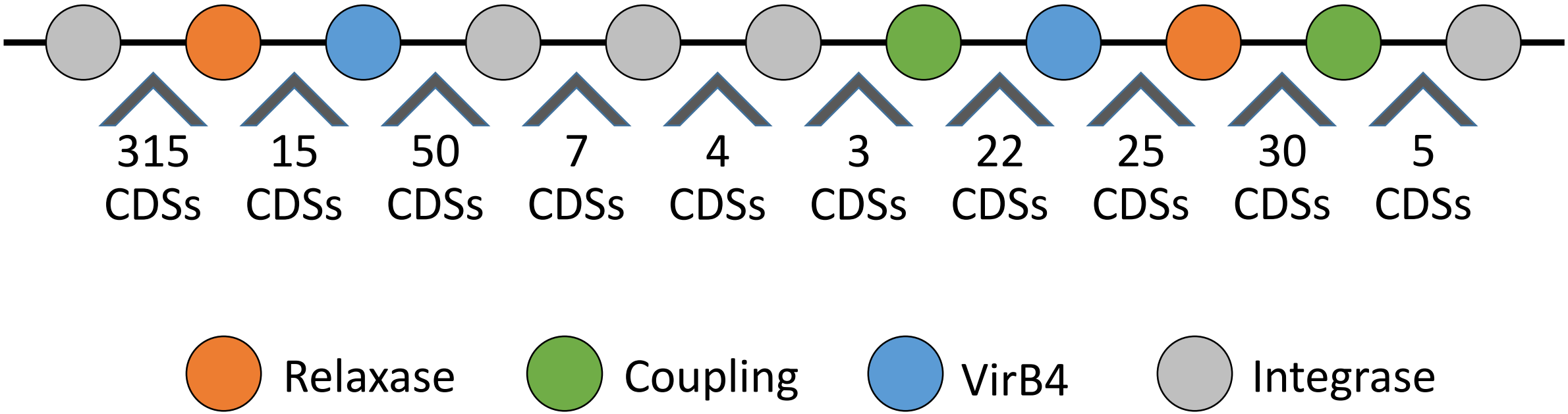
Algorithm for the detection of ICEs/IMEs structures

- Based on seed extension (like blast) and fusion of compatible seeds.
- Object-oriented implementation to facilitate data structuring.

Author : Thomas Lacroix

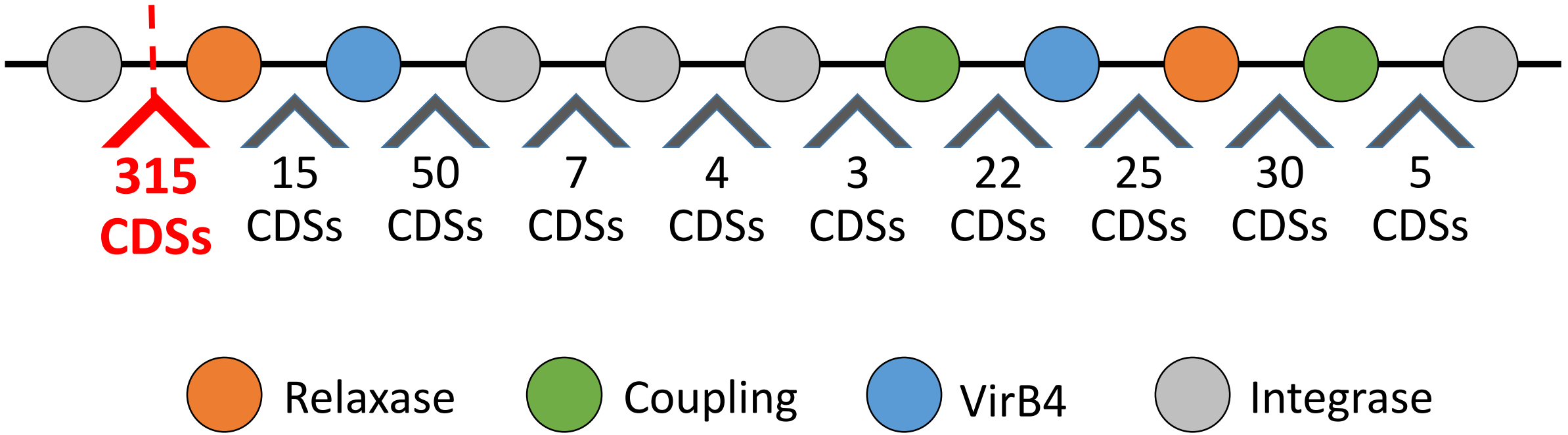
Input data

- Sequence of signature proteins ordered on the genome.



1st step: ICEs / IMEs cannot be too large

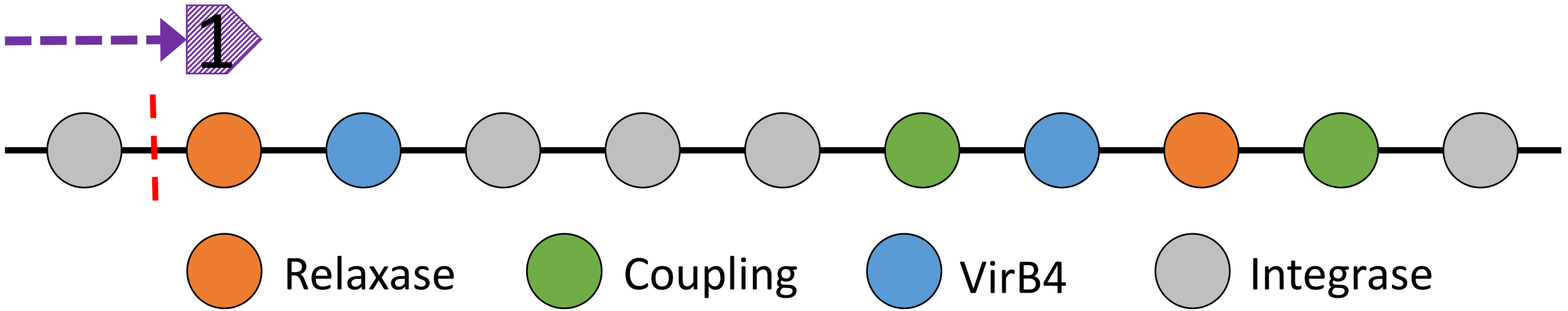
- the sequence is cut if >100 CDSs between 2 signature proteins.



2nd step: rules for creating a seed

The sequence is scanned from left to right (→). When either one of the 3 signature proteins relaxase, coupling, or virB4 is found → seed start (1).

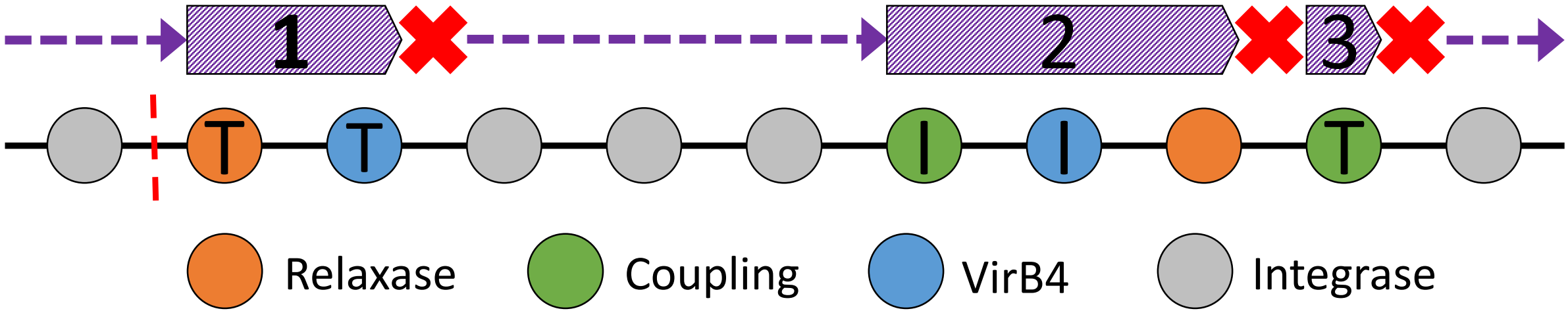
- If an element contains a virB4, it indicates an ICE (complete or partial).
- If an element contains a coupling or relaxase and at least one other signature proteins, it indicates an ICE or IME.
- If an integrase is found → less specific of ICEs / IMEs (may be other EMs, i.e. transposons). The integrase is always at the border of the element.



3rd step: rules for extending a seed

The sequence continue to be scanned from left to right. An ICE / IME seed cannot contain (conditions for stopping the extension) :

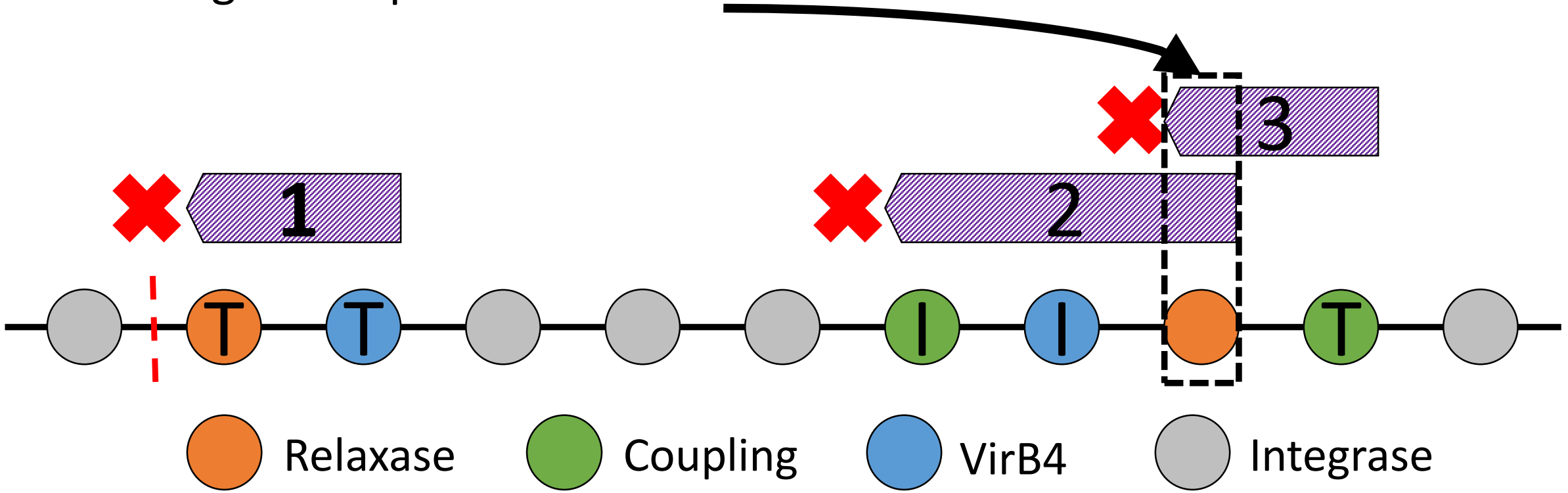
- 2 signature proteins separated from more than 100 CDSs (step 1).
- 2 signature proteins of the same type (unless they are adjacent on the genome and of type relaxase or coupling).
- Signature proteins of different families (i.e. **I** = ICESt3, **T** = Tn916).
- Integrase : integrase are always at the border of the element, they will be dealt with subsequently.



4th step: extending seeds from right to left

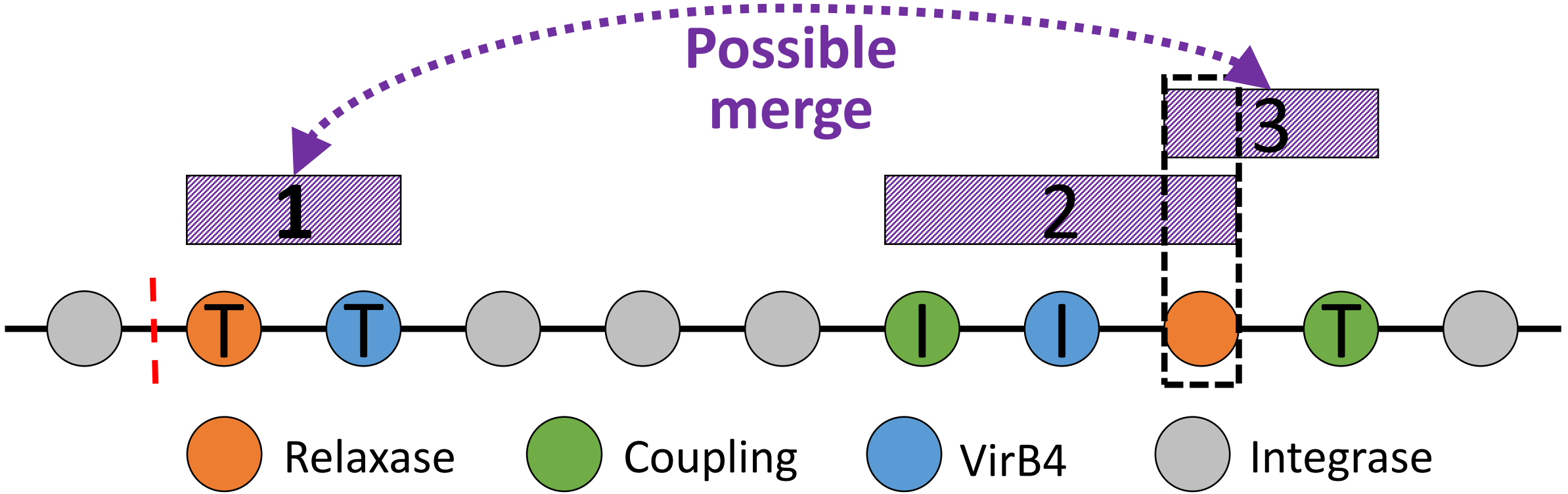
After the 3rd step (creation and extension of seeds from left to right), each seed is extended from right to left (same stopping conditions).

- ICEs / IMEs have no direction.
- Algorithm consistent and independent of the choice of scanning direction.
- Possible signature protein in "conflict" attached to 2 different seeds.



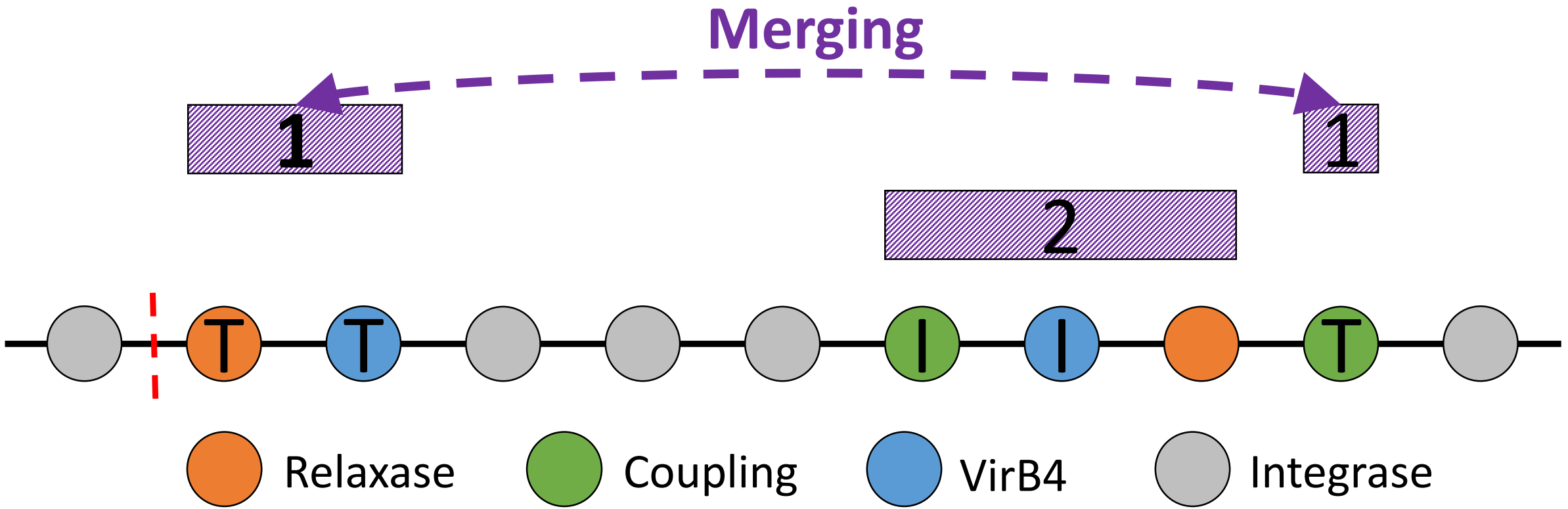
5th step: merging of seeds

- Exhaustive: all combinations of merging are tested. The priority is given to the merging of the nearest seeds 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 a seed.



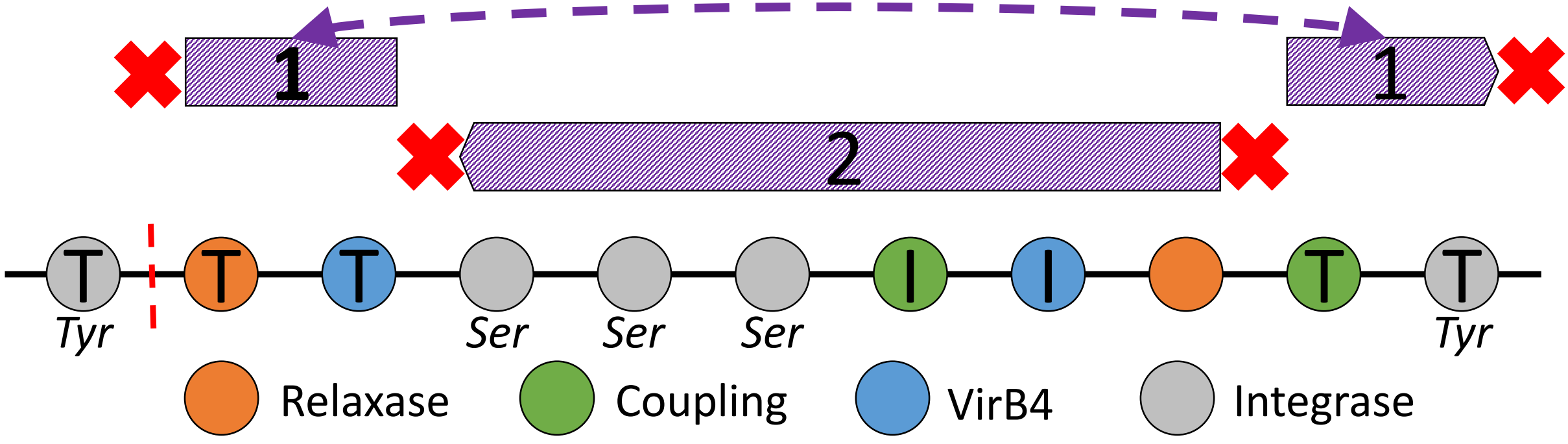
5th step: merging of seeds

- This step can help resolve signature proteins in “conflict” (attached to 2 different seeds).



6th step: rules for adding the integrase(s) to seeds

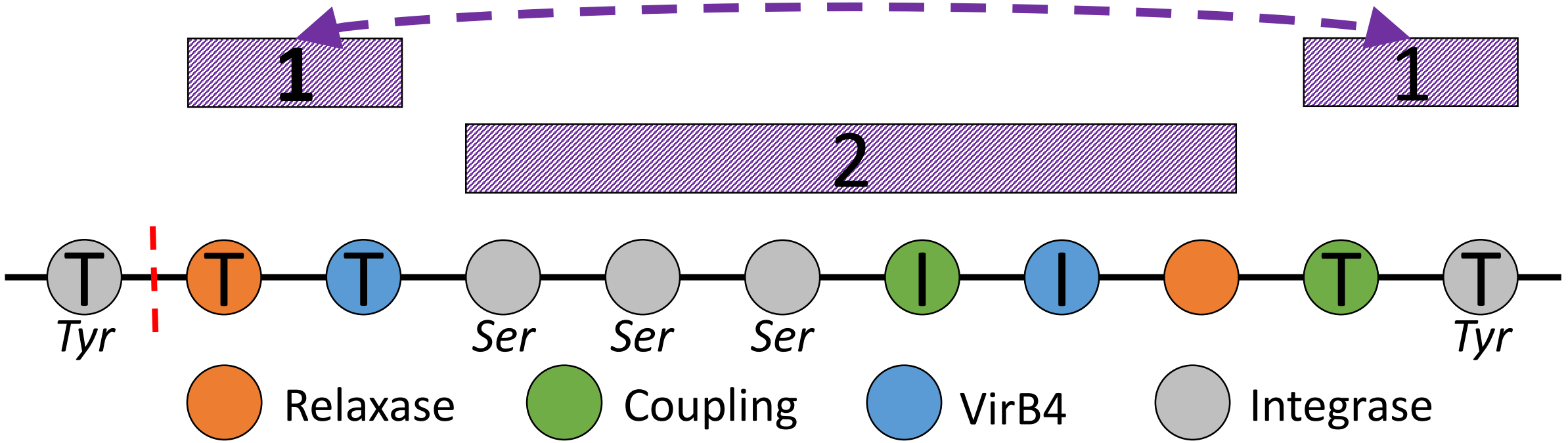
- Integrase(s) can be up or downstream within the 100 CDS limit (step 1).
- Priority is given to (1) integrase()s from the same family (i.e. I = ICESt3, T = Tn916) than the other signature proteins in the seed and (2) integrase(s) adjacent to the seed (if nested ICEs/IMEs, there can be distant integrases).



6th step: rules for adding the integrase(s) to seeds

Special cases :

- Adjacent Ser integrases on the genome.
- Upstream ICE → integrase strand - ; downstream ICE → strand +.
- The algorithm may not be able to choose between an upstream or a downstream integrase.



7th step: classification of different types of ICEs / IMEs

Complete, partial, to be verified experimentally, nested, etc. :

- Complete ICE: R+C+V+I
- Conjugation module: R+C+V
- Partial ICE: V + other signature proteins
- Complete IME: R+I or R+C+I with distance < 10 CDS
- Mobilizable element: R+C with distance < 10 CDS
- Other partial element: R+C>10 CDS, R+V, V+C

Test sets of 89 ICEs / IMEs

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

- Signatures proteins : 356
- Complete ICEs: 23
- Conjugation modules: 8
- Partial ICEs: 11
- Complete IMEs: 37
- Mobilizable elements (R+C < 10 CDS) : 3
- Other partial elements (R+C>10 CDS, R+V, V+C) : 7
- Nested elements: 47