plan

HMM setup

In the simplest case we have a read that we can align to two references (ref_A and ref_B). The read can potentially be a recombinant of the two references.

```
diff  =     x    x x        x        x            x          x
ref_A = ACTTCTGAGTATCGTCTTCAGTCGCTAGCTATATTATCGATCAATAATCTATTTTTGC
ref_B = ACTTTTGAGGAACGTCTTCAATCGCTAGCAATATTATCGATCGATAATCTATTATTGC

qry   = ACTTCTGAGGATCGTTTTCAGTCGCTAGCTATATTGTCGATCGATAATCTATTTTTGC
state = ....a....b.a...^....a........a.....^......b..........b....
```

The two references are similar and only differ in few sites (`x`). The value of the query on these sites gives us information on whether this region is likely to be vertically inherited from one reference or the other.

The vertical origin of each site is the hidden state in our hidden markov model (`A` or `B`). The probability of jumping from one state to the next depends on a transition matrix:

	A	B
A	$1-x$	x
B	x	$1-x$

where x is the transition probability.

I have three (or four) possible visible states:

- `a`: agrees with ref_A
- `b`: agrees with ref_B
- `.`: agrees both with ref_A and ref_B (they are the same)
- `^`: agrees with neither of the references.

Depending on the value of the hidden state, visible states have different probabilities:

$P(V\|H)$	A	B
`a`	γ	ϵ
`b`	ϵ	γ
`.` or `^`	$1-\gamma-\epsilon$	$1-\gamma-\epsilon$

Where:

- γ is the probability of emitting evidence for the correct state.
- ϵ is the (small) probability of emitting evidence for the wrong state

For the purpose of hidden state inference, the visible states . and ^ can be considered as the same.

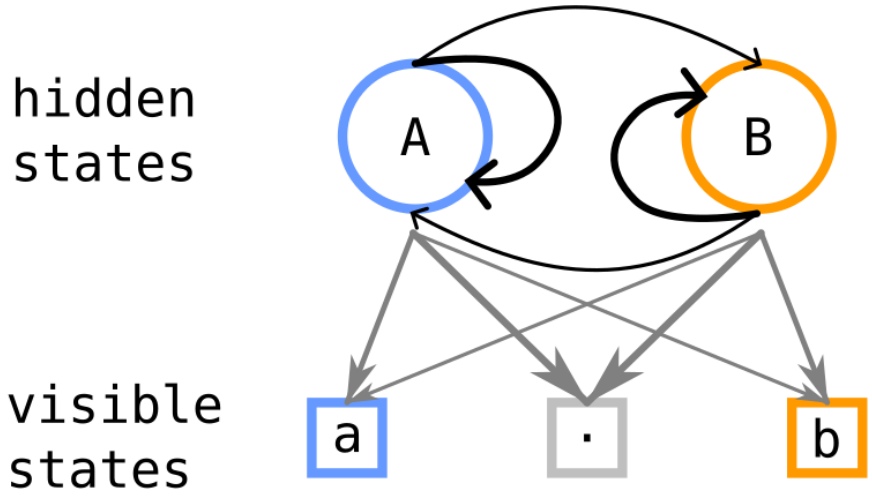

Figure 1: hmm

parameter inference

- the value of x depends on our expectation of the likelihood of observing a recombination event as a function of the read length. This should be reasonably small, but we can test how the result changes for few values of x.

model improvements

- how to deal with in/dels? They offer evidence. We can add another visible state? And a corresponding hidden state?
- add sequence position in HMM? evidence positions are not random, they can only be where the references differ.

Tasks

- ☐ Viterbi algorithm
  - ☐ Find a package that implements the Viterbi algorithm
  - ☐ Simulate the HMM with sensible values of the parameters
  - ☐ Apply the Viterbi algorithm on simulated data and verify that things work as we expect them
- ☐ Infer the visible states

- ☐ decide how to derive vectors of visible states from the reads. This will involve selecting reads that align to both references, and then extract visible values from the alignment.
- ☐ apply this to reads from a single clone and not population culture. This will allow us to estimate γ and ϵ for the two hidden states, and verify that they are similar (not ϵ_A and ϵ_B).

☐ Once the parameters are inferred, run the inference on the population reads with few values of x, and analyze cases of reference change.