Figures and Legends for

"Blueprint of the "<u>A</u>lignment <u>N</u>eighborhood <u>Ex</u>plorer" (ANEX) (tentatively named)"

by Kiyoshi Ezawa

(Finished on November 24th, 2015; added CC4 statement on August 14th, 2020)

Figures and Legends

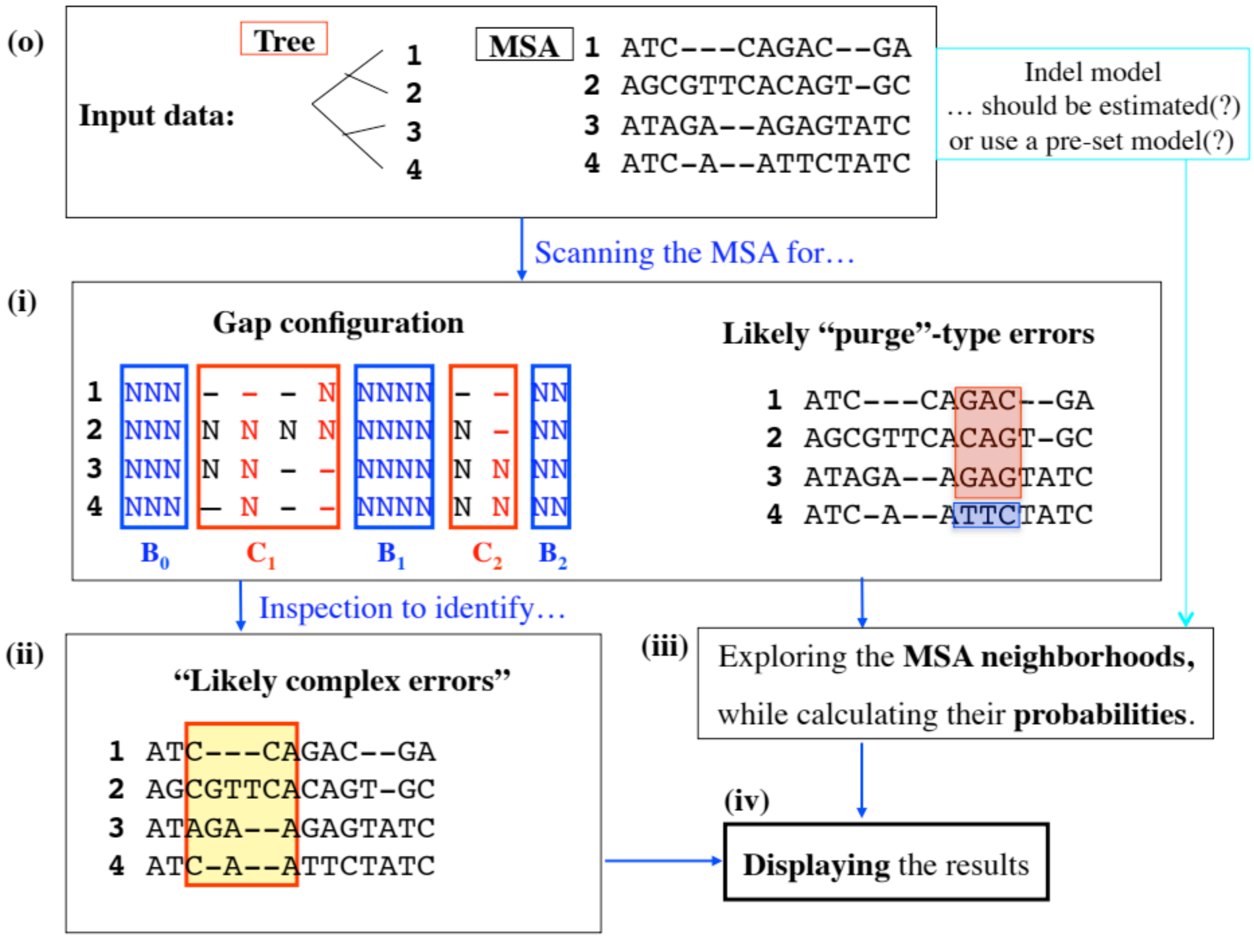


Figure 1. Flowchart explaining the overall workflow of the program.

In the gap configuration in (i), the 'N' and '-' represent the presence of a residue and a gap (*i.e.*, the absence of a residue), respectively.

And the blue and red rectangles enclose the gapless and gapped segments, respectively.

On the right of (i), the alignment of the regions shaded in red and blue indicates a "purge"-type error.

In (ii), the yellow-shaded region likely contains a "complex" error.

[I will rewrite the figure later, especially by fleshing out the steps (iii) and (iv).]

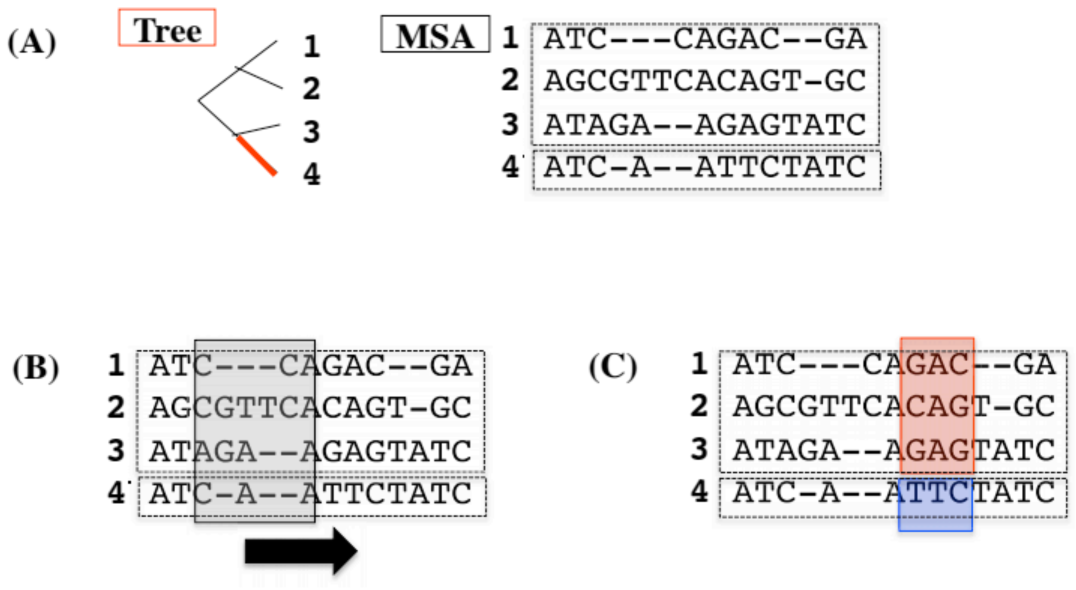


Figure 2. Sliding-window analysis to identify regions that likely harbor "purge"-type errors.

A. In this example, we search for possible "purge"s between sub-alignments enclosed by dashed rectangles, which are mutually separated by the red bold branch (on the left).

B. The window (shaded gray) contains a specified number of residue pairs, and slides from left to right (indicated by the black solid arrow).

C. This dual-colored window shows more predicted substitutions than expected from the branch length.

Thus, the program suspects that a "purge"-type error likely resulted in this region.

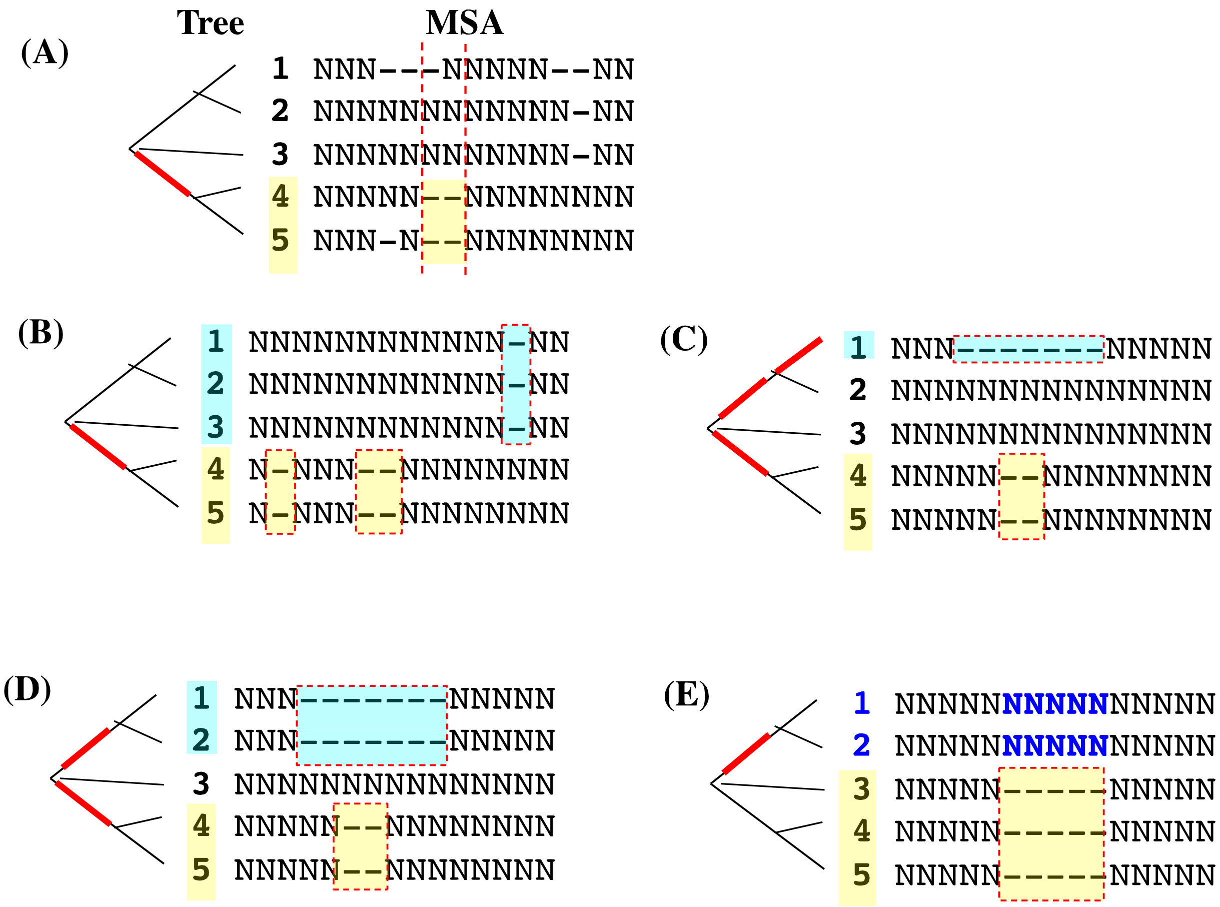


Figure 3. Gap-block, "isolated" gap-block, and sequence-block.

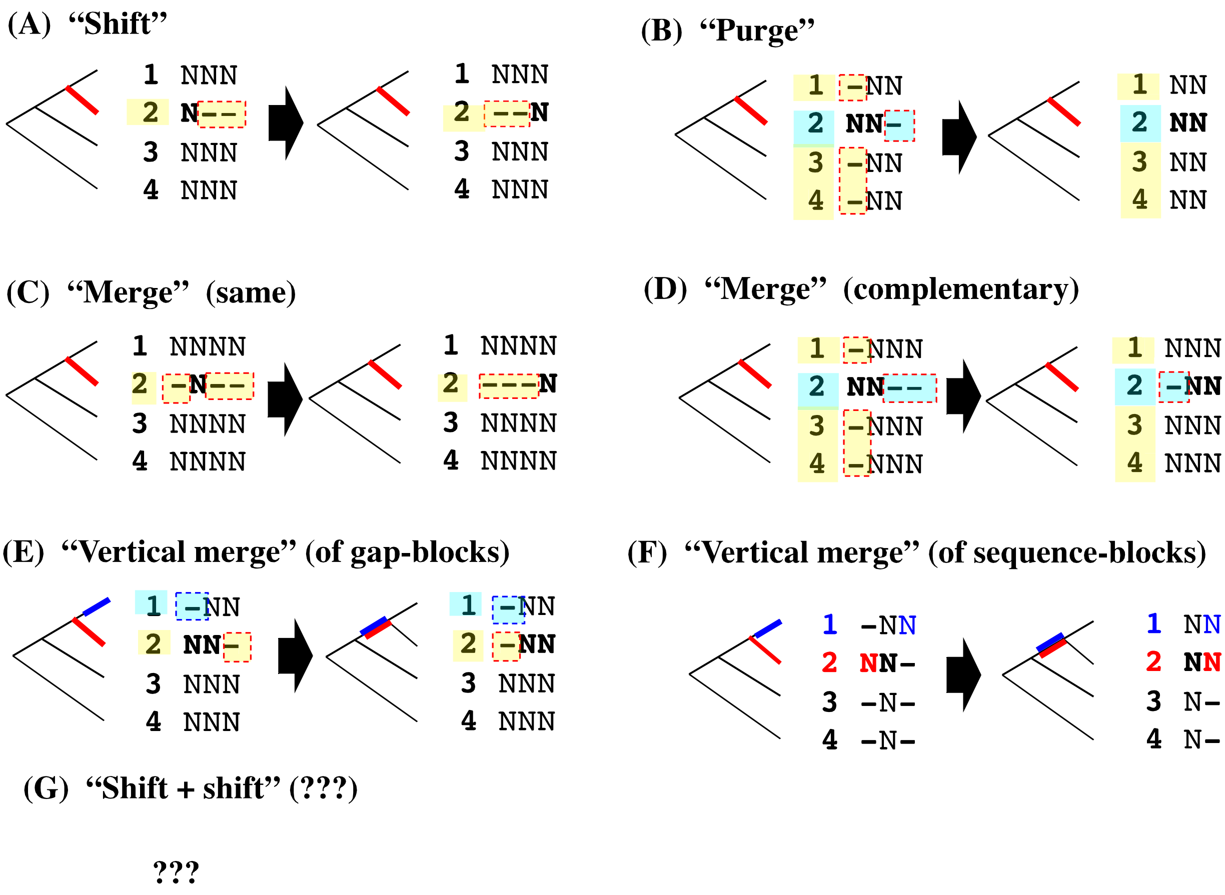
To focus on the gap-pattern, each residue was represented as 'N' regardless of its identity.

A. The yellow-shaded portion of the MSA is a gap-block. It is delimited by two inter-column positions (the vertical red dashed lines) and a branch in the tree (the red branch). (The identifiers of the involved sequences are also shaded in yellow.)

B, C. "Isolated" gap-blocks (shaded portions enclosed by dashed boxes). In panel B, the isolation is obvious because the horizontal positions of the gap-blocks do not overlap. In panel C, although the two gap-blocks overlap in their horizontal positions, they are separated with each other by three branches (in red).

D. A pair of "non-isolated" gap-blocks, which horizontally overlap and are separated by only two branches (in red).

E. A sequence-block (in blue), which is the complement of a gap-block (shaded in yellow).



(A) “Shift”

(B) “Purge”

(C) “Merge” (same)

(D) “Merge” (complementary)

(E) “Vertical merge” (of gap-blocks)

(F) “Vertical merge” (of sequence-blocks)

(G) “Shift + shift” (???)

Figure 4. Elementary moves that program will attempt.

The bold “N”s represent the residues that moved.

A. A “shift” of a single gap-block (shaded portion enclosed by the dashed box).

B. A “purge” of two gap-blocks affecting the complementary sets of sequences. (Its reverse is an “ex-nihilo.”)

C. A “merge” of two gap-blocks affecting the same set of sequences. (Its reverse is a “split” of a gap-block.)

D. A “merge” of gap-blocks affecting the complementary sets of sequences. (Its reverse is a “split” of a gap-block.)

E. A “vertical-merge” of gap-blocks. (Its reverse is a “vertical-split” of a gap-block.)

F. A “vertical-merge” of sequence-blocks. (Its reverse is a “vertical-split” of a sequence-block.)

G. “Shift + Shift” …???

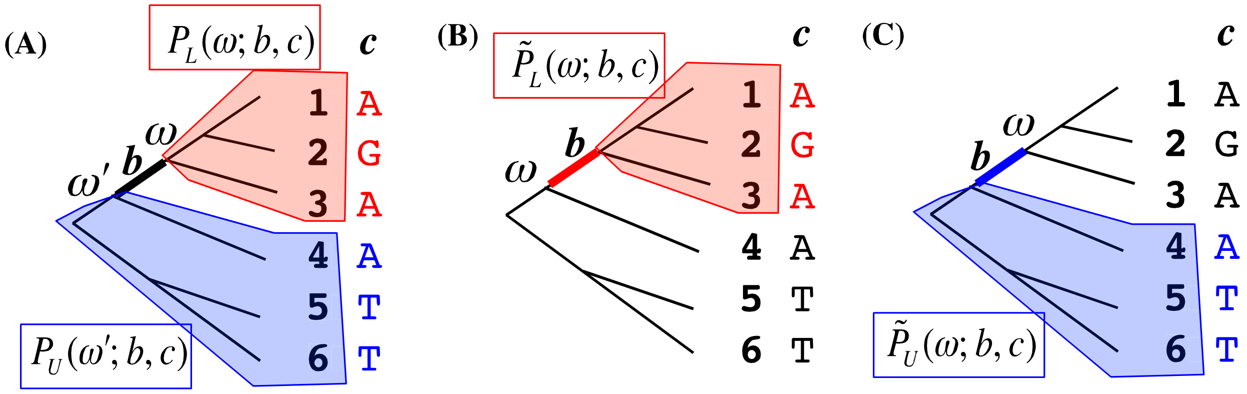


Figure 5. Four sets of important probabilities, assigned to each branch-column combination, for fast calculation of MSA residue-pattern probabilities.

A. Probabilities, $P_L(\omega; b, c)$ and $P_U(\omega'; b, c)$, for the residue configurations of two complementary sequence sets in a column (c). The branch b (bold) separates the sequence sets. The portions of the tree yielding $P_L(\omega; b, c)$ and $P_U(\omega'; b, c)$ are shaded in red and blue, respectively. And, in column c, the color of each residue indicates which of the probabilities it contributes to.

B. The extension of $P_L(\omega; b, c)$.

C. The extension of $P_U(\omega'; b, c)$.

In each panel, the numbers assigned to the external nodes also specify the sequences in the MSA.

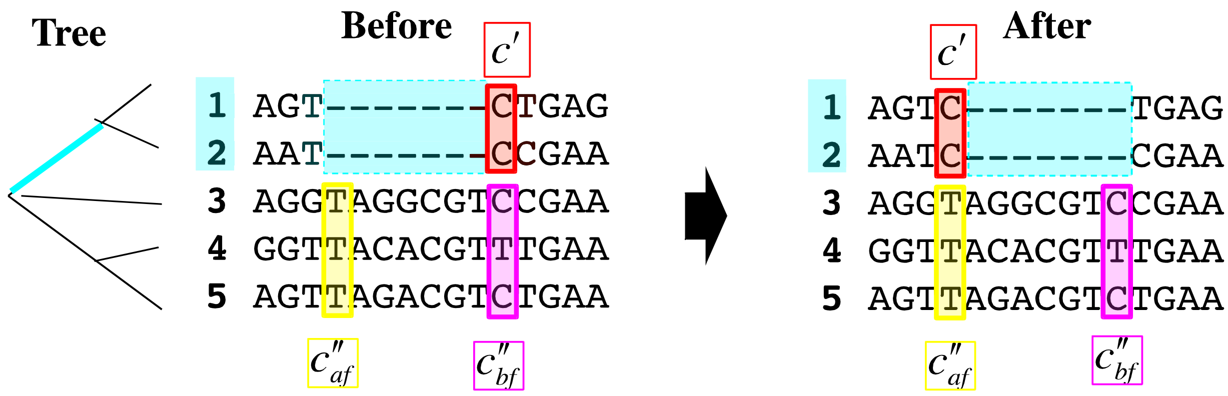


Figure 6. Change in residue component of MSA probability when gap-block shifts.

The gap-block in question and the sequences it affects are shaded in cyan.

The red shaded rectangle encloses "semi-column" c'.

The magenta and yellow shaded rectangles enclose "semi-columns" c''_{bf} and c''_{bf}, respectively.

The cyan branch phylogenetically delimits the gap-block.

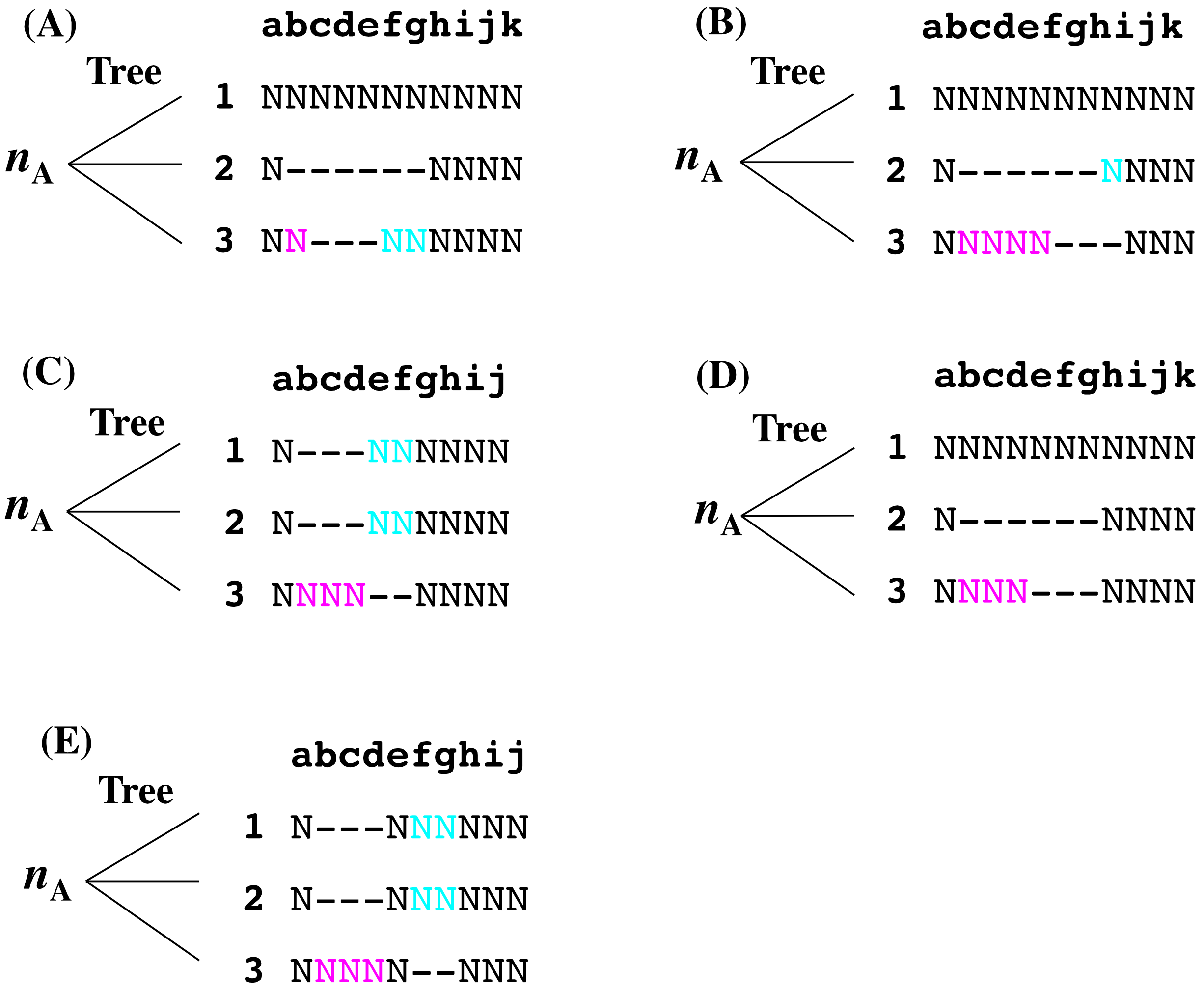


Figure 7. Typical examples of gap-configurations of 3 sequences connected via 3-OTU tree.

A. A short gap in the 3rd OTU (labeled “3”) is nested in a long gap in the 2nd OUT (“2”).

B. Gaps in two OTUs overlap each other in a non-nested manner.

C. Two non-homologous sequence-blocks neighboring each other.

D. This pattern is considered as topologically the same as in panel A but not as in panel B, because of the (parsimonious) indel histories that can create the patterns.

E. This pattern is considered as topologically different from that in panel C.

In each panel, the “ n_A ” denotes the “most recent common ancestor (MRCA)” of the three OTUs, and the lower-case letters above the alignment represent the ancestry indices of the sites.

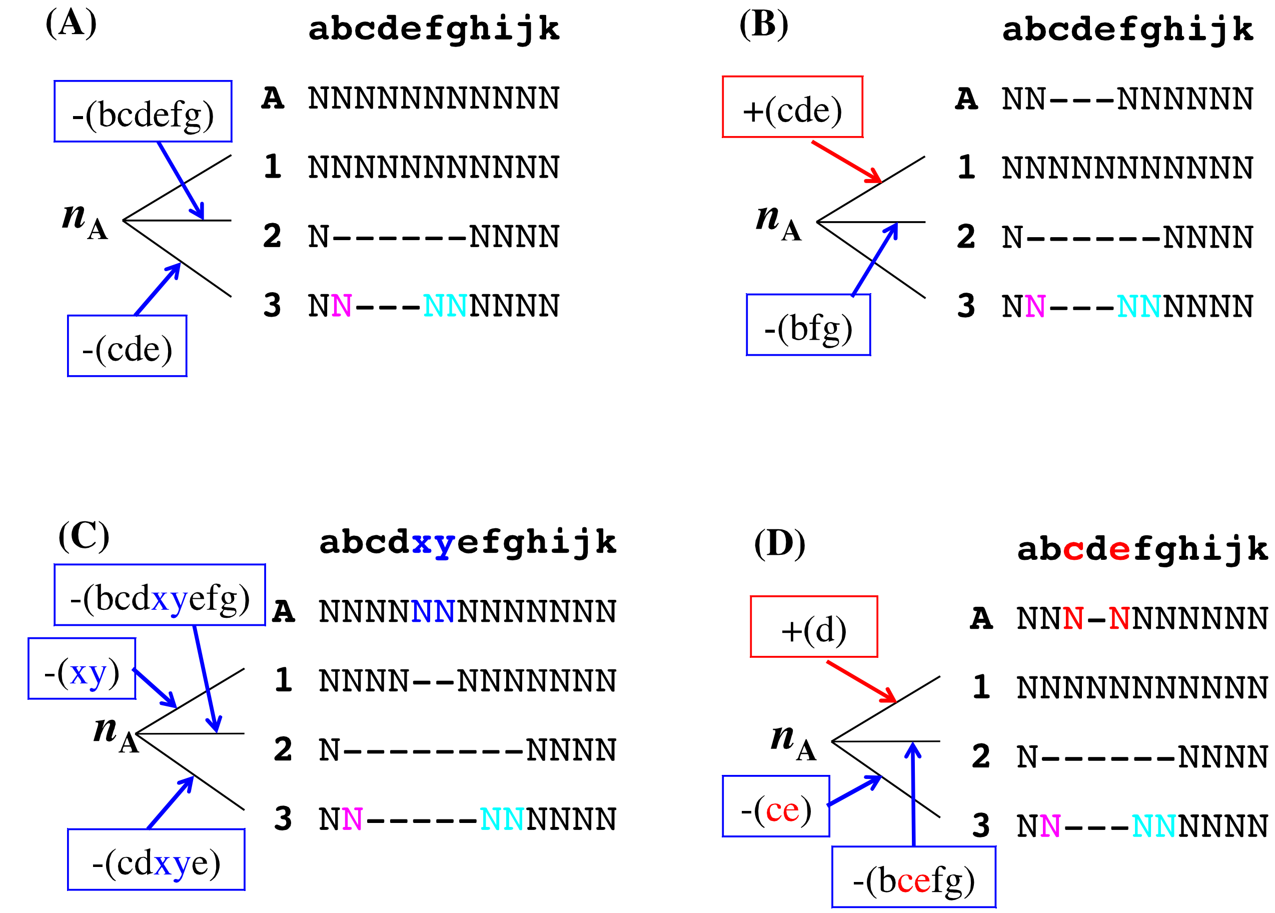


Figure 8. Indel histories that can create pattern A in Figure 7.

Panels A and **B** show the parsimonious histories.

Panels **C** and **D** show some next-to-parsimonious histories.

The 'A' on the top-left corner of the alignment indicates the ancestral sequence (at node n_A).

"-(abc)" in a blue rectangle shows that the sites a, b and c were deleted along the branch it points, and "+(def)" in a red rectangle shows that the sites d, e and f were inserted.

Some residues and ancestries were colored in order to facilitate the comparisons among the histories.

NOTE: Here we omitted all next-to-parsimonious histories in each of which two or more indels occur along a branch.

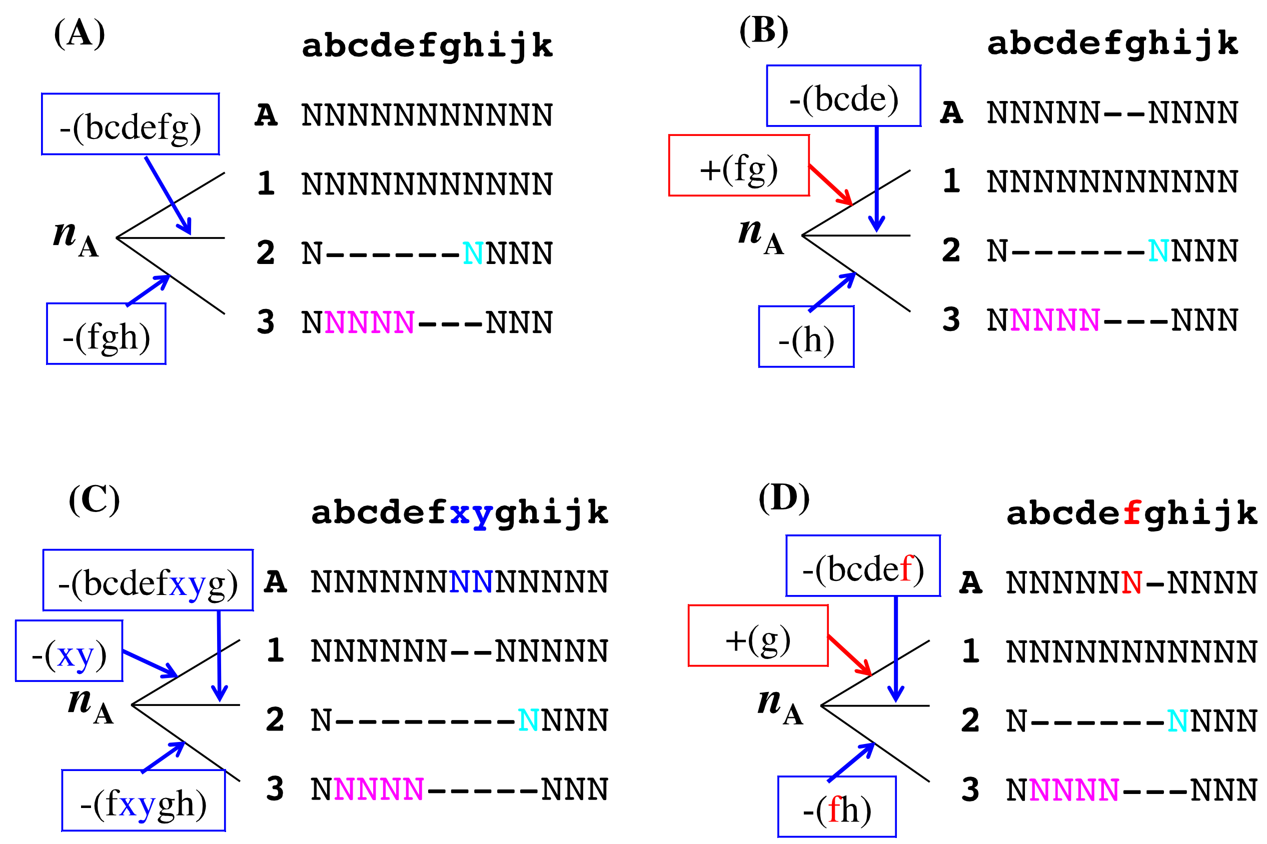


Figure 9. Indel histories that can create patten B in Figure 7.

Panel **A** shows the parsimonious indel history.

Panels **B, C** and **D** show some next-to-parsimonious histories.

Notations are the same as that for Figure 8.

NOTE: Here, again, we omitted the next-to-parsimonious histories in each of which two indels occur along a branch.

NOTE2: Panel B is actually the history created by concatenating the column-wise Dollo parsimonious indel histories (Farris 1977; Ezawa, Graur and Landan, Part III).

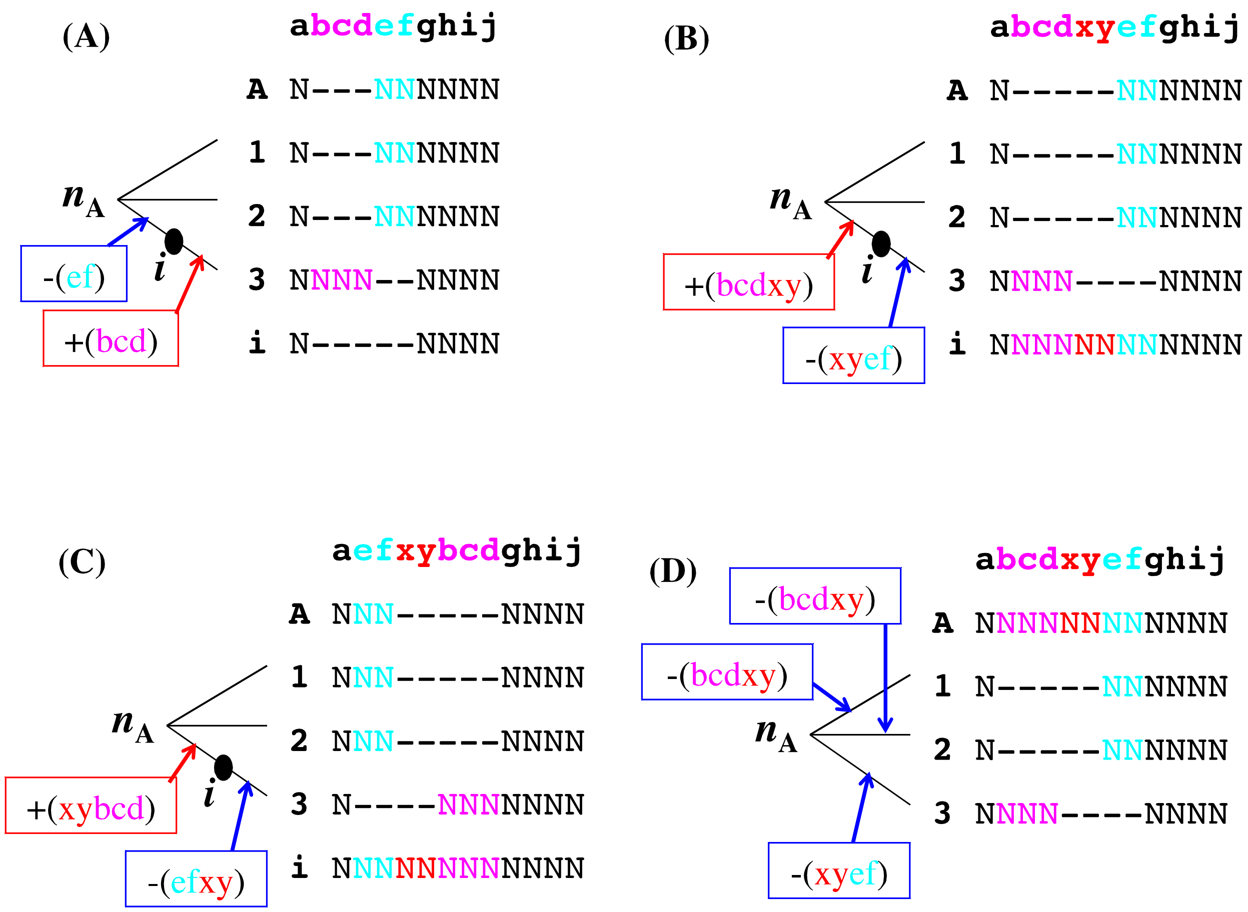


Figure 10. Indel histories that can create pattern C in Figure 7.

Panels **A**, **B**, and **C** show the three types of parsimonious indel histories.

Panel **D** shows a next-to-parsimonious indel history, which was derived from panel B via a "branching" operation (Ezawa, Graur and Landan, Part III).

(Note that a different history could be derived from panel C in a similar way.)

The notations are basically the same as in Figure 8. In addition, the sequence labelled "*i*" at the bottom of each MSA (in panels A, B and C) is the "intermediate" state at the point marked with a solid circle.

NOTE: The MSA of the extant sequences in panel C is equivalent to that in panel A (and in panel B).

NOTE2: We omitted next-to-parsimonious histories in each of which 3 indels occur along a branch.

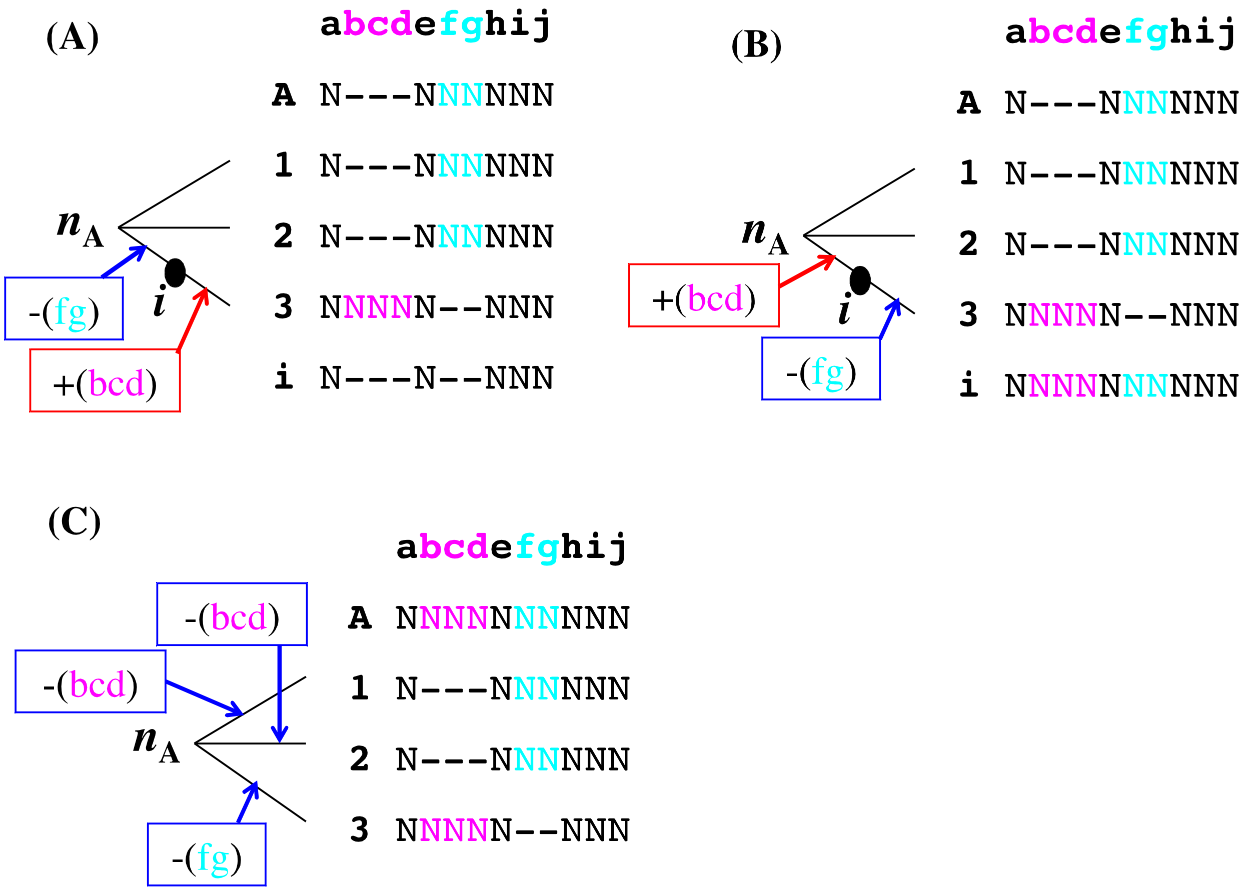


Figure 11. Indel histories that can create pattern in Figure 7E.

Panels **A** and **B** show the two parsimonious histories.

Panel **C** shows a next-to-parsimonious history. (It was obtained by applying a “branching” operation to panel B.)

The notations are basically the same as in Figure 8.

NOTE: We omitted those next-to-parsimonious histories each of which requires more than 2 indels in a branch.

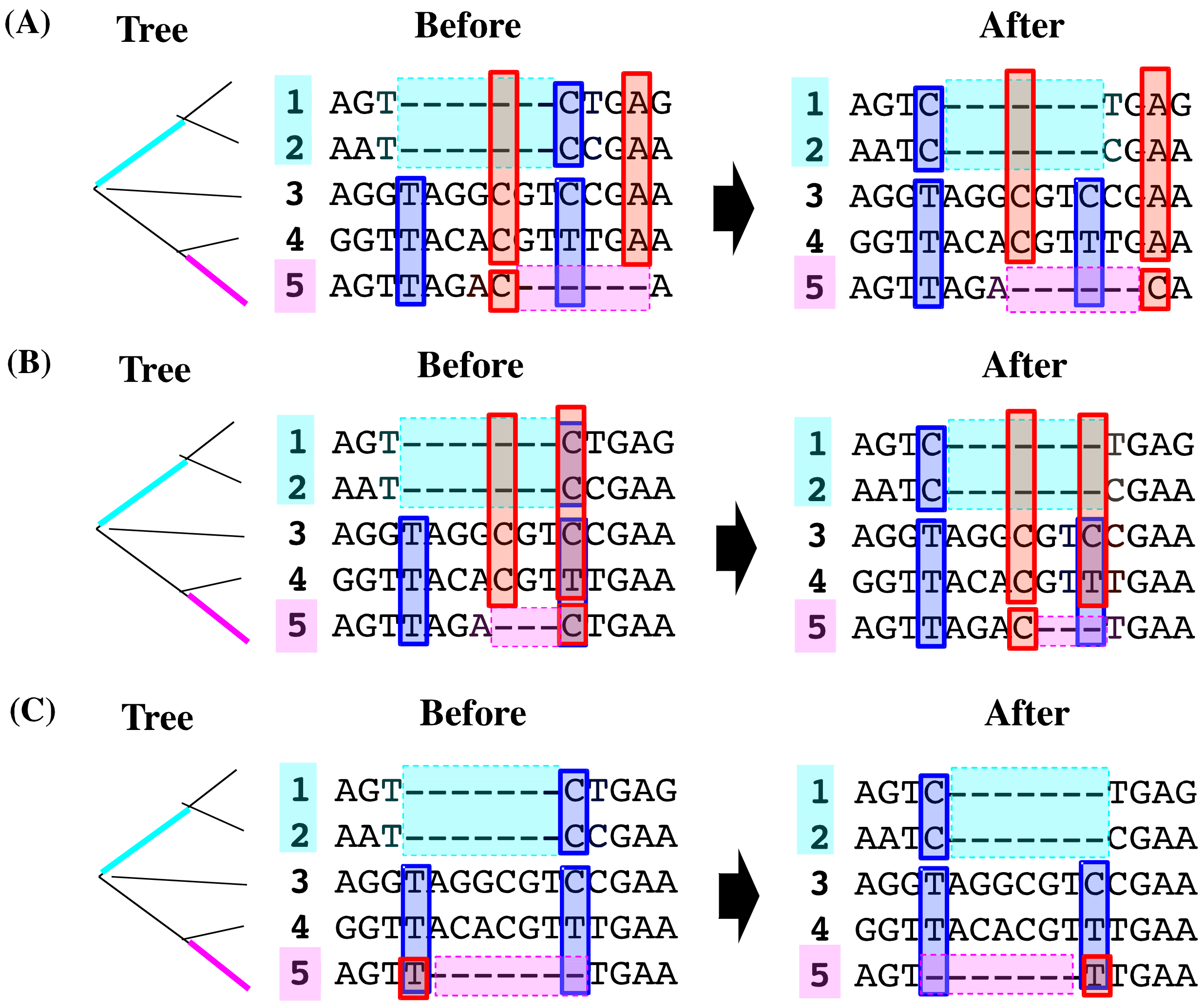


Figure 12. Effects of simultaneous shifts of two isolated gap-blocks on substitution component of MSA probability.

A. The shifts affect different columns independently.

B. The shifts affect the same column simultaneously.

C. The shifts affect the same pair of columns simultaneously.

In each column, shifts of the gap-blocks shaded in cyan and magenta affect the (semi-)columns shaded in blue and red, respectively.

The colored branches in the tree phylogenetically delimits the gap-blocks shaded in the same color.

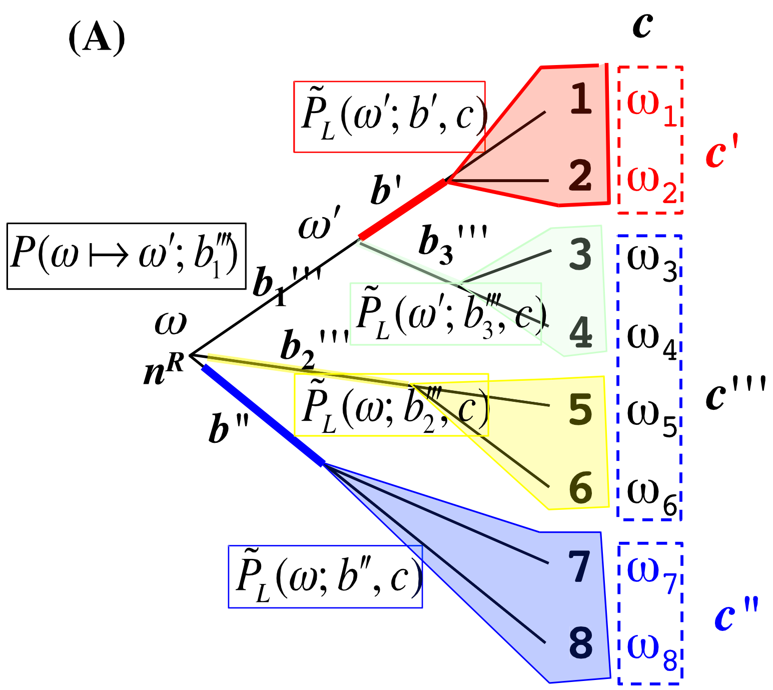

(A)
c
$\tilde{P}_L(\omega'; b', c)$
1
2
ω1
ω2
c'
b'
ω'
$P(\omega \mapsto \omega'; b_1''')$
b_3'''
3
4
ω3
ω4
$\tilde{P}_L(\omega'; b_3''', c)$
b_1'''
ω
n^R
b_2'''
$\tilde{P}_L(\omega; b_2''', c)$
5
6
ω5
ω6
c'''
b''
7
8
ω7
ω8
c''
$\tilde{P}_L(\omega; b'', c)$

Figure 13. Model situation for examining whether effects of shifts of two "isolated" gap-blocks are nearly independent or not.

Each portion of the tree (shaded in respective color) is assigned a building-block probability (enclosed by a rectangle of the same color).

The column (c) was divided into three parts: c', c'' and c'''.

Branches b' and b'' delimit c' and c'', respectively. And branches b_i''' (i =1,2,3) are important factors determining c'''.

The ω and ω' at both ends of branch b_1''' are the residue states over which the probabilities will be summed.

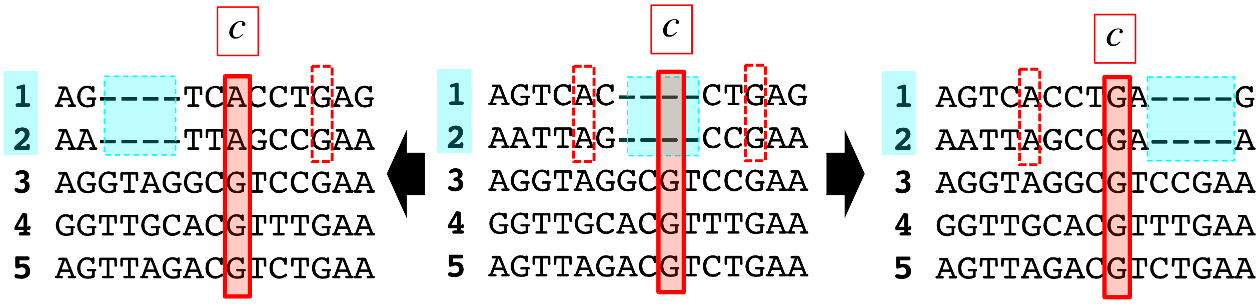


Figure 14. Possible residue configurations of each column caused by shifts of single gap-block. The shifts of a single gap-block (cyan-shaded) give rise to at most 3 possible configurations in each column (red-shaded), because they are "rigid" moves of the gap-block.

The red dotted rectangles enclose the original "ingredients" of the columns that resulted from the shifts.

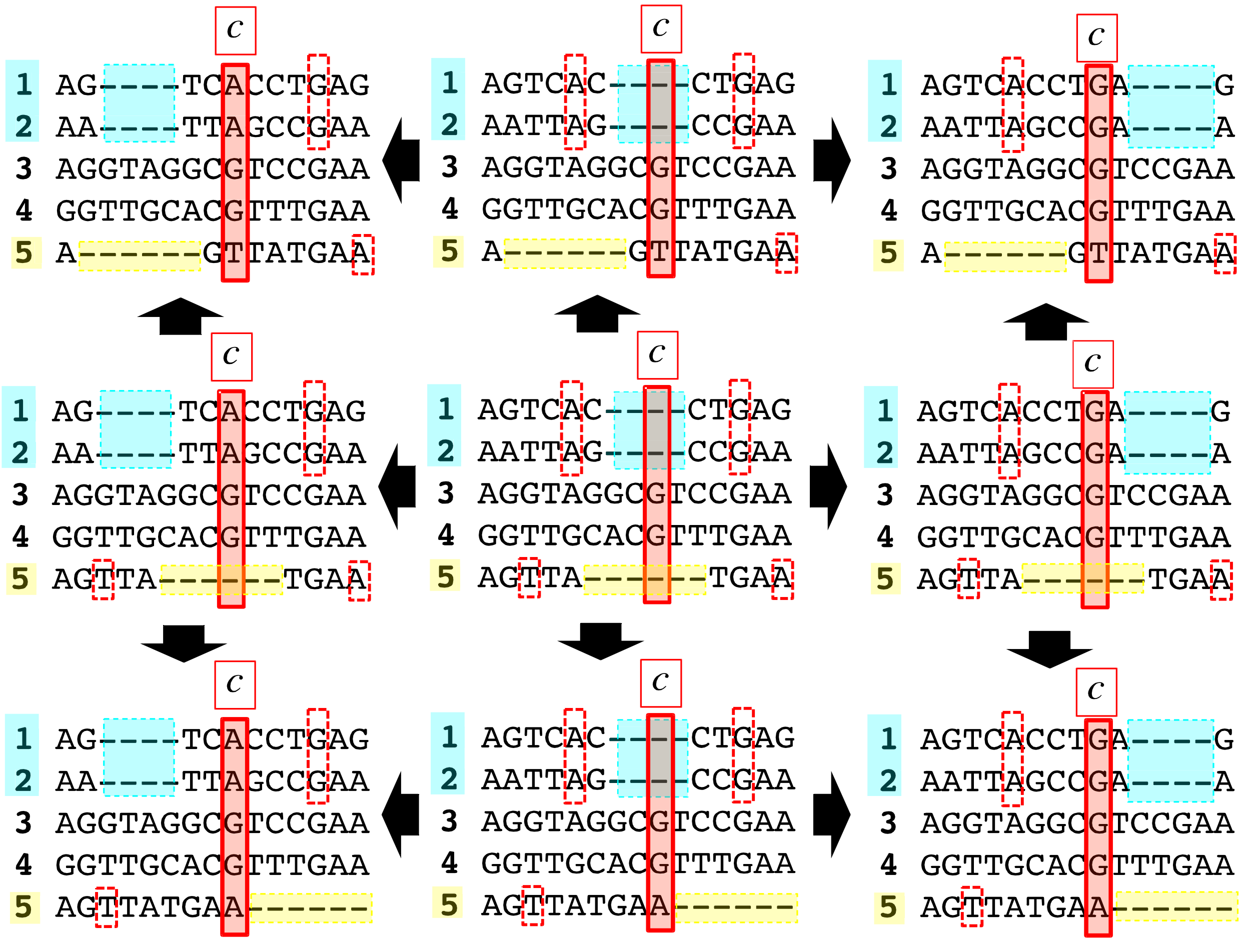


Figure 15. Possible residue configurations of each column caused by double-shifts of two gap-blocks.

The double-shifts of two gap-blocks (cyan- and yellow-shaded) give rise to at most 9 possible configurations in each column (red-shaded),

The red dotted rectangles enclose the original “ingredients” of the columns that resulted from the shifts.