Pairwise Alignment

Mark Voorhies

3/29/2011

Review

- FASTA files

```
>Name Free-form annotation
MGCLLIMKEGGPGRKHKLIVMLYLDENQ
EHELPIMTRAPPEDINADNAMACHINEW
NQEDLYMNILKHGPPGEDEDRKHEDEDG
```

Review

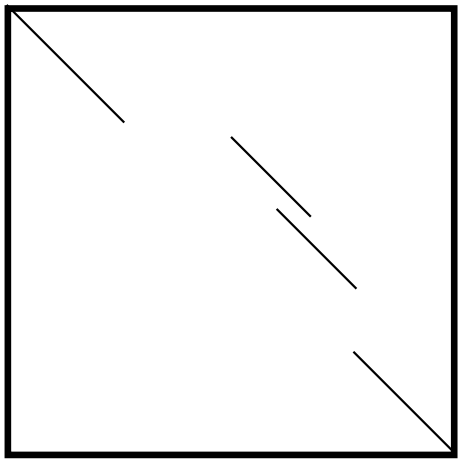
- FASTA files

```
>Name Free-form annotation
MGCLLIMKEGGPGRKHKLIVMLYLDENQ
EHELPIMTRAPPEDINADNAMACHINEW
NQEDLYMNILKHGPPGEDEDRKHEDEDG
```

- Dotplots: unbiased plot of all possible ungapped alignments of two sequences.

Pairwise Alignment

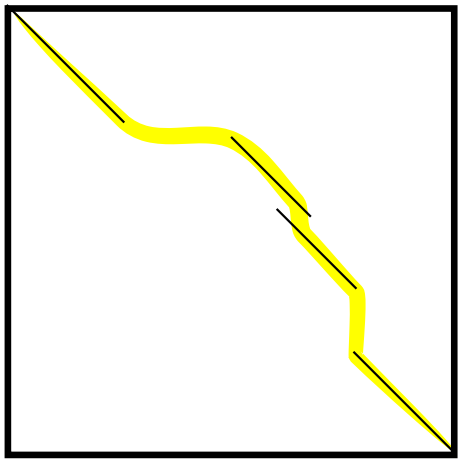
How can we automate our dotplot protocol to find the "best" gapped alignment of our sequences?



Pairwise Alignment

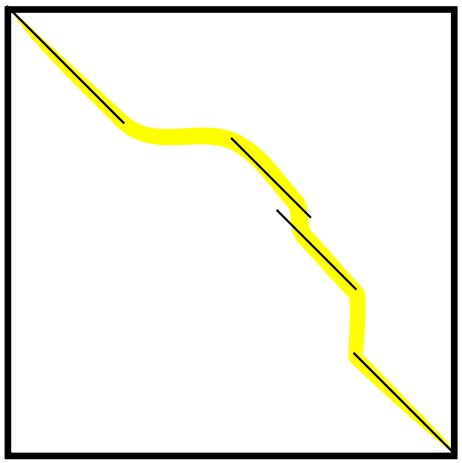
How can we automate our dotplot protocol to find the "best" gapped alignment of our sequences?

What do we mean by best?



Pairwise Alignment

How can we automate our dotplot protocol to find the "best" gapped alignment of our sequences?

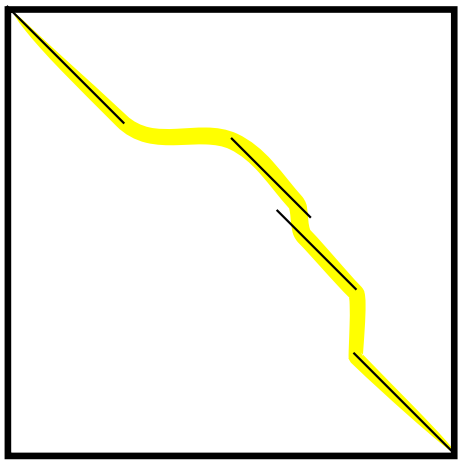


What do we mean by best?

- Residues with equivalent functional roles are paired

Pairwise Alignment

How can we automate our dotplot protocol to find the "best" gapped alignment of our sequences?

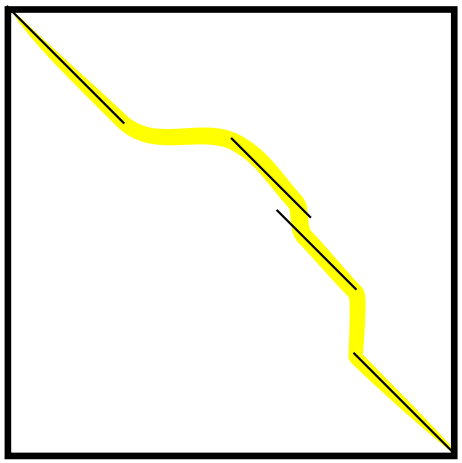


What do we mean by best?

- Residues with equivalent functional roles are paired
- Residues that derive from the same position in the common ancestor are paired (homology)

Pairwise Alignment

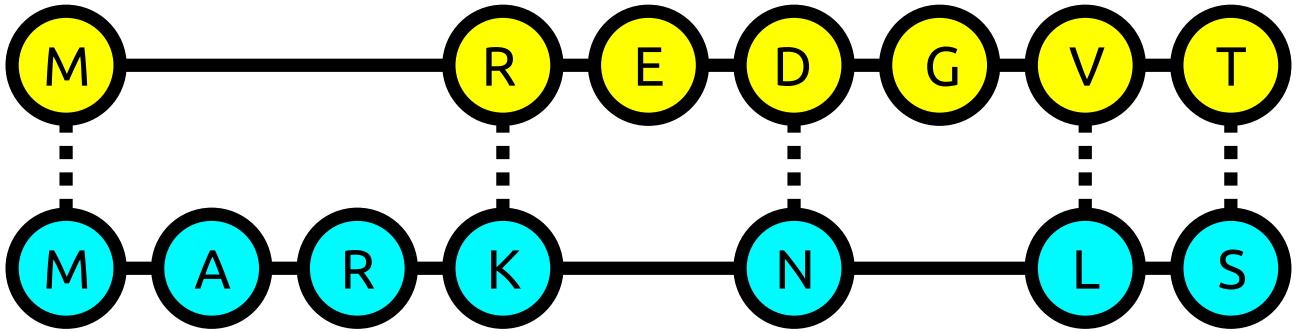
How can we automate our dotplot protocol to find the "best" gapped alignment of our sequences?



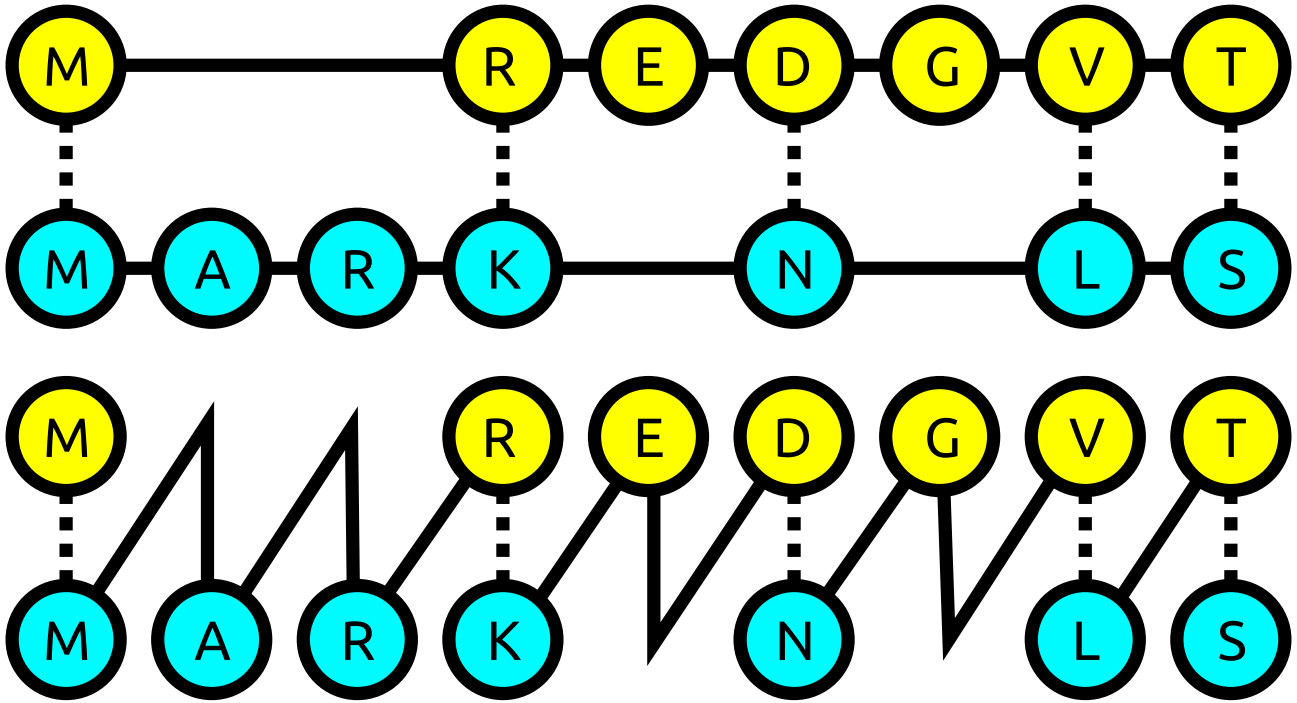
What do we mean by best?

- Residues with equivalent functional roles are paired
- Residues that derive from the same position in the common ancestor are paired (homology)
- The sequence alignment maximizes a similarity function

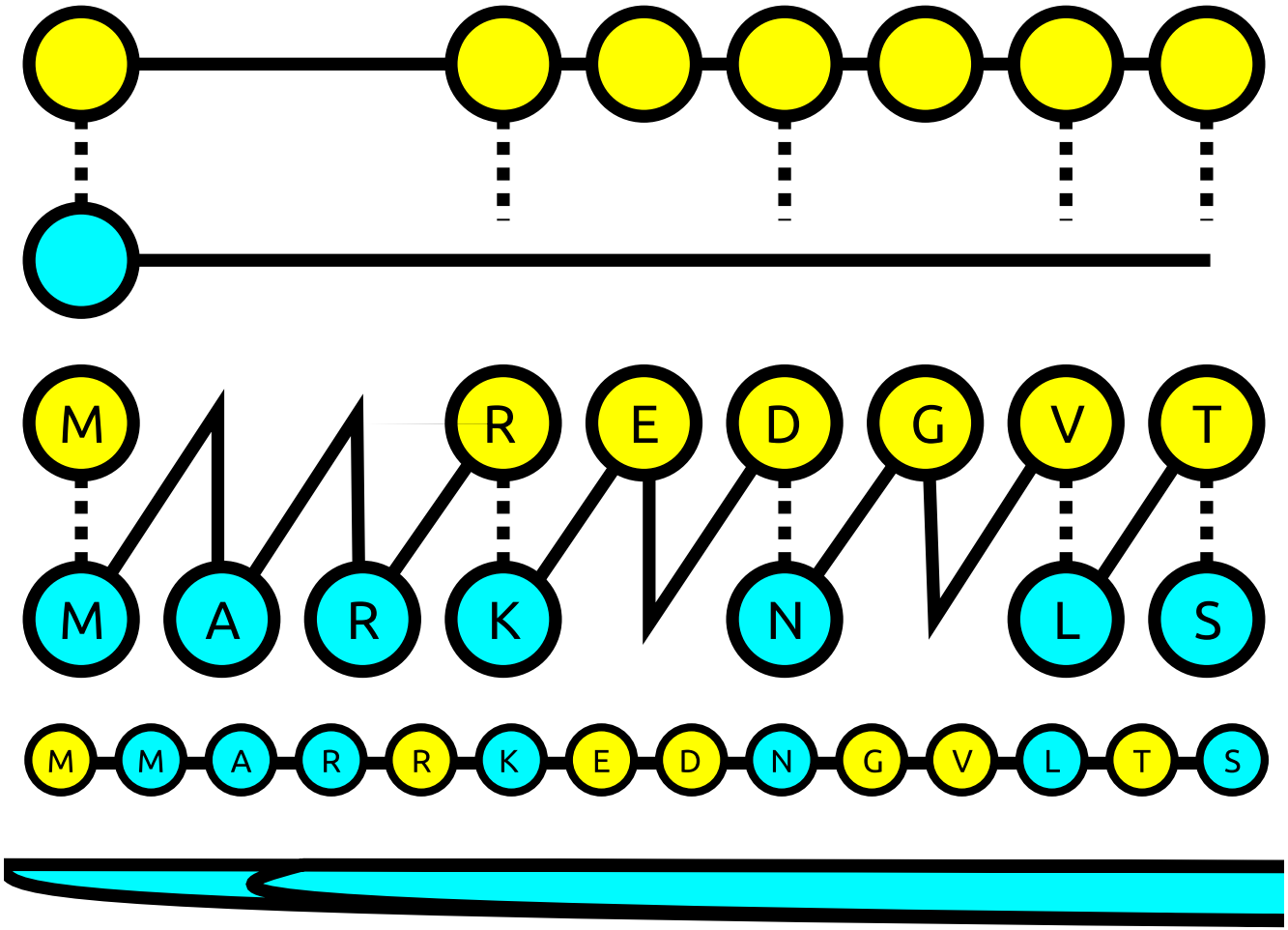
How many ways can we align two sequences?



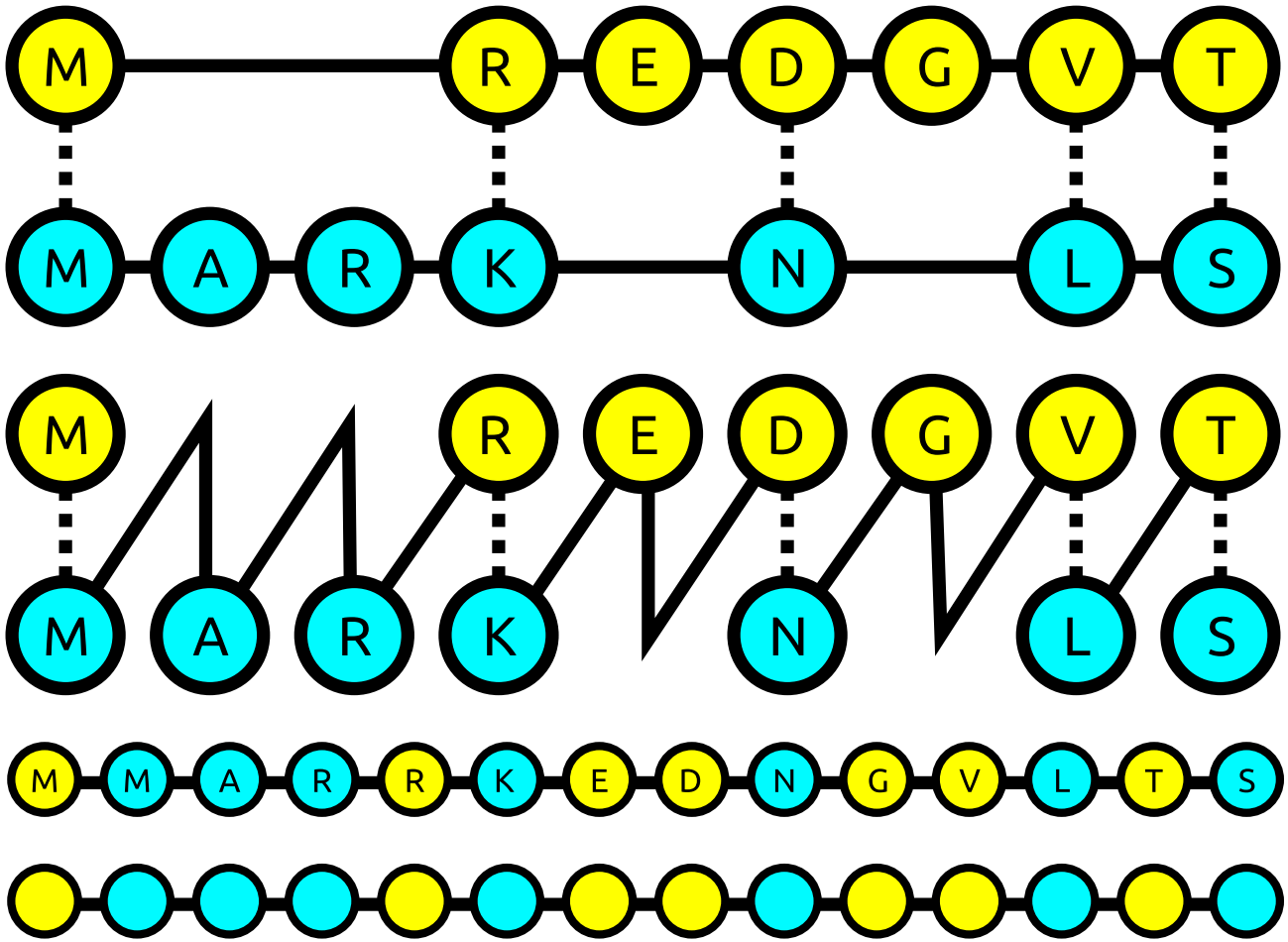
How many ways can we align two sequences?



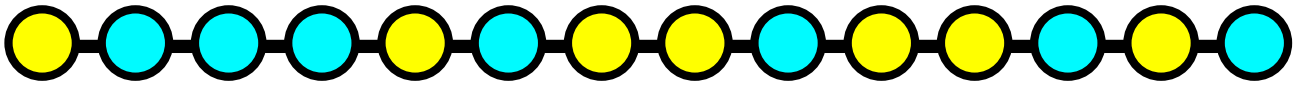
How many ways can we align two sequences?



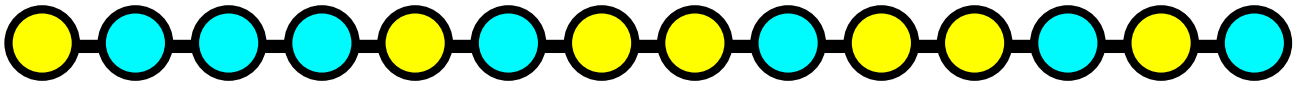
How many ways can we align two sequences?



How many ways can we align two sequences?



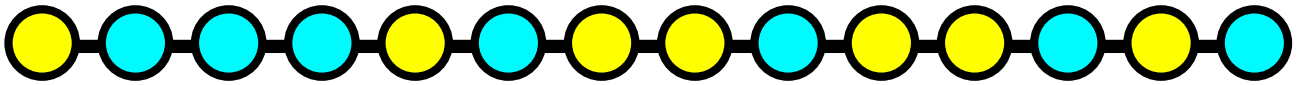
How many ways can we align two sequences?



Binomial formula:

$$\binom{k}{r} = \frac{k!}{(k-r)!r!} \tag{1}$$

How many ways can we align two sequences?

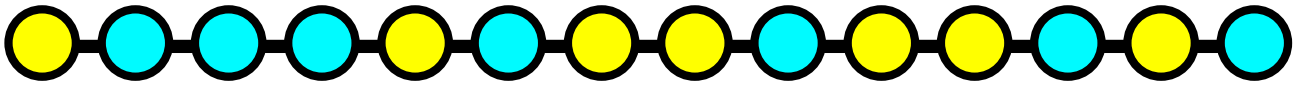


Binomial formula:

$$\binom{k}{r} = \frac{k!}{(k-r)!r!} \tag{1}$$

$$\binom{2n}{n} = \frac{(2n)!}{n!n!} \tag{2}$$

How many ways can we align two sequences?



Binomial formula:

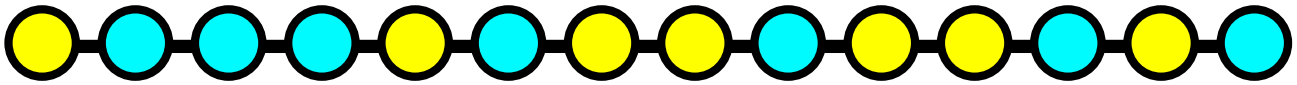
$$\binom{k}{r} = \frac{k!}{(k-r)!r!} \tag{1}$$

$$\binom{2n}{n} = \frac{(2n)!}{n!n!} \tag{2}$$

Stirling's approximation:

$$x! \approx \sqrt{2} \quad x^{x+\frac{1}{2}} \quad e^{-x} \tag{3}$$

How many ways can we align two sequences?



Binomial formula:

$$\binom{k}{r} = \frac{k!}{(k-r)!r!} \tag{1}$$

$$\binom{2n}{n} = \frac{(2n)!}{n!n!} \tag{2}$$

Stirling's approximation:

$$x! \approx \sqrt{2} \quad x^{x+\frac{1}{2}} \quad e^{-x} \tag{3}$$

$$\binom{2n}{n} \approx \frac{2^{2n}}{\sqrt{n}} \tag{4}$$

Log-odds scores

Frequency of residue i:

$$p_i \tag{5}$$

Log-odds scores

Frequency of residue i:

$$p_i \tag{5}$$

Frequency of residue i aligned to residue j:

$$q_{ij} \tag{6}$$

Log-odds scores

Frequency of residue i:

$$p_i \tag{5}$$

Frequency of residue i aligned to residue j:

$$q_{ij} \tag{6}$$

Expected frequency if i and j are independent:

$$p_i p_j \tag{7}$$

Log-odds scores

Frequency of residue i:

$$p_i \tag{5}$$

Frequency of residue i aligned to residue j:

$$q_{ij} \tag{6}$$

Expected frequency if i and j are independent:

$$p_i p_j \tag{7}$$

Ratio of observed to expected frequency:

$$\frac{q_{ij}}{p_i p_j} \tag{8}$$

Log-odds scores

Frequency of residue i:

$$p_i \tag{5}$$

Frequency of residue i aligned to residue j:

$$q_{ij} \tag{6}$$

Expected frequency if i and j are independent:

$$p_i p_j \tag{7}$$

Ratio of observed to expected frequency:

$$\frac{q_{ij}}{p_i p_j} \tag{8}$$

Log odds (LOD) score:

$$s(i;j) = \log \frac{q_{ij}}{p_i p_j} \tag{9}$$

PAM (Dayhoff) and BLOSUM matrices

- PAM1 matrix originally calculated from manual alignments of highly conserved sequences (myoglobin, cytochrome C, etc.)

PAM (Dayhoff) and BLOSUM matrices

- PAM1 matrix originally calculated from manual alignments of highly conserved sequences (myoglobin, cytochrome C, etc.)
- We can think of a PAM matrix as evolving a sequence by one unit of time.

PAM (Dayhoff) and BLOSUM matrices

- PAM1 matrix originally calculated from manual alignments of highly conserved sequences (myoglobin, cytochrome C, etc.)
- We can think of a PAM matrix as evolving a sequence by one unit of time.
- If evolution is uniform over time, then PAM matrices for larger evolutionary steps can be generated by multiplying PAM1 by itself (so, higher numbered PAM matrices represent greater evolutionary distances).

PAM (Dayhoff) and BLOSUM matrices

- PAM1 matrix originally calculated from manual alignments of highly conserved sequences (myoglobin, cytochrome C, etc.)
- We can think of a PAM matrix as evolving a sequence by one unit of time.
- If evolution is uniform over time, then PAM matrices for larger evolutionary steps can be generated by multiplying PAM1 by itself (so, higher numbered PAM matrices represent greater evolutionary distances).
- The BLOSUM matrices were determined from automatically generated ungapped alignments. Higher numbered BLOSUM matrices correspond to *smaller* evolutionary distances. BLOSUM62 is the default matrix for BLAST.

Fun with logarithms

In log space, multiplication and division become addition and subtraction:

$$\begin{aligned} \log(xy) &= \log(x) + \log(y) \\ \log(x{=}y) &= \log(x) - \log(y) \end{aligned}$$

Fun with logarithms

In log space, multiplication and division become addition and subtraction:

$$\begin{aligned} \log(xy) &= \log(x) + \log(y) \\ \log(x{=}y) &= \log(x) - \log(y) \end{aligned}$$

Therefore, exponentiation becomes multiplication:

$$\log(x^y) = y \log(x)$$

Fun with logarithms

In log space, multiplication and division become addition and subtraction:

$$\begin{aligned} \log(xy) &= \log(x) + \log(y) \\ \log(x{=}y) &= \log(x) - \log(y) \end{aligned}$$

Therefore, exponentiation becomes multiplication:

$$\log(x^y) = y \log(x)$$

Also, we can change of the base of a logarithm like so:

$$\log_A(x) = \log(x){=}\log(A)$$

Scoring an alignment

Multiplying independent probabilities is equivalent to adding independent log probabilities.

Scoring an alignment

Multiplying independent probabilities is equivalent to adding independent log probabilities.
Therefore, for an ungapped alignment can be scored as:

$$S(x; y) = \sum_{i} s(x_i; y_i) \tag{10}$$

Scoring an alignment

Multiplying independent probabilities is equivalent to adding independent log probabilities.
Therefore, for an ungapped alignment can be scored as:

$$S(x; y) = \sum_{i} s(x_i; y_i) \tag{10}$$

What about gaps?

Scoring an alignment

Multiplying independent probabilities is equivalent to adding independent log probabilities.
Therefore, for an ungapped alignment can be scored as:

$$S(x; y) = \sum_{i} s(x_i; y_i) \tag{10}$$

What about gaps?

- Probability of an insertion/deletion event (gap opening, G)
- Length distribution of insertions/deletions (gap extension, E)

Scoring an alignment

Multiplying independent probabilities is equivalent to adding independent log probabilities.
Therefore, for an ungapped alignment can be scored as:

$$S(x; y) = \sum_{i}^{N} s(x_i; y_i) \tag{10}$$

What about gaps?

- Probability of an insertion/deletion event (gap opening, G)
- Length distribution of insertions/deletions (gap extension, E)

$$S_{gapped}(x; y) = S(x; y) + \sum_{i}^{gaps} (G + E * L_i) \tag{11}$$

Scoring an alignment

Multiplying independent probabilities is equivalent to adding independent log probabilities.
Therefore, for an ungapped alignment can be scored as:

$$S(x; y) = \sum_{i}^{N} s(x_i; y_i) \tag{10}$$

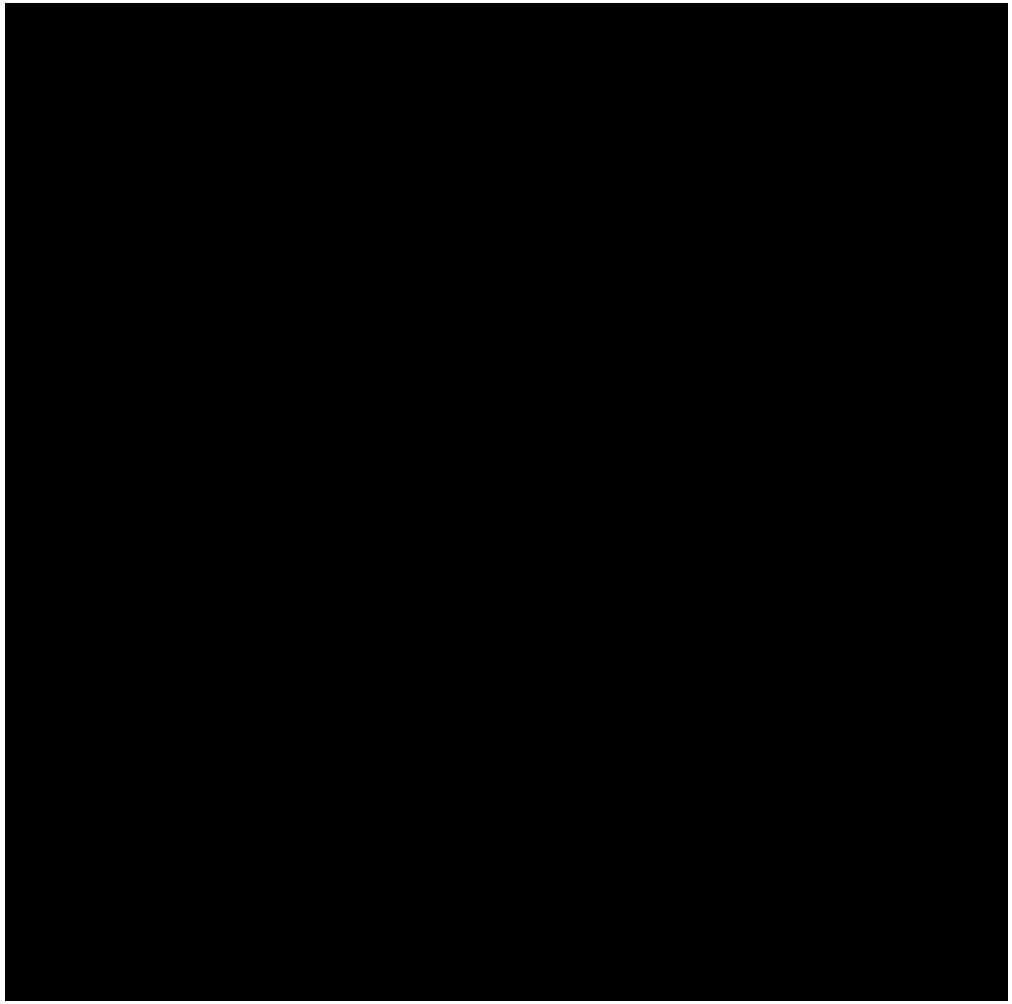
What about gaps?

- Probability of an insertion/deletion event (gap opening, G)
- Length distribution of insertions/deletions (gap extension, E)

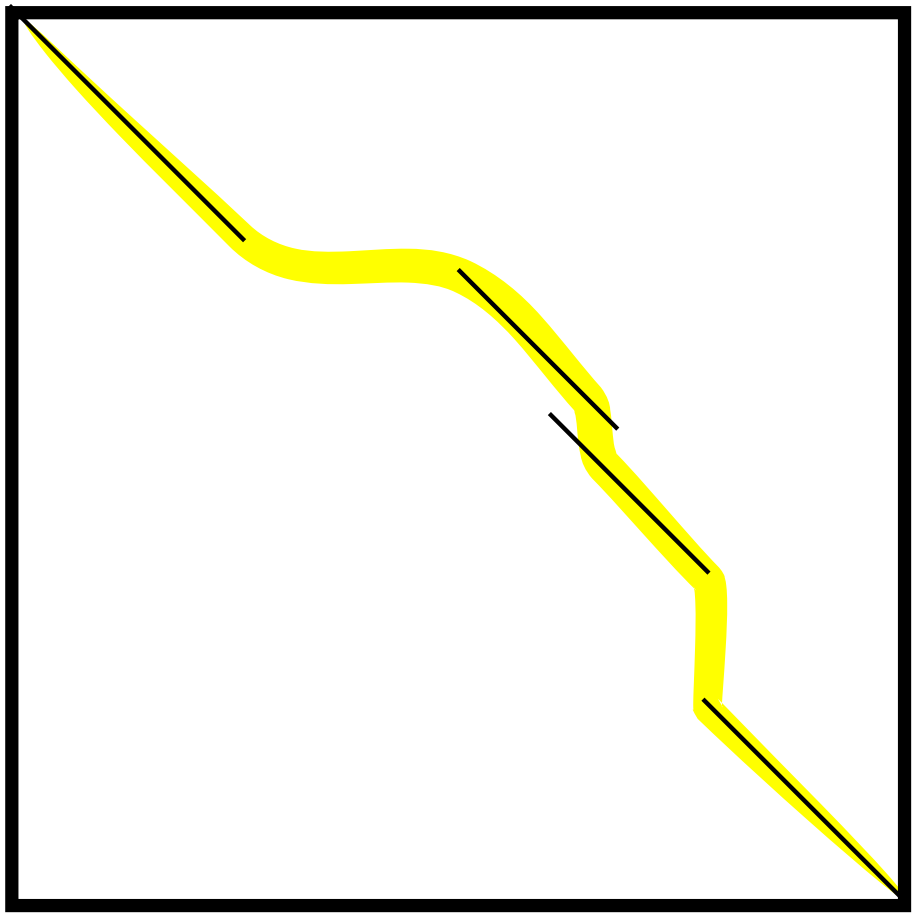
$$S_{gapped}(x; y) = S(x; y) + \sum_{i}^{gaps} (G + E * L_i) \tag{11}$$

The best alignment of any pair of subsequences is independent of the global alignment.

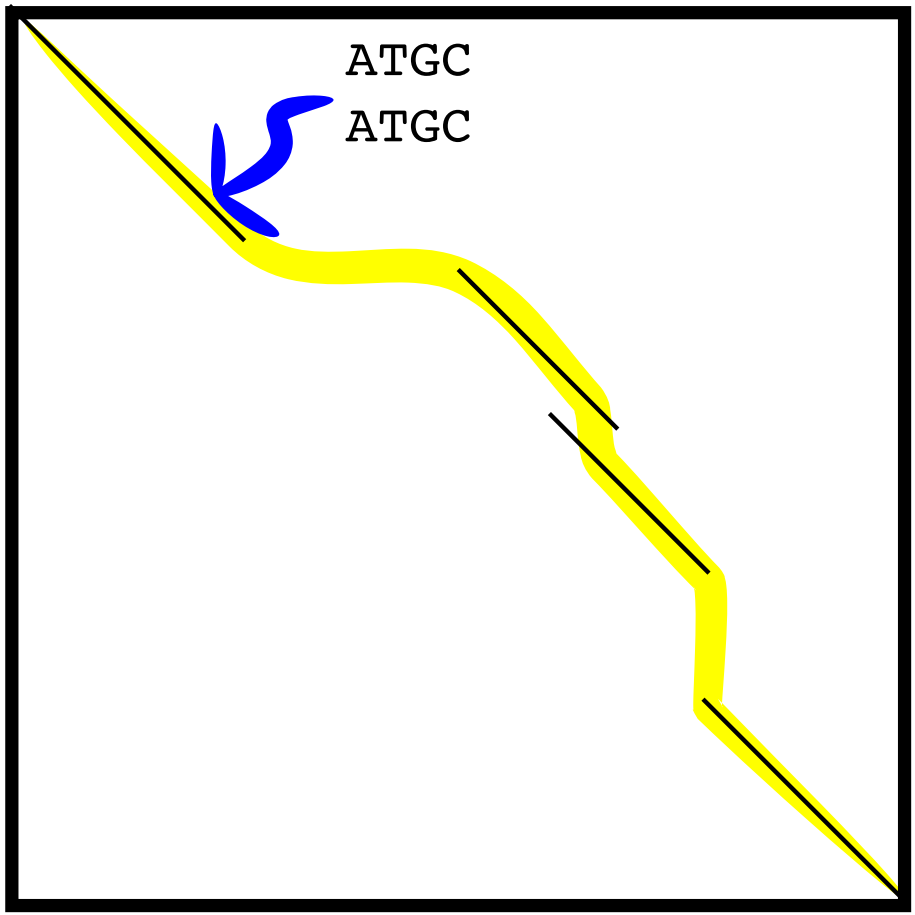
Dynamic Programming



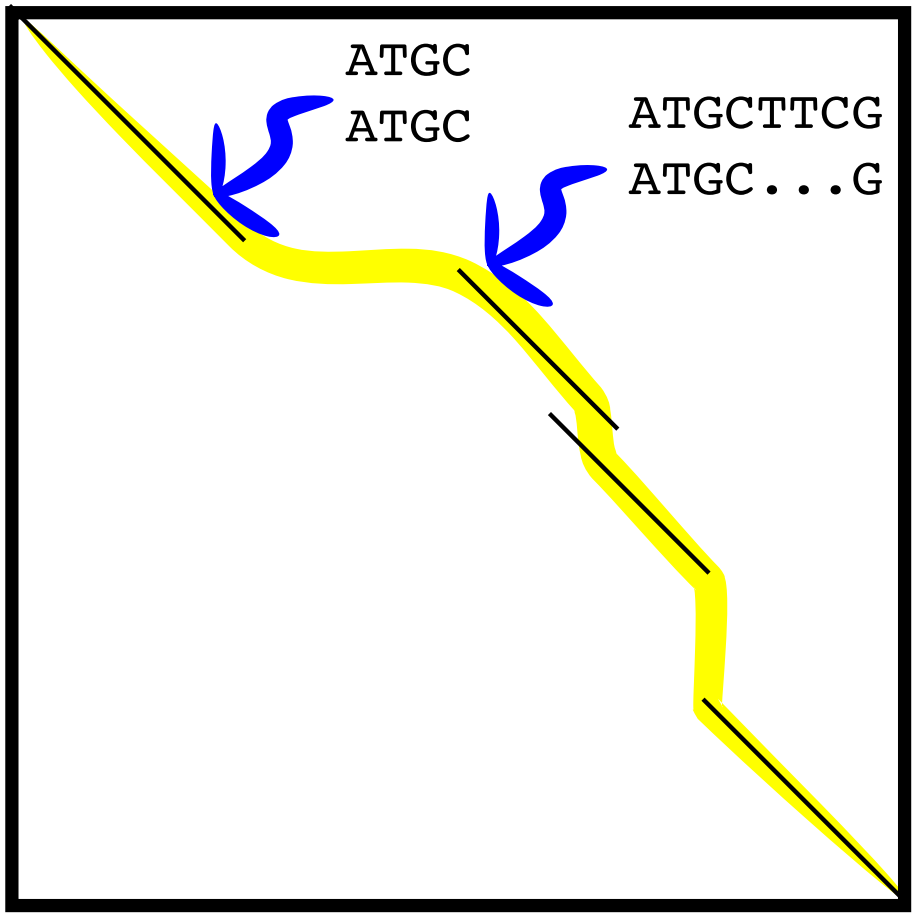
Needleman-Wunsch Global Alignment



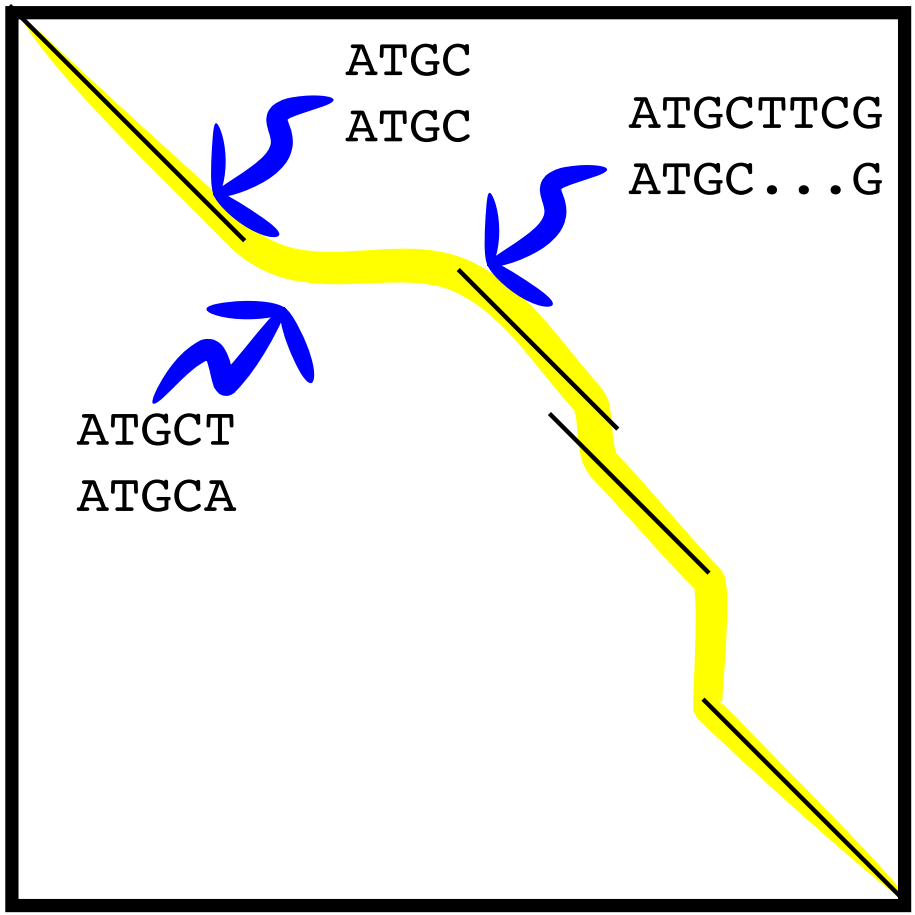
Needleman-Wunsch Global Alignment



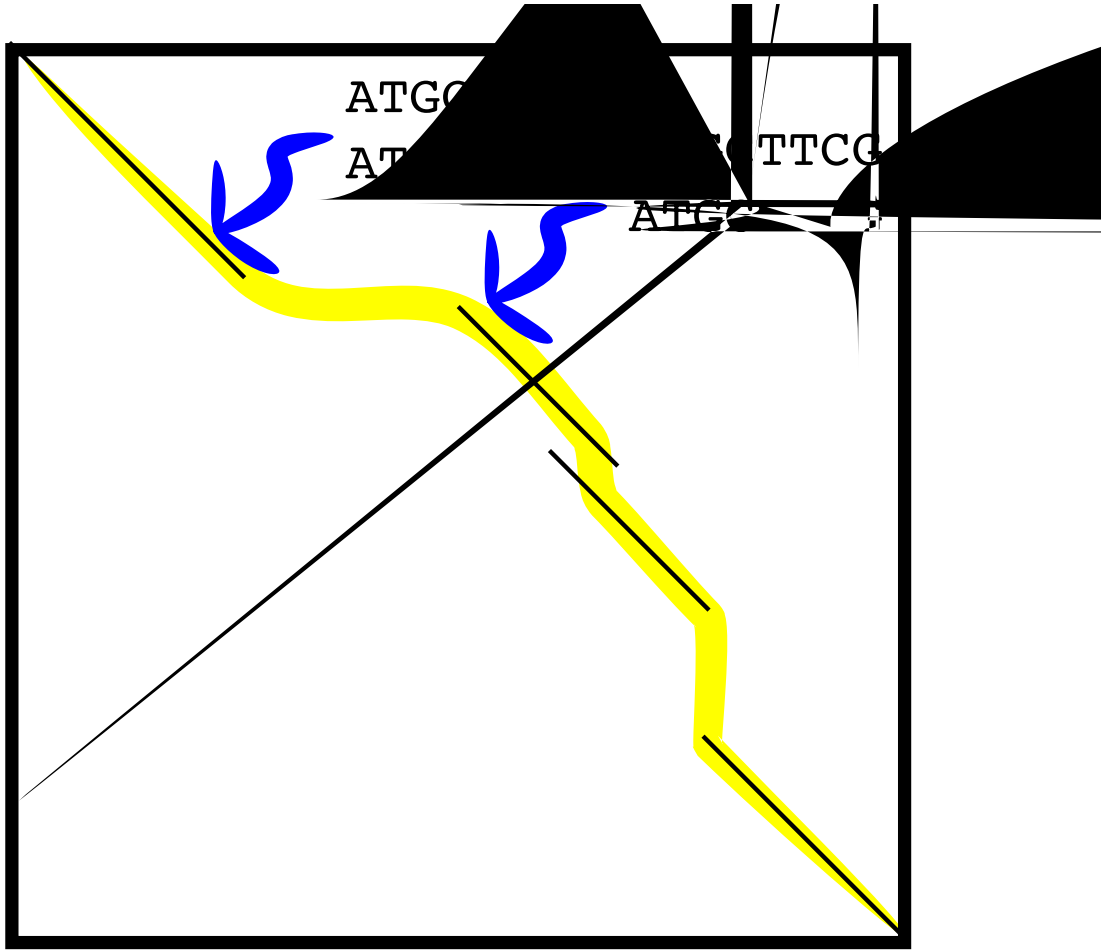
Needleman-Wunsch Global Alignment



Needleman-Wunsch Global Alignment



Needleman-Wunsch Global Alignment



Alignment speeds

- DOTTER: $O(n^2)$
- Exhaustive search: $\frac{2^{2n}}{\sqrt{n}}$
- Dynamic programming: $O(n^2)$ to $O(n^3)$

Homework

- Play with some of your favorite sequences in CLUSTALX
- Experiment with varying the scoring matrices and gap parameters
- Which sequences are easier or harder to align?