - Slow, may be better with weak signals
- Other Alternatives?
- **Profiles**: A statistically grounded model describing the properties of the member sequences
  - Motifs
  - Position-Specific Scoring Matrices (PSSMs)
  - Profiles
  - Hidden Markov Model Profiles

Constructing a Profile

- Let $\mathbf{Prof}_{ra}$ be the score of aligning amino acid a (of the sequence) to the position r (of the profile)
- When computing Prof_{ra} consider
  - Observed amino acids at position r
  - Number of observations at position r
  - The similarity of a to the residues seen at position r
  - The background distribution of amino acids
  - The weight of each sequence (as in MSA)
  - Number of gaps at position r

$$\mathrm{Prof}_{ra} = \sum_{b \in \mathcal{M}} R_{ba} V_{rb}$$

where

$\mathcal{M}$	Alphabet of Amino Acids
R_{ba}	Scoring Matrix
$\boxed{V_{rb}}$	Weight of b at position r

Constructing a Profile

- Consider V_{rb}
  - Ideally

$$V_{rb} = 0 \text{ when } T_{rb} = 0$$

$$V_{rb} = 1 \text{ when } T_{rb} = m_r$$

num occurrences of b in position r of alignment

num residues (not gaps) at position r of alignment

  - Can interpolate in various ways, or weight:

$$V_{rb} = \frac{\sum_{i=1}^{m_r} w_i \delta_b}{\sum_{i=1}^{m_r} w_i}$$

$$\delta_b = 1 \text{ if } \bar{s}_r^i = b$$
$$\delta_b = 0 \text{ if } \bar{s}_r^i \neq b$$

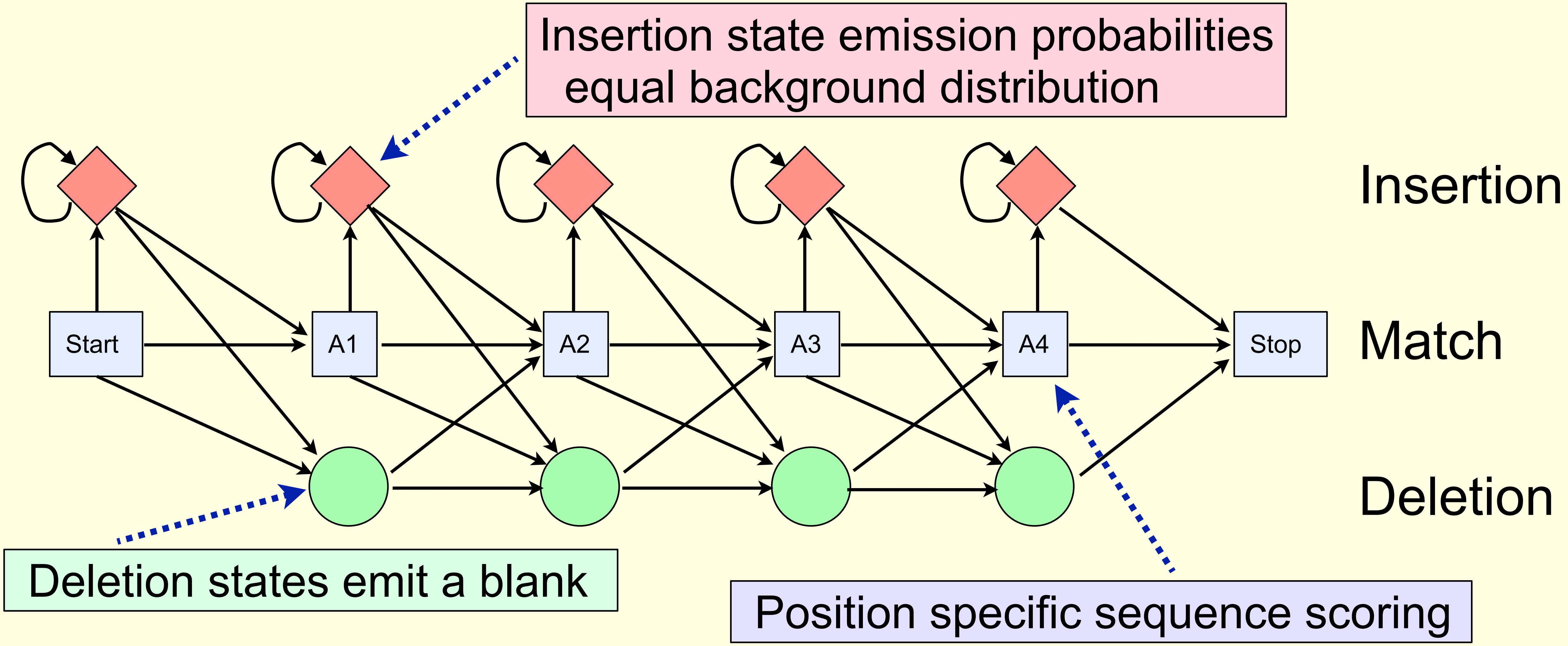
- **Gaps**
  - If *sufficient* number of gaps in position of MSA, remove that position from profile

PSI-BLAST

- Position Specific Iterated BLAST (PSI-BLAST
- Idea
  - **Perform BLAST** with original query q
  - **Construct profile** P from sequences aligned to q (pivot)
  - **Perform another BLAST** with P as new query
  - **Construct new profile** P' from sequences
  - Repeat until convergence or max iterations
- Capable of identifying more distant homologs
- Useful in structure prediction

Profile HMM

- Consider an HMM, constructed to emit sequences similar to some known profile
- Such an HMM might have states corresponding to insertions, matches, and deletions

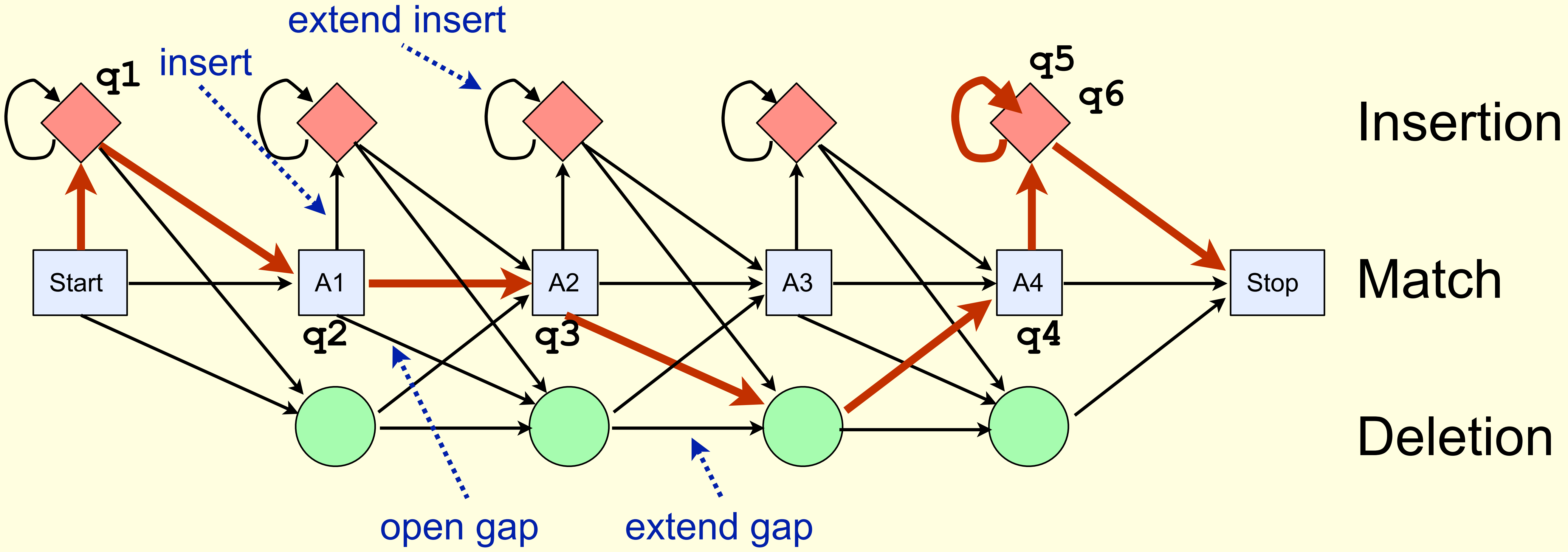


Insertion state emission probabilities equal background distribution

Deletion states emit a blank

Position specific sequence scoring

Profile HMM



Consider query sequence and best profile alignment

```
q1  q2  q3   -  q4  q5  q6
 -  A1  A2  A3  A4   -   -
```

Profile HMM

- Best path through HMM is found via DP (what were you expecting?)
  - **Viterbi algorithm**, recurrence is:

$$D_{i+1,j} = \max_{0 \leq k \leq j-1} (D_{i,k} P(k,j) P(q_{i+1}|j))$$

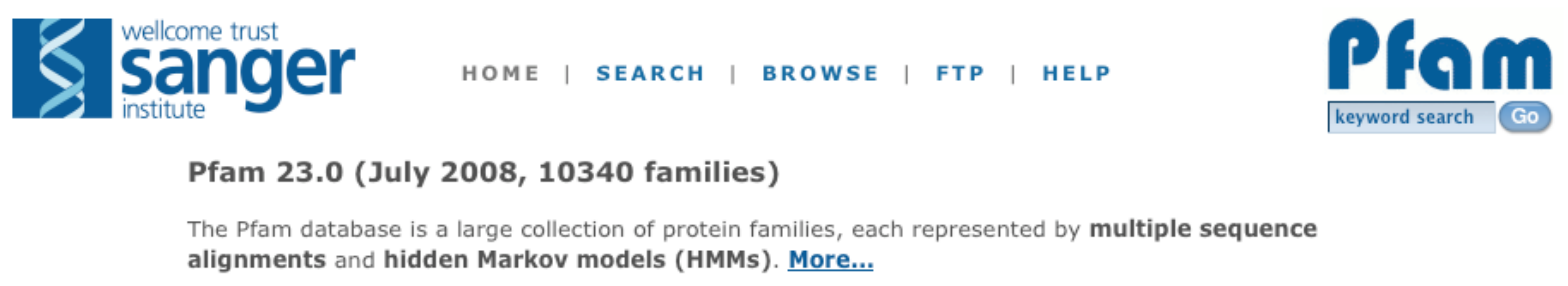
- Note that for Profile HMMs, most transition probabilities are zero.
- This leads to efficient implementations.
- One may be interested in the sum of probabilities over all paths capable of generating the query sequence.

Profile Databases

Many protein family databases with signatures for each family

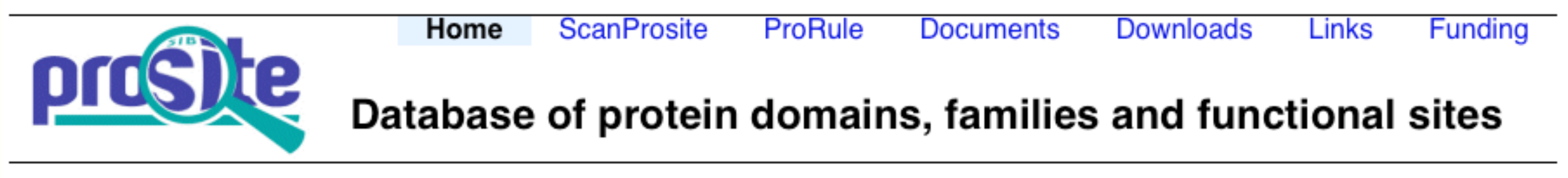
PFam, PROSITE, PRINTS, BLOCKS, ...

- When searching a query against the database, compute the sequence emission probability for each family's HMM
- Report those above a threshold



Pfam 23.0 (July 2008, 10340 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. More...



Database of protein domains, families and functional sites