Algorithm for the detection of ICEs/IMEs structures

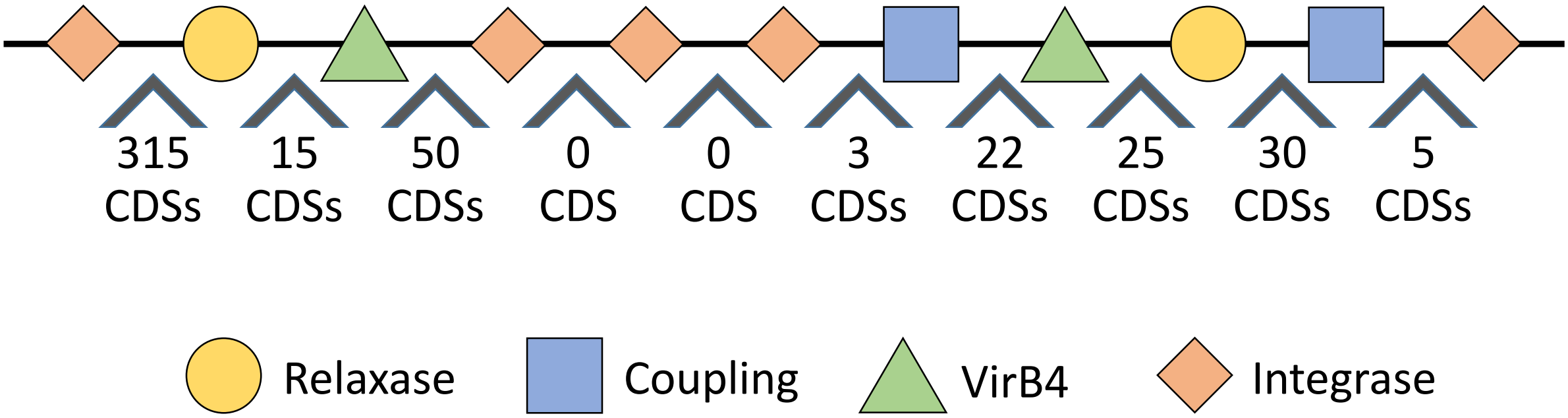
The rules implemented for detecting the ICEs/IMEs structures are based on the biological nature of ICEs/IMEs and empirical evidences deduced from many curated structures by the DynAMic team. The three main steps of the algorithm are :

1. Find anchors of signature proteins (SPs) from the conjugation module and extend them sequentially and bi-directionally
2. Eventually merge distant compatible anchors to find nested structures
3. Find the integrases that belong to the structures

Author : Thomas Lacroix

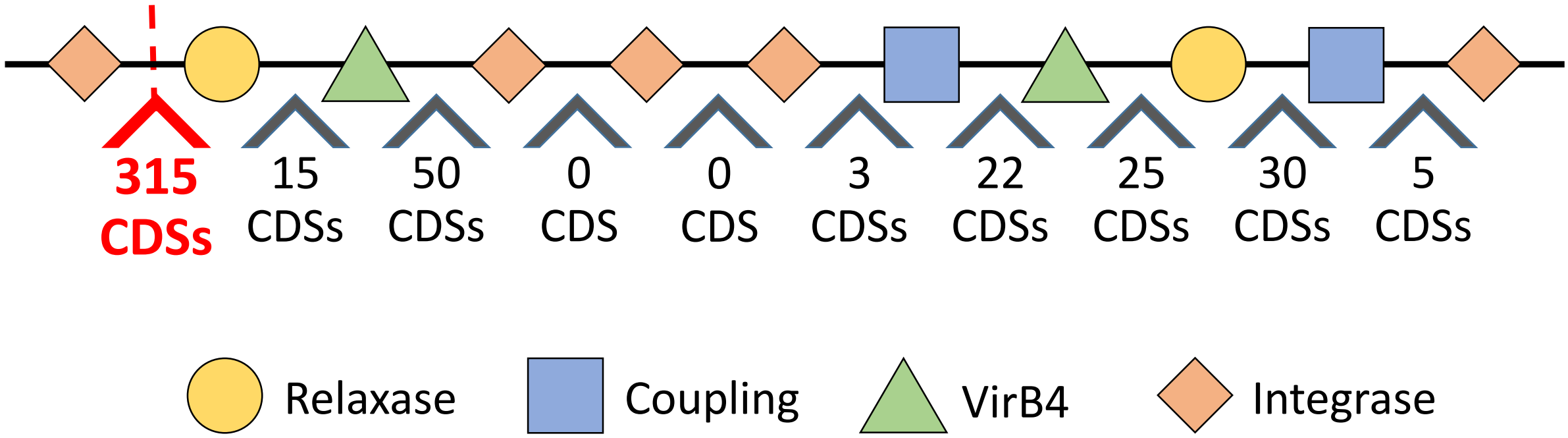
Input data

- Sequence of signature proteins (SPs) ordered on the genome.



Step #1: ICEs / IMEs cannot be too large

- The sequence is cut in segments if >100 CDSs between 2 successive SPs.

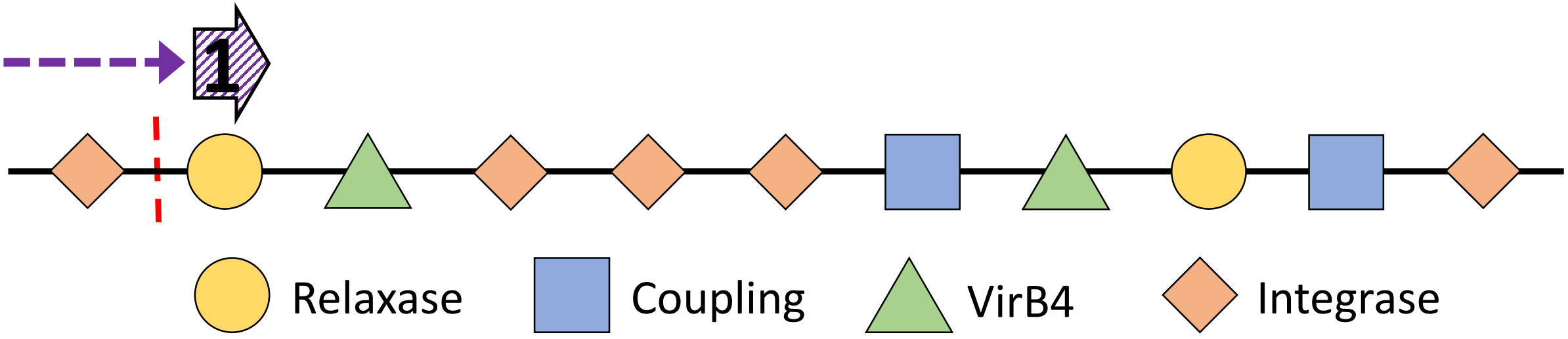


Step #2 : build anchors based on sequential SPs from the conjugation module

- At first, only sequential SPs are considered to build anchors. Building of non-sequential structures (i.e. nested structures) will be dealt with subsequently (merging of anchors).
- SPs from the conjugation module used to build anchors are relaxase, coupling, and virB4. They are quite specific of ICEs / IMEs conjugation modules if they are found in combination with at least one other SP within a short genomic region (<100kb).
- Integrases are not part of the conjugation module and are less specific of ICEs / IMEs structures as they may also relate to other mobile elements (i.e. prophages for Tyr or Ser, transposons or IS for DDE). Integrase are always found at the border of the mobile element and are dealt with at a latter stage.

Step #2.1 : rules for creating an anchor

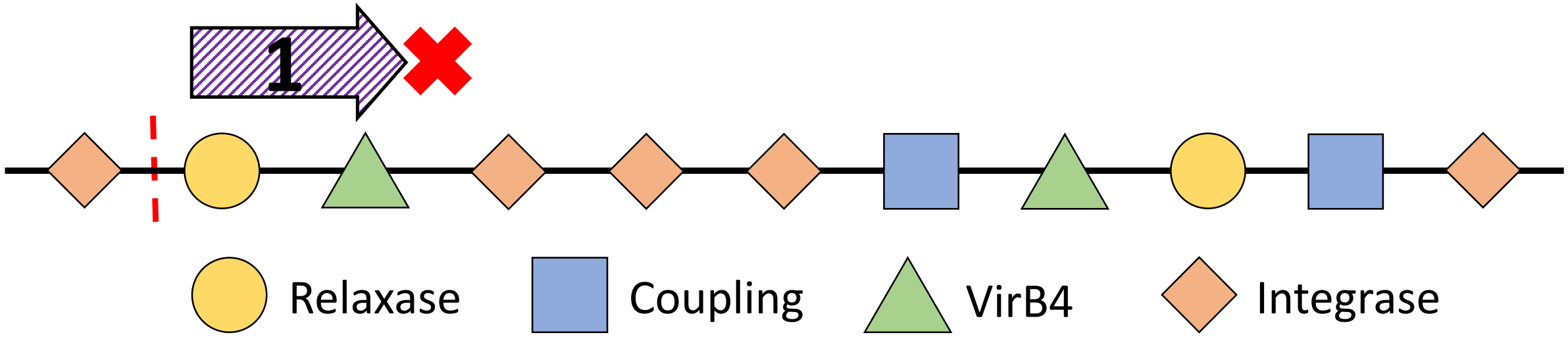
The sequence is scanned from left to right (→). When either one of the conjugation module SPs (relaxase, coupling, or virB4) is found, the anchor starts (1).



Step #2.2: extension of anchor from left to right (1/2)

The sequence continues to be scanned from left to right. An ICE / IME anchor cannot contains (conditions for stopping the extension):

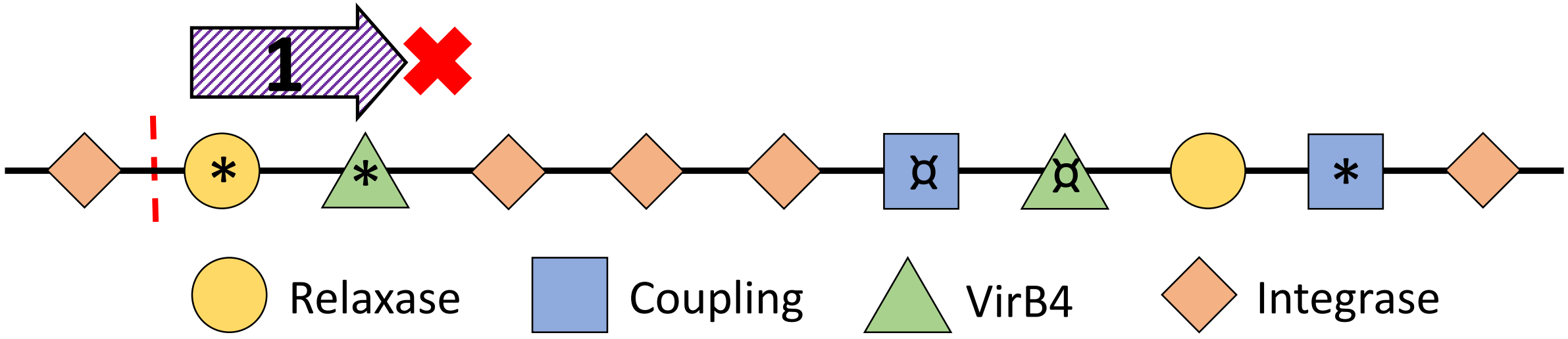
- 2 SPs separated from more than 100 CDSs (step #1).
- 2 virB4 or 2 coupling
- 2 relaxase unless they are adjacent on the genome or separated by one CDS.



Step #2.2: extension of anchor from left to right (2/2)

An ICE / IME anchor cannot contain (conditions for stopping the extension):

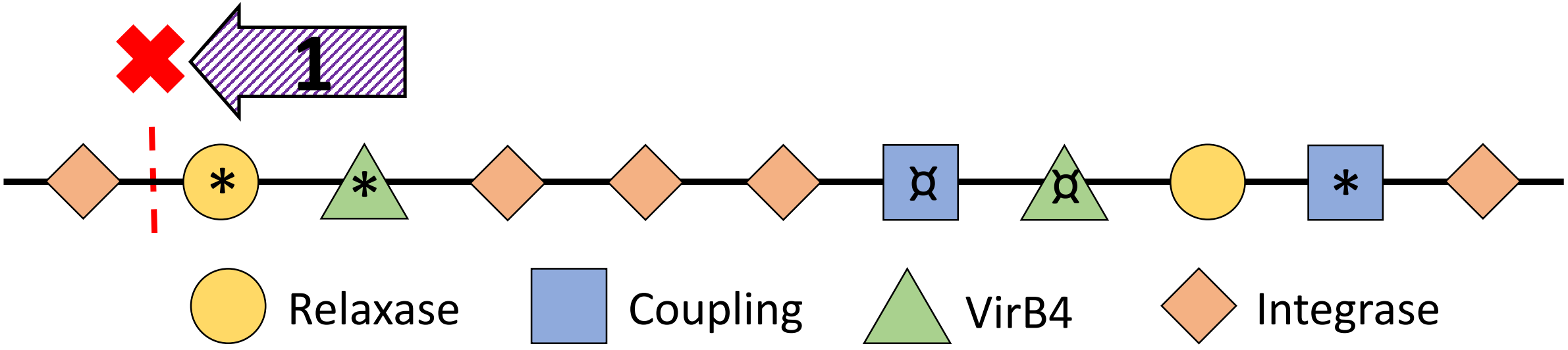
- SPs of different families (i.e. * = ICESt3, ¤ = Tn916). Families of ICEs and IMEs are curated known elements in *Streptococcus*. BlastP hits of the same family are preferably grouped within an anchor while BlastP hits of different families are separated. SPs without any family information (i.e. HMM hits) can be added to an anchor regardless of the family criterion.
- Integrase (they will be dealt with at a latter stage).



Step #2.3: extension of anchor from right to left

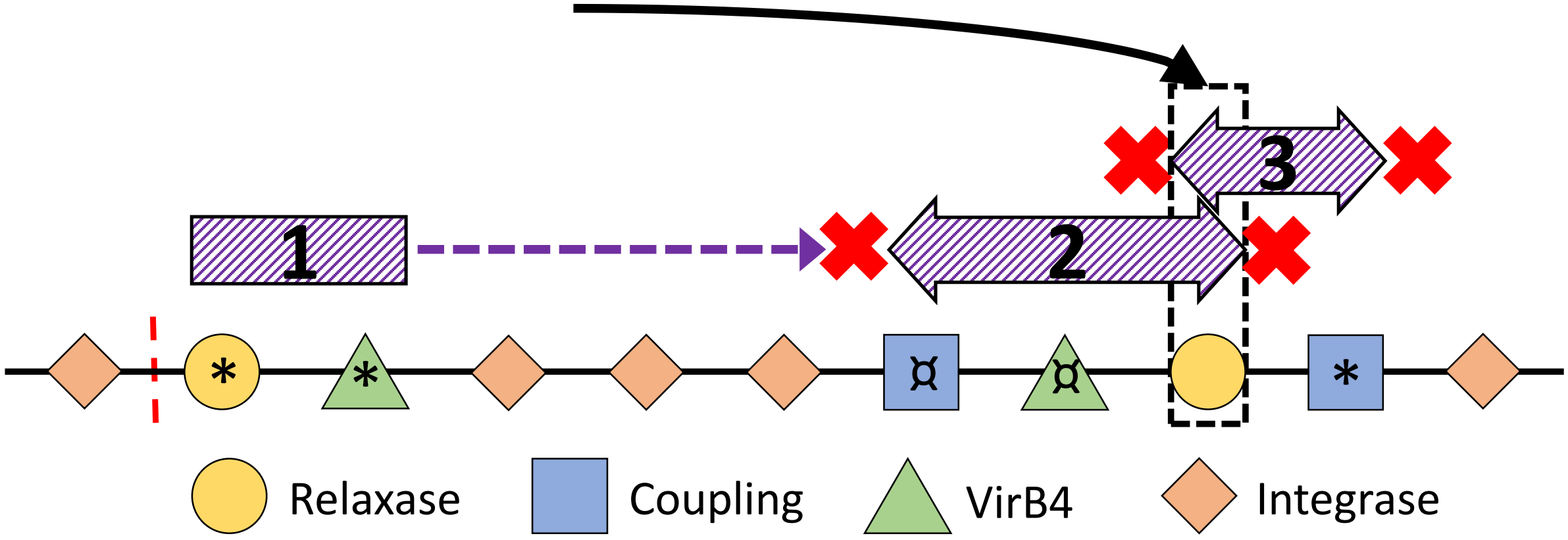
Once an anchor has been created and extended from left to right, it is then extended from right to left (same stopping conditions).

- ICEs / IMEs have no direction so the algorithm must be independent of the choice of the initial scanning direction.



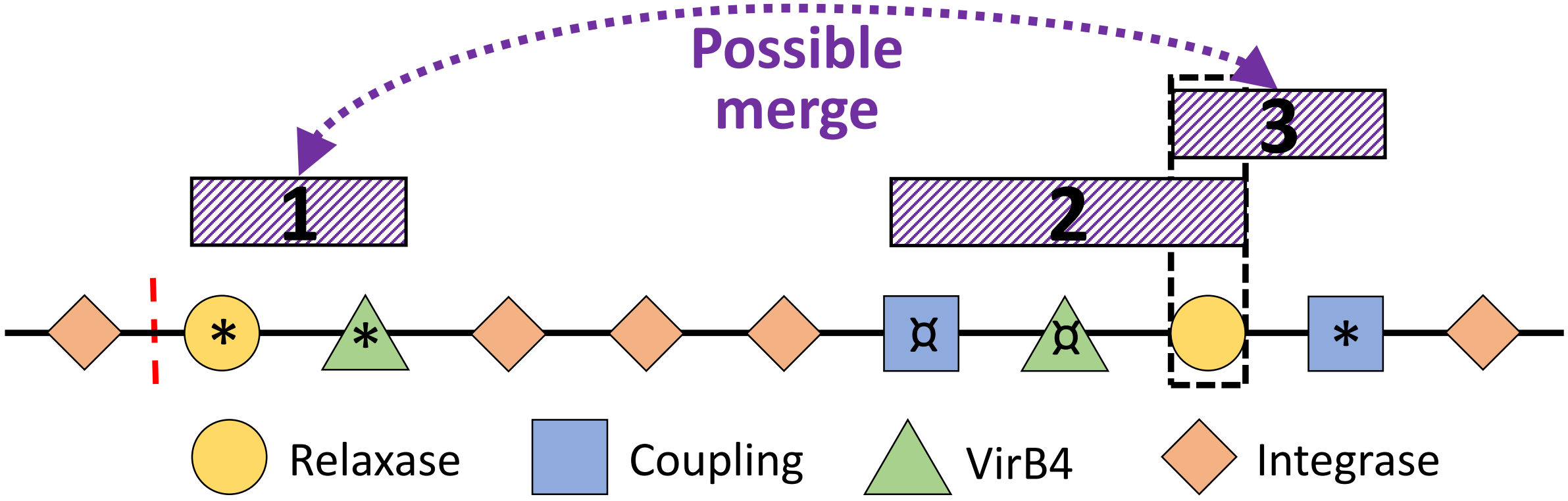
Repeat steps #2.1 to #2.3

- Starting from the sequential SP right of the anchor that was just built, the steps #2.1 to #2.3 are repeated until the whole sequence of SPs is scanned.
- Some SPs can be attributed to 2 different anchors.



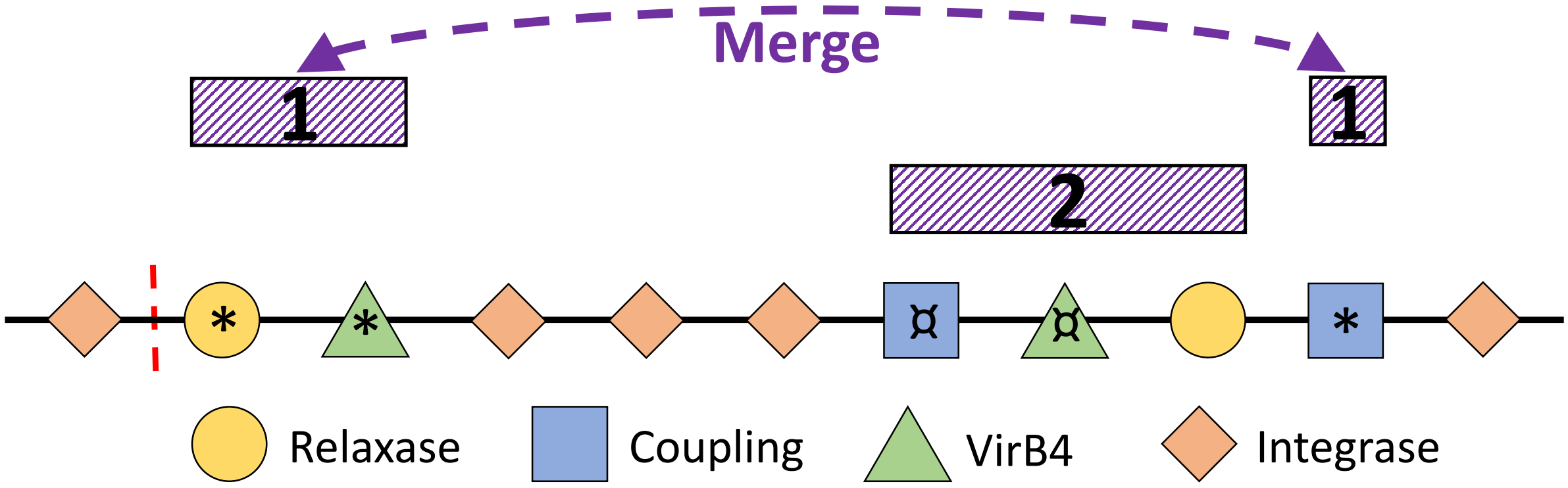
Step #5: merging of anchors (1/2)

- Exhaustive: all combinations of merging are tested. The priority is given to the merging of the nearest anchors if there is multiple possibilities.
- Recursive: detection of cases with multiple levels of nesting and/or when the ICEs / IMEs are "split apart" in more than 2 pieces (rare case).
- The rules for merging are identical to the rules for extending an anchor.



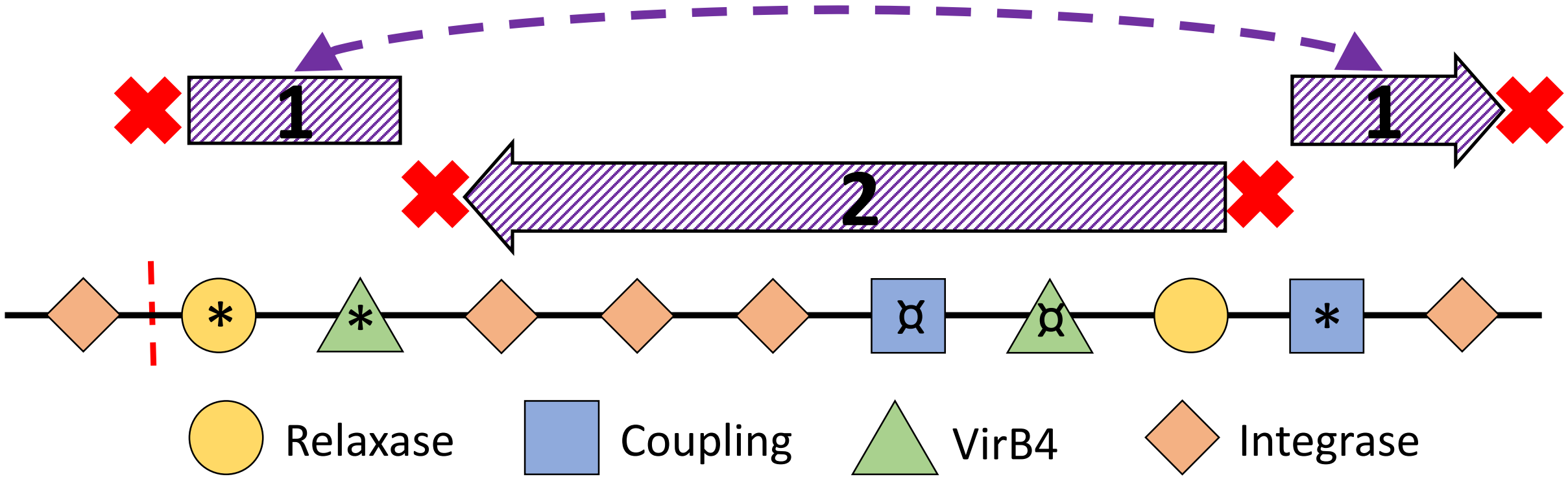
Step #5: merging of anchors (2/2)

- This step can help resolve SPs previously attributed to 2 different anchors.



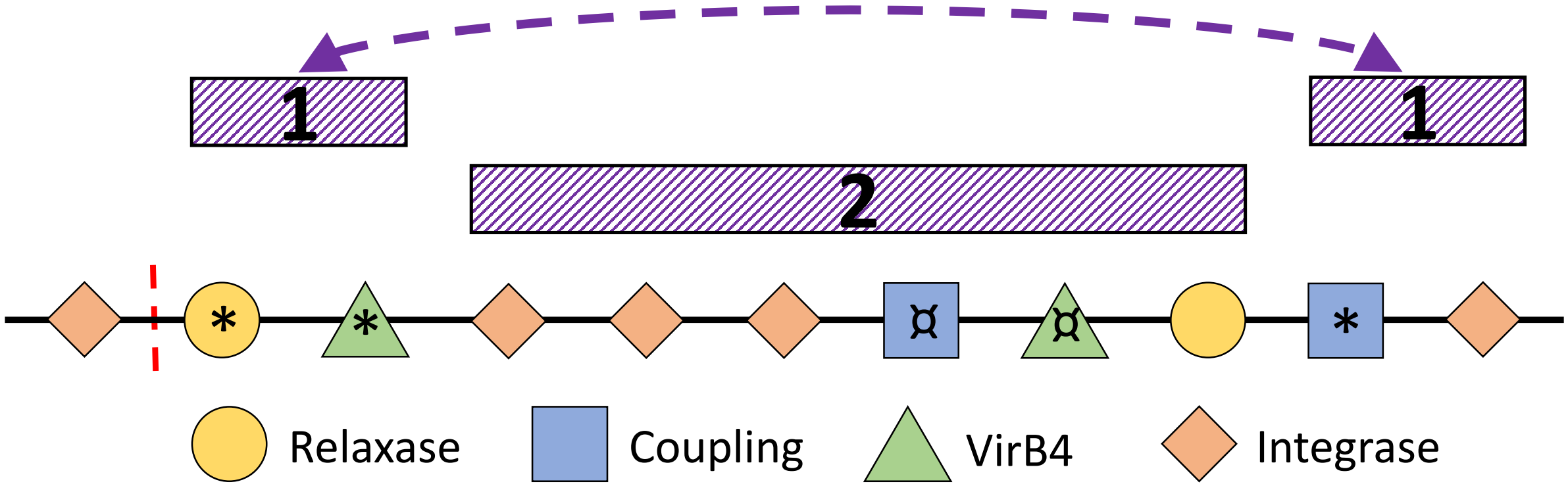
Step #6: adding an integrase to an anchor

- Integrases are always at the border of the mobile element.
- Integrase can be up or downstream within the 100 CDS limit (step 1).
- Integrase sequential to the anchor are good candidate but there can be more distant integrase in case of nested ICEs/IMEs

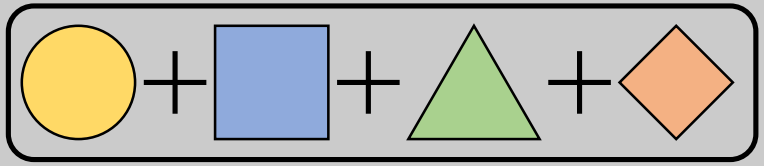
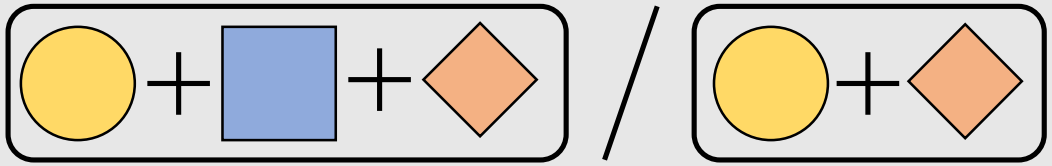
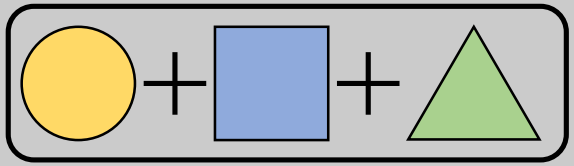
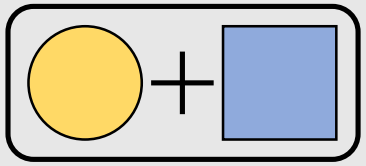


Special cases regarding the integrases

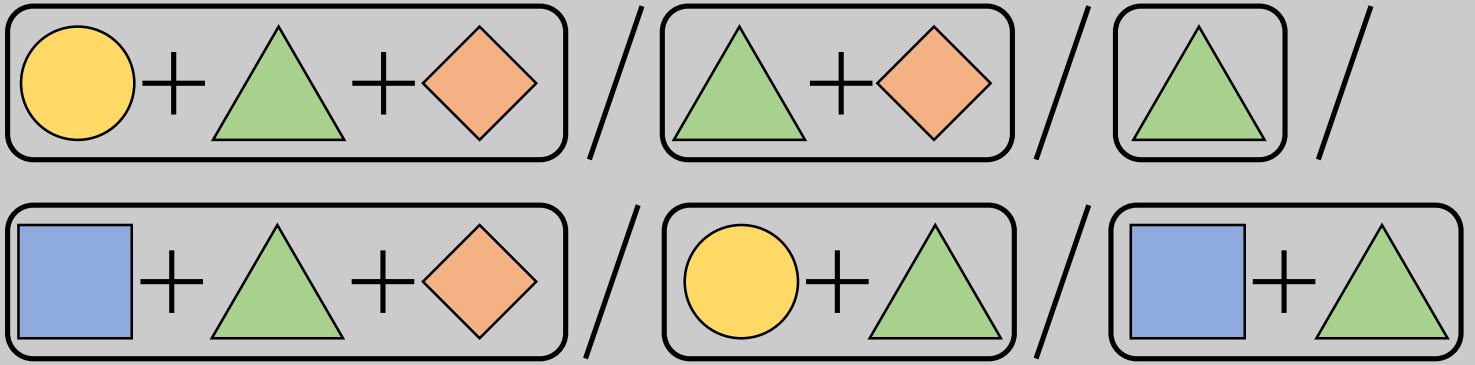
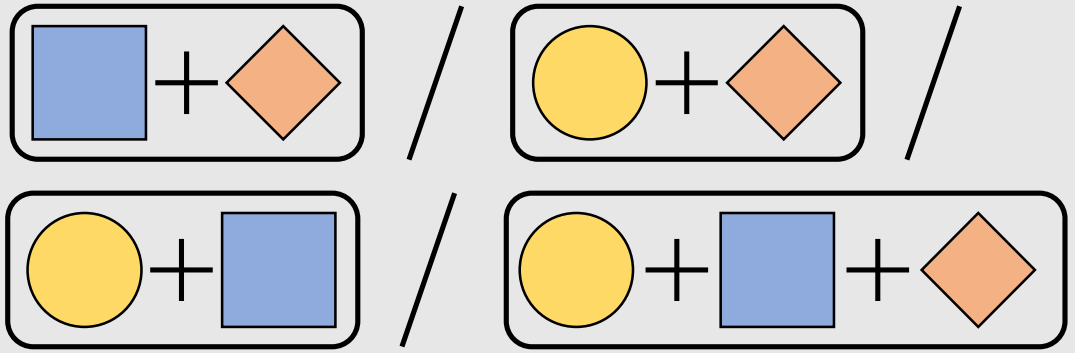
- Adjacent trio or duo of integrases (may be separated by a CDS).
- Upstream ICE → integrase strand - ; downstream ICE → strand +.
- The algorithm may not be able to choose between upstream and downstream good candidate integrases.



Step #7: classification types of ICEs / IMEs (1/2)

Element type	Max size between two sequential SPs	Combinations of SPs
ICE	≤ 100 CDS	
IME	≤ 10 CDS	
Conjugation module	≤ 100 CDS	
Mobilizable element	≤ 10 CDS	

Step #7: classification types of ICEs / IMEs (1/2)

Element type	Max size between two sequential SPs	Combinations of SPs
Partial ICE	≤ 100 CDS	
Other partial element	≥ 10 CDS and ≤ 100 CDS	

Test sets of 124 ICEs / IMEs structures

Manually adapted from real cases to test the algorithm on a variety of complex cases:

- Signatures proteins : 425
- Complete ICEs: 34
- Conjugation modules: 6
- Partial ICEs: 10
- Complete IMEs: 57
- Mobilizable elements (R+C < 10 CDS) : 2
- Other partial elements (R+C>10 CDS, R+V, V+C) : 15
- Nested elements: 24