Multiple Sequence Alignment

by Ahmet Sacan

Images are from "Understanding Bioinformatics" and Wikipedia.

Multiple alignment

- Improve confidence in homology
- Correct an alignment
- Find conserved residues and regions
- Predict protein secondary/tertiary structures

Multiple sequence alignment

- Given k sequences, find an alignment that maximizes the alignment score.

```
VSLSCTGSSSNIGAGNHVKWYQQLPG
VTISCTGTSSNIGSITVNWYQQLPG
ATLVCLISDFYPGASVTVAWKADS
AALGCLVKDYFPEPVTVSWNSG
LTCLVKGFYPSDIAVEWESNG
```

```
VSLSCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTSSNIG--SITVNWYQQLPG
ATLVCLISDFYPGA-SVTVAWKADS--
AALGCLVKDYFPEP--VTVSWNSG---
--LTCLVKGFYPSD--IAVEWESNG--
```

Multiple sequence alignment

- Given k sequences, find an alignment that maximizes the alignment score.

```
VSLSCTGSSSNIGAGNHVKWYQQLPG
VTISCTGTSSNIGSITVNWYQQLPG
ATLVCLISDFYPGASVTVAWKADS
AALGCLVKDYFPEPVTVSWNSG
LTCLVKGFYPSDIAVEWESNG
```

```
VSLSCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTSSNIG--SITVNWYQQLPG
ATLVCLISDFYPGA-SVTVAWKADS--
AALGCLVKDYFPEP--VTVSWNSG---
--LTCLVKGFYPSD--IAVEWESNG--
```

Multiple Alignment Score

V
V
A
A
–

- Column score = Sum of all-pairs:

$$S_{VV} + S_{VA} + S_{VA} + S_{V-} + S_{VA} + S_{VA} + S_{V-} + S_{AA} + S_{A-} + S_{A-}$$

- Multiple Alignment score = Sum of all column scores

Multiple Alignment Score

- Pairwise sequence alignment score:

  $S_{ab} = S_{VV} + S_{ST} + S_{LI} + S_{SS} + \cdots$

- Multiple Alignment Score: sum of all pairwise alignments

$$S_{ab} + S_{ac} + S_{ad} + S_{ae}$$
$$+S_{bc} + S_{bd} + S_{be}$$
$$+S_{cd} + S_{ce} + S_{de}$$

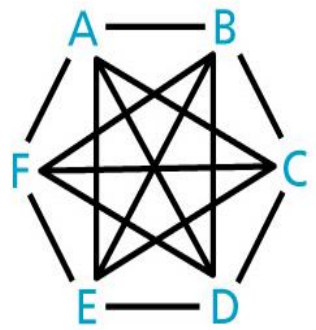
```
a: VSLSCTGSSSNIGAG-NHVKWYQQLPG
b: VTISCTGTSSNIG--SITVNWYQQLPG
c: ATLVCLISDFYPGA-SVTVAWKADS--
d: AALGCLVKDYFPEP--VTVSWNSG---
e: --LTCLVKGFYPSD--IAVEWESNG--
```

Multiple Alignment Score

- Sum of scores of pairwise alignments:

All-pairs

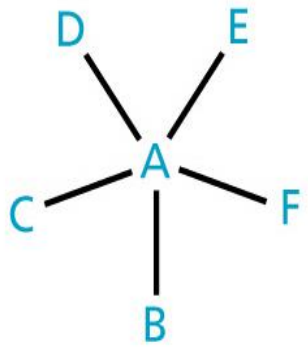
(C)



$$\text{score} = S_{AB} + S_{AC} + S_{AD} + S_{AE} + S_{AF} + S_{BC} + S_{BD} + S_{BE} + S_{BF} + S_{CD} + S_{CE} + S_{CF} + S_{DE} + S_{DF} + S_{EF}$$

Star

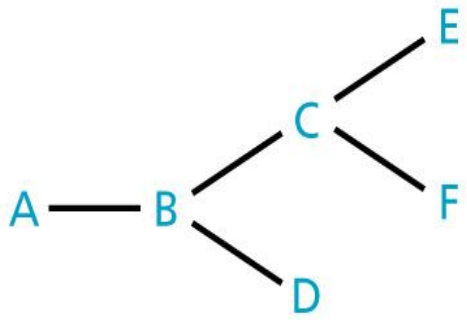
(B)



$$\text{score} = S_{AB} + S_{AC} + S_{AD} + S_{AE} + S_{AF}$$

Phylogenetic-tree

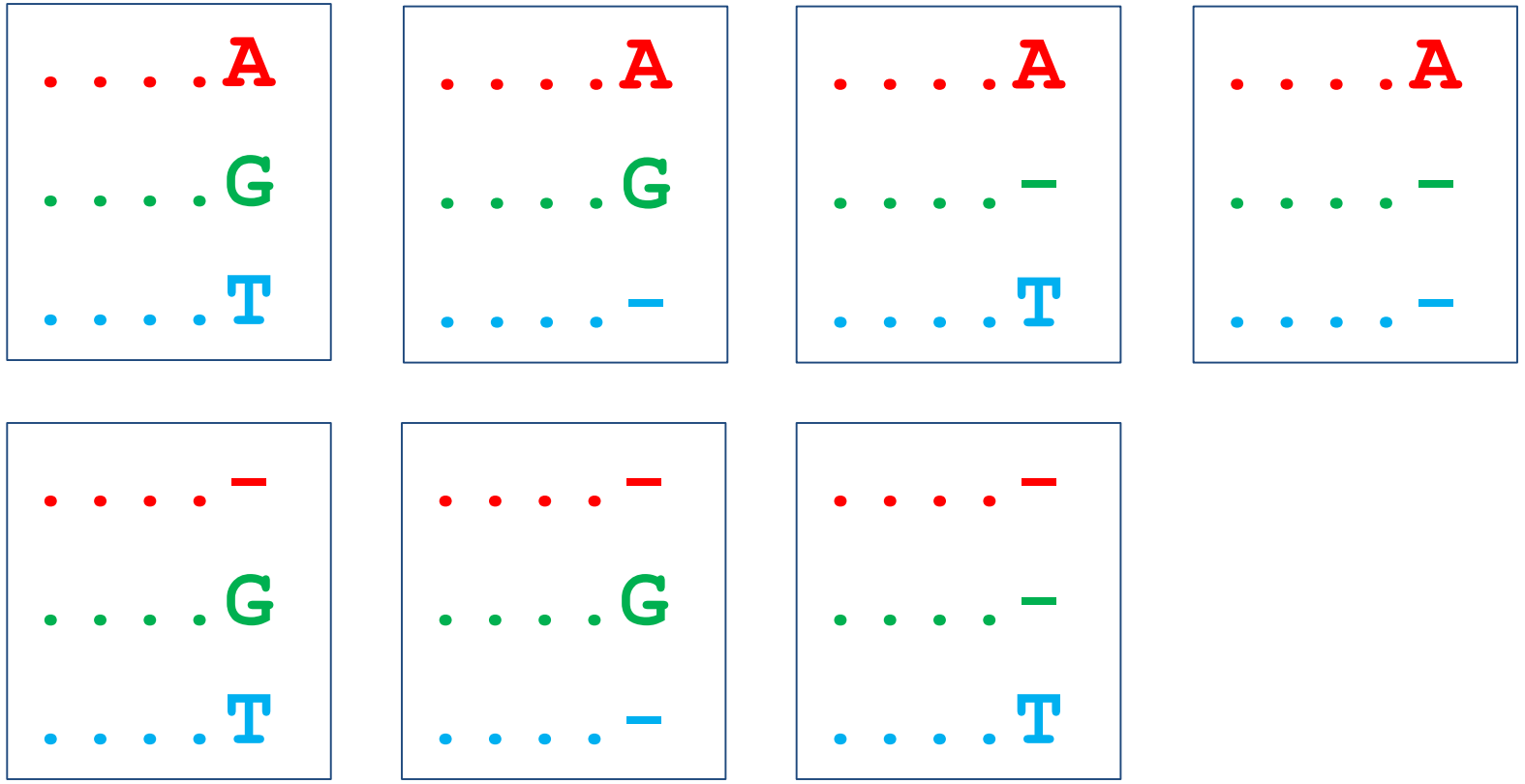
(A)



$$\text{score} = S_{AB} + S_{BC} + S_{BD} + S_{CE} + S_{CF}$$

Aligning 3 sequences

- The last column of the alignment could be one of:



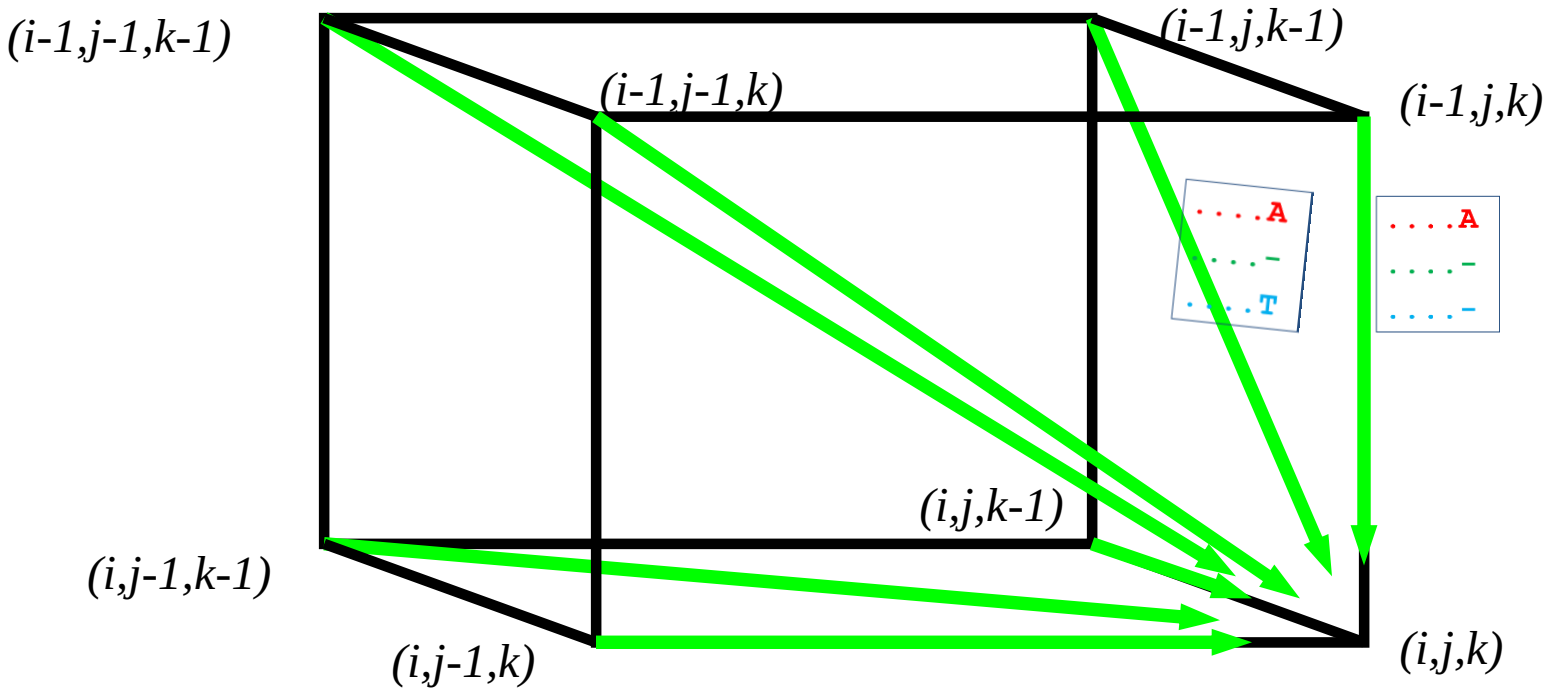
Dynamic Programming Multiple alignment

$$S_{i,j,k} = max \begin{cases} S_{i-1,j-1,k-1} + \delta(a_i, b_j, c_k) \\ S_{i,j-1,k-1} + \delta(\_, b_j, c_k) \\ S_{i-1,j,k-1} + \delta(a_i, \_, c_k) \\ S_{i-1,j-1,k} + \delta(a_i, b_j, \_) \\ S_{i,j,k-1} + \delta(\_, \_, c_k) \\ S_{i,j-1,k} + \delta(\_, b_j, \_) \\ S_{i-1,j,k} + \delta(a_i, \_, \_) \end{cases}$$

cube diagonal: no indels

face diagonal: one indel

edge diagonal: two indels



Dynamic Programming Multiple alignment: Computational Complexity

- For 3 sequences of length n, run time is:

$$O(7n^3)$$

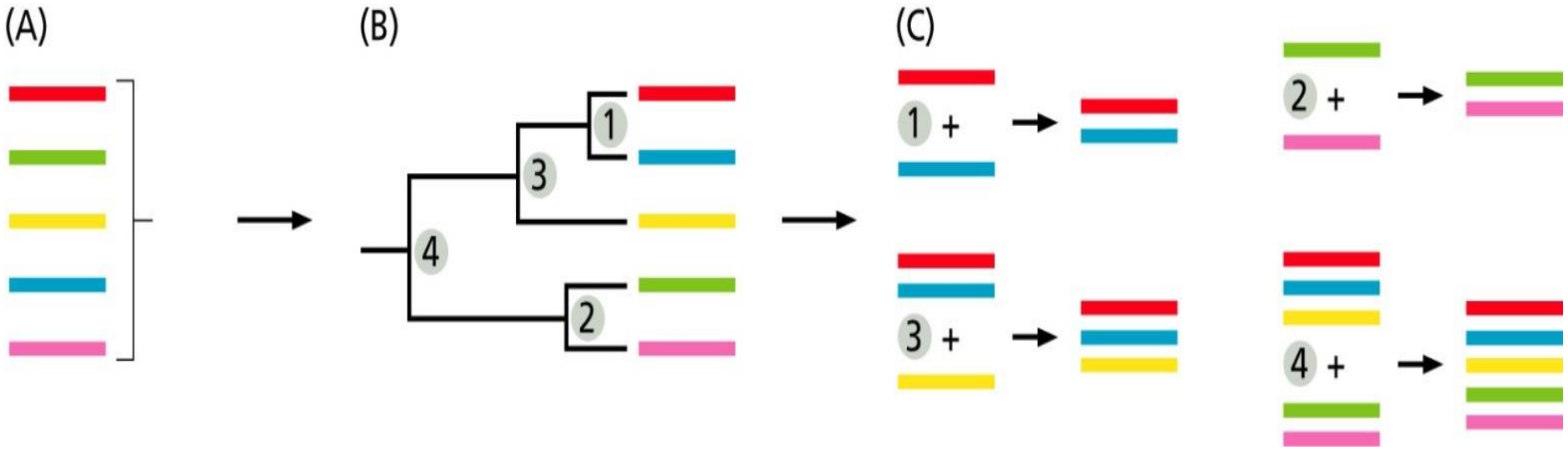
- For 4 sequences:

$$O(15n^4)$$

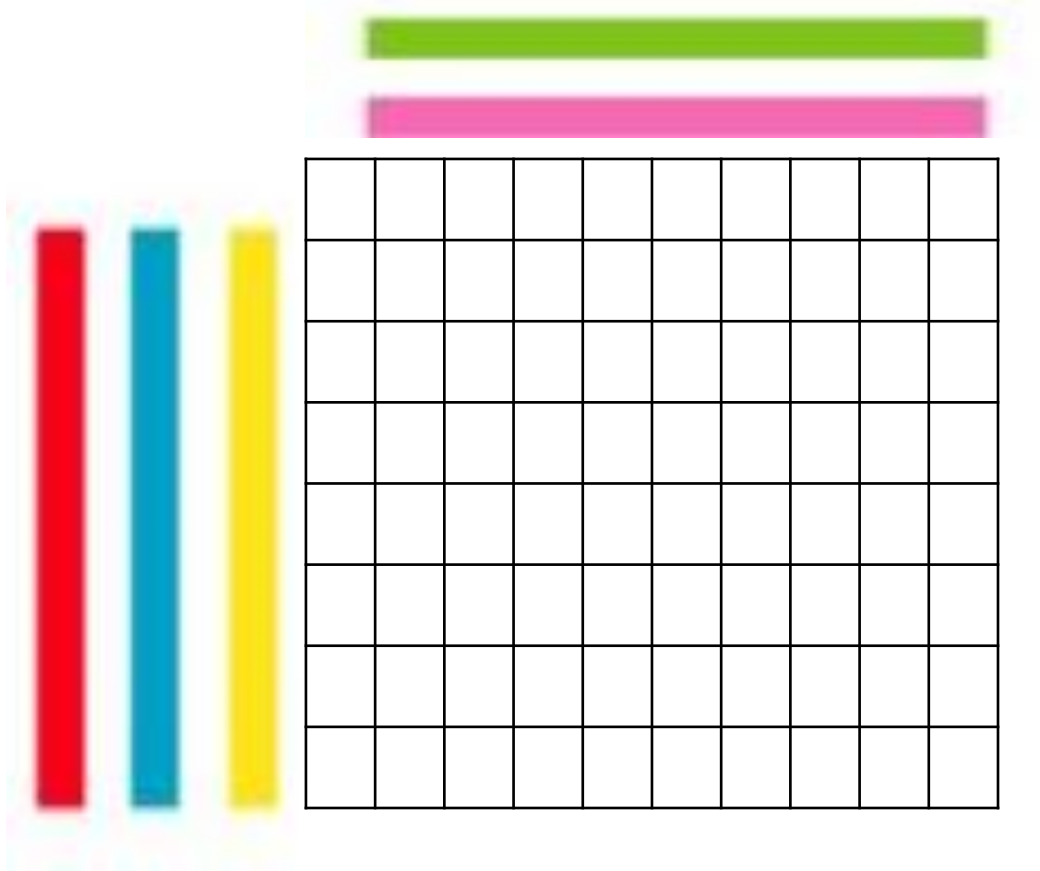
- For k sequences:
  - $O((2^k-1)n^k) = O(2^k n^k)$

Progressive alignment

- Do pairwise alignment at each step
  - Align a sequence to a sequence
  - Align a sequence to a multiple-alignment
  - Align a multiple alignment to a multiple alignment

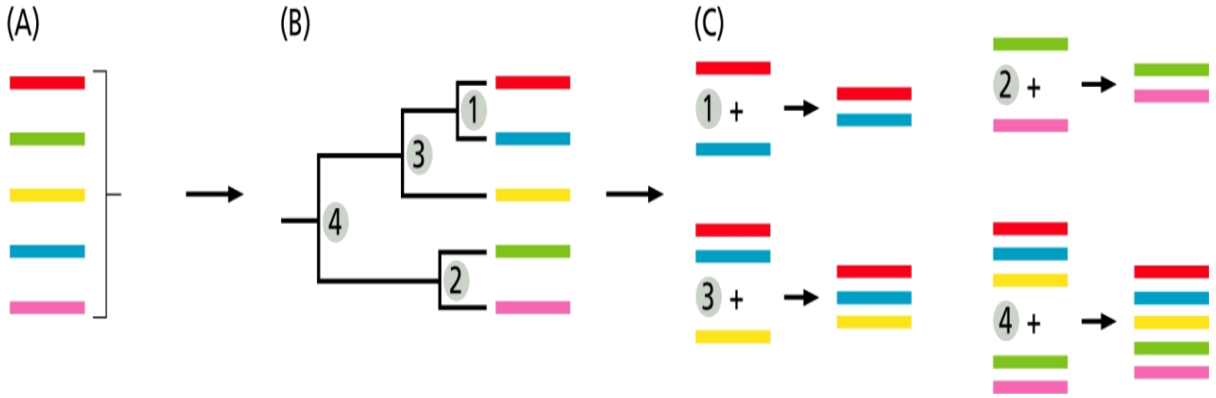


Dynamic Programming to Align Alignments



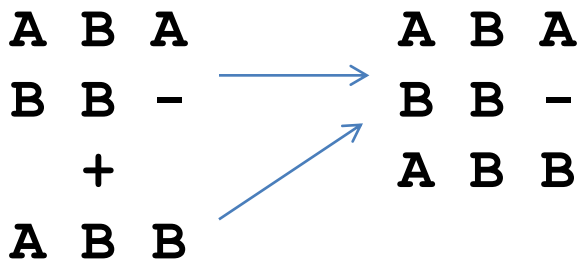
Progressive alignment: Computational Complexity

- k sequences, each with length n
- Each pairwise alignment: $O(n^2)$
- Building the phylogenetic tree
  - $O(k^2)$ pairwise comparisons
  - $O(k^2 n^2)$ time
  - Done once, to construct a phylogenetic tree
- Number of merge steps
  - $O(k)$ steps
  - $O(k\,n^2)$ time
- Overall time:
  - $O(k^2 n^2 + k n^2) = O(k^2 n^2)$

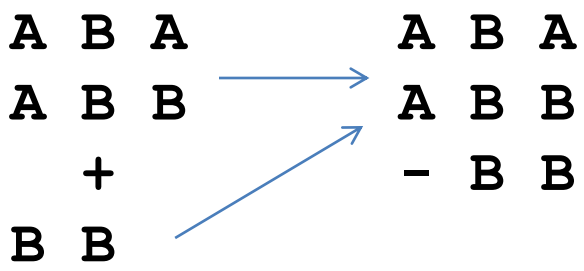


The order of alignment matters

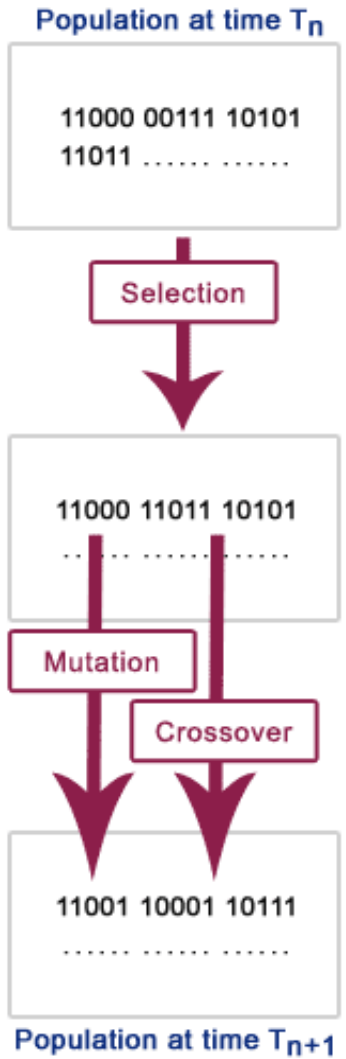
- Align ABA, BB, and ABB
- If we align ABA and BB, we may get:



- If we align ABA and ABB first, we don't have this problem:



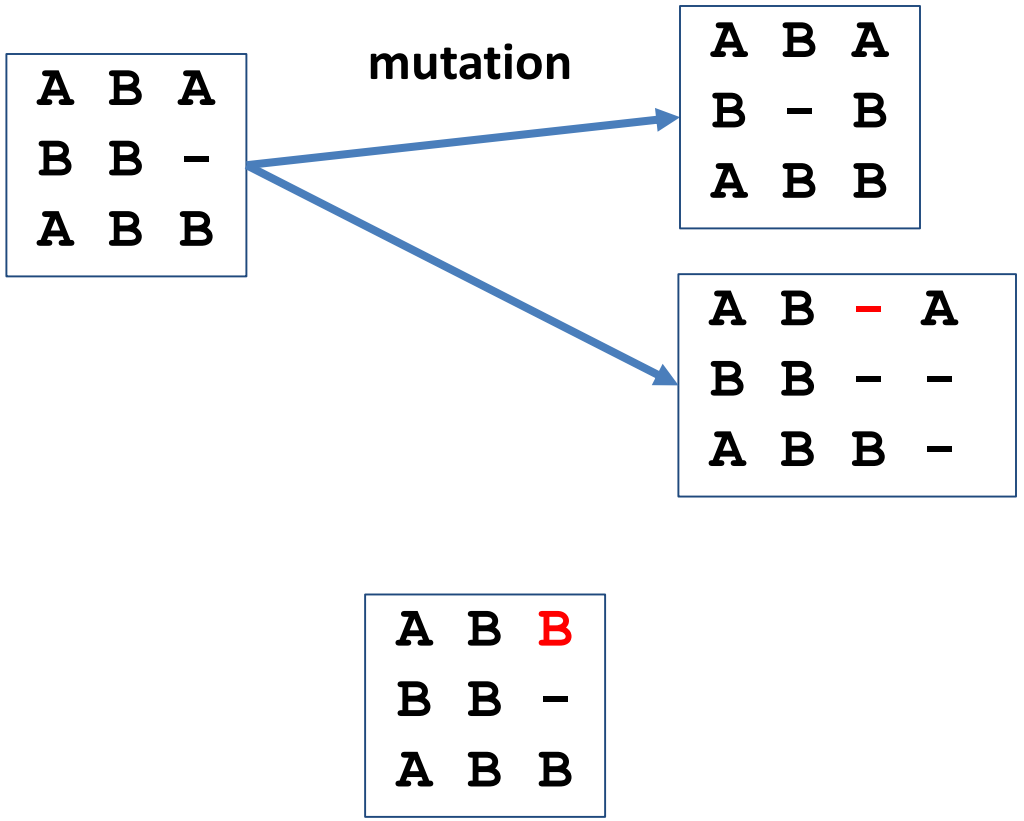
Alignment using Genetic algorithm



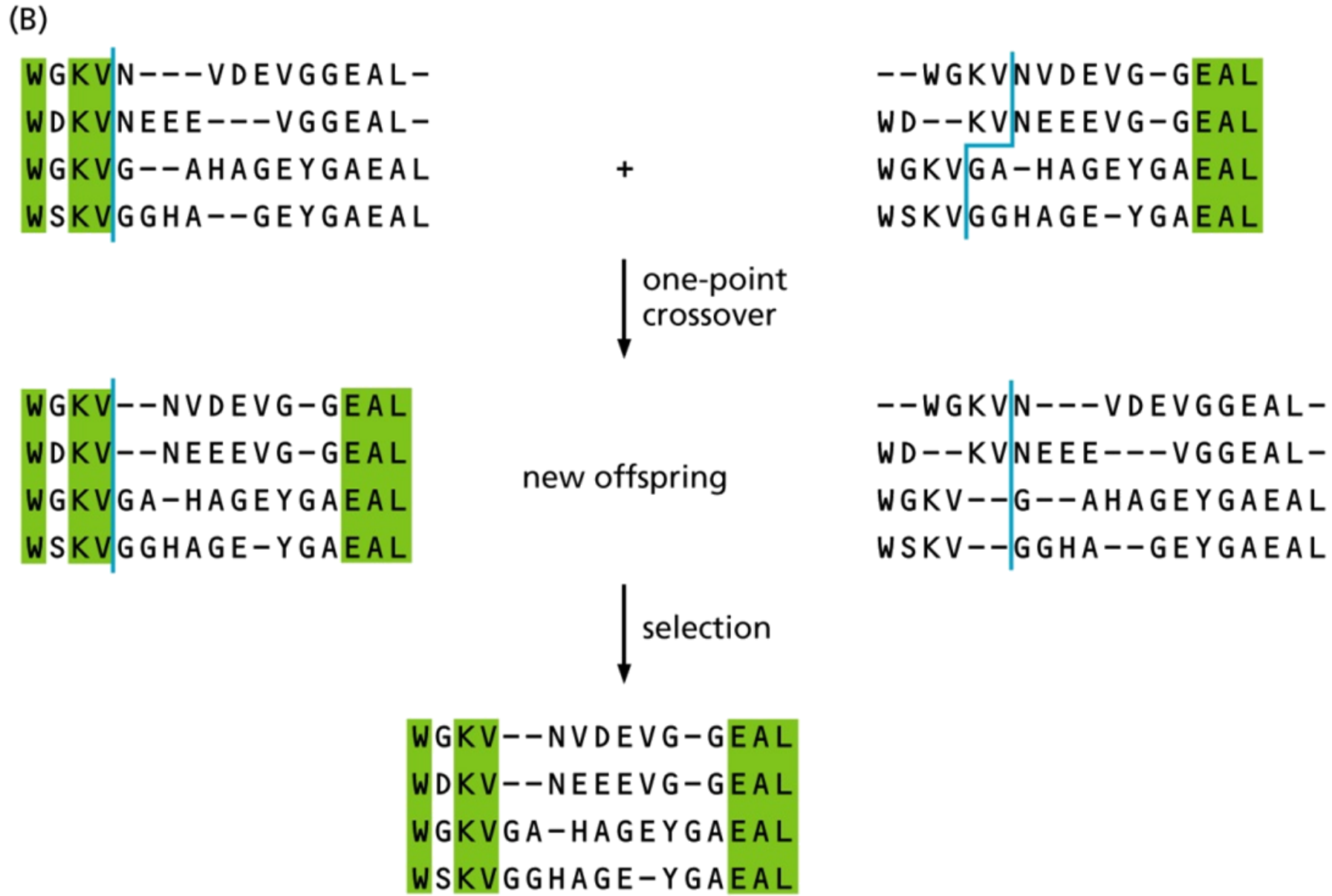
Fitness values:

F(11000) = 0.8
F(00111) = 0.2
F(10101) = 0.6
F(11011) = 0.7

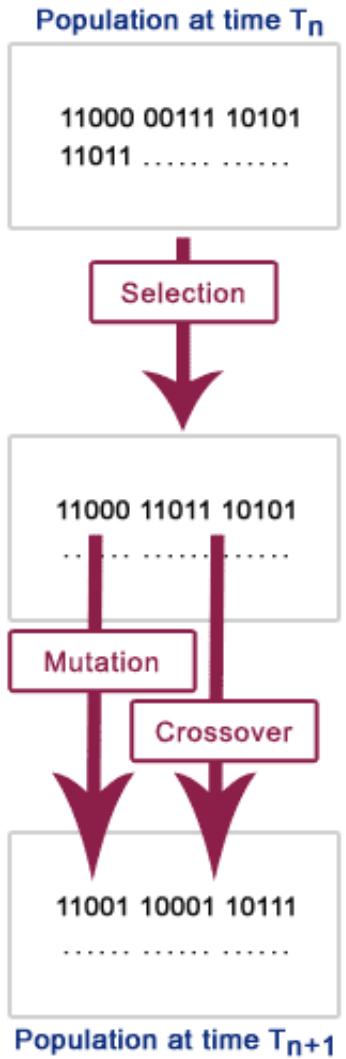
Alignment using Genetic algorithm



Alignment using Genetic algorithm



Alignment using Genetic algorithm



Fitness values:

F(11000) = 0.8
F(00111) = 0.2
F(10101) = 0.6
F(11011) = 0.7

Summary

- Multiple sequence alignment is more accurate than pairwise alignment
  - Used to create sequence profiles
- Multiple Alignment is a computationally difficult problem
  - Finding optimal solution is not feasible
- Heuristic approaches find an answer efficiently, but do not guarantee it is the best one
- Optimization methods can be employed to obtain better solutions.