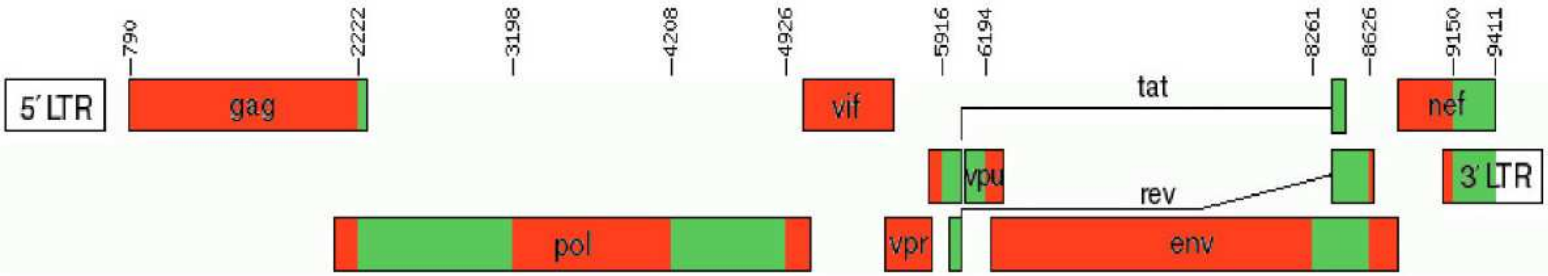
Biological applications: HIV recombination detection

Different subtypes of HIV (A, B, C, . . .)

Sometimes genomes recombine

Labels: one for each subtype



Source: Schultz et al 2006; subtypes A and G

Jumping HMMs for HIV recombination detection

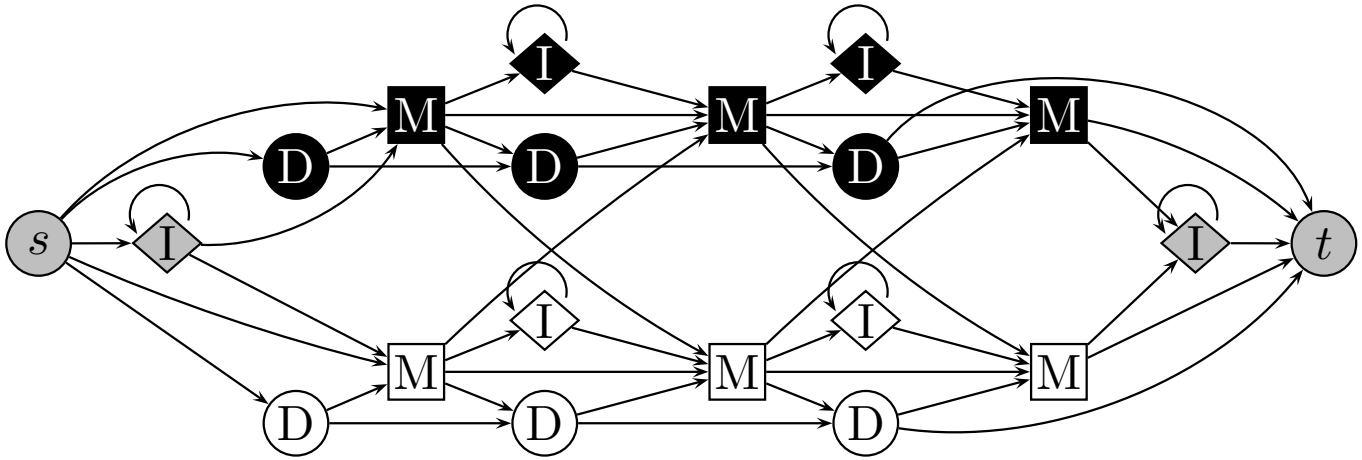
[Schultz et al 2006]

Profile HMM for each subtype

— represents multiple alignment of sequences in the subtype

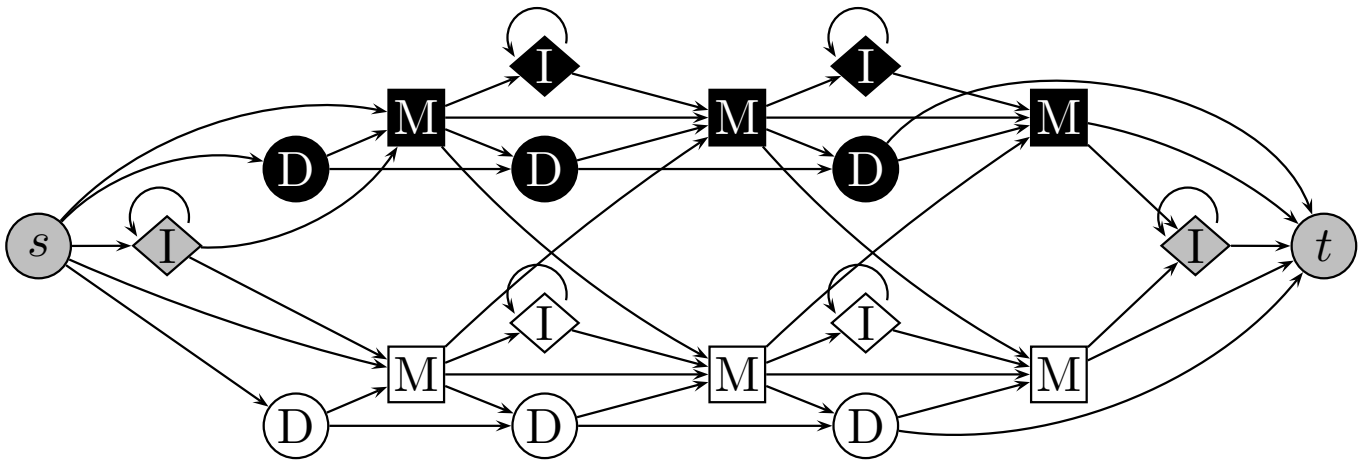
— match, insert, delete state for each alignment column

— profiles connected by jumping transitions




```
TTTTGGCTGAGGCAATGAGCCAAGCAACAAATGC
TTTTGGCCGAGGCAATGAGTCAAGCA---AATTC
TTTTGGCTGAGGCAATGAGCCAAGCA---AATAC
```

Annotation issues in jumping HMMs



State path: alignment of sequence to subtype profiles

Annotation: segments of inputs emitted by subtype profiles

Problems with most probable annotation:

— probably hard to decode

— many annotations with slightly shifted boundaries

Change the objective function for decoding

Gain function [Hamada et al. 2009]

$G(A, A')$ measures accuracy of A wrt. correct annotation A'

Examples:

Identity: score 1 iff A completely correct, 0 otherwise

Pointwise: score +1 for every correct label in A

Boundary: score +1 for every correct boundary, $-\gamma$ for incorrect boundary

	Identity	Pointwise	Boundary
$A =$  $A' =$ 	1	5	4
$A =$  $A' =$ 	0	4	$3 - \gamma$

Optimizing expected gain

Goal: find annotation A that maximizes

$$E_{A'|X}[G(A, A')] = \sum_{A'} G(A, A')P(A'|X)$$

Identity gain function: Viterbi algorithm

Pointwise gain function: Posterior decoding (forward-backward)

Boundary gain function: [Gross et al. 2007]

The choice of gain function is application-dependent