Nanopore automata

Ian Holmes[1,2,*]

1 Lawrence Berkeley National Laboratory, Berkeley, CA, USA

2 Department of Bioengineering, University of California, Berkeley, CA, USA

Contents

1 Abstract

State machine algorithms for aligning Nanopore reads.

2 Specification

Initial goal (Preliminary Results) is simple reusable code for aligning a segmented nanopore read (with segment currents summarized) to a reference sequence.

Longer-term goals (Specific Aims) include

- quasi-hierarchical series of models for processed→raw data (raw, FAST5, FASTQ, FASTA)
- transducer intersection-style models for read-pair alignment
- systematic strategies for approximation/optimization algorithms, climbing the hierarchy (starting with k-mer or FM-index approaches)
- transducer intersection models for aligning reads from different sequencing technologies

2.1 Parameterization algorithm

Given the following inputs

- Reference genome (FASTA)
- Segment-called reads (FAST5/HDF5)

Perform the following steps

- Perform Baum-Welch to fit a rich model

Rich model incorporates segment statistics.

2.2 Reference search algorithm

Given the following inputs

- Reference genome
- Segment-called reads (FAST5/HDF5)
- Parameterized rich model

Perform the following steps

- Perform Viterbi alignment

2.3 Implementation

Libraries etc.

HDF5...

2.4 Evaluation

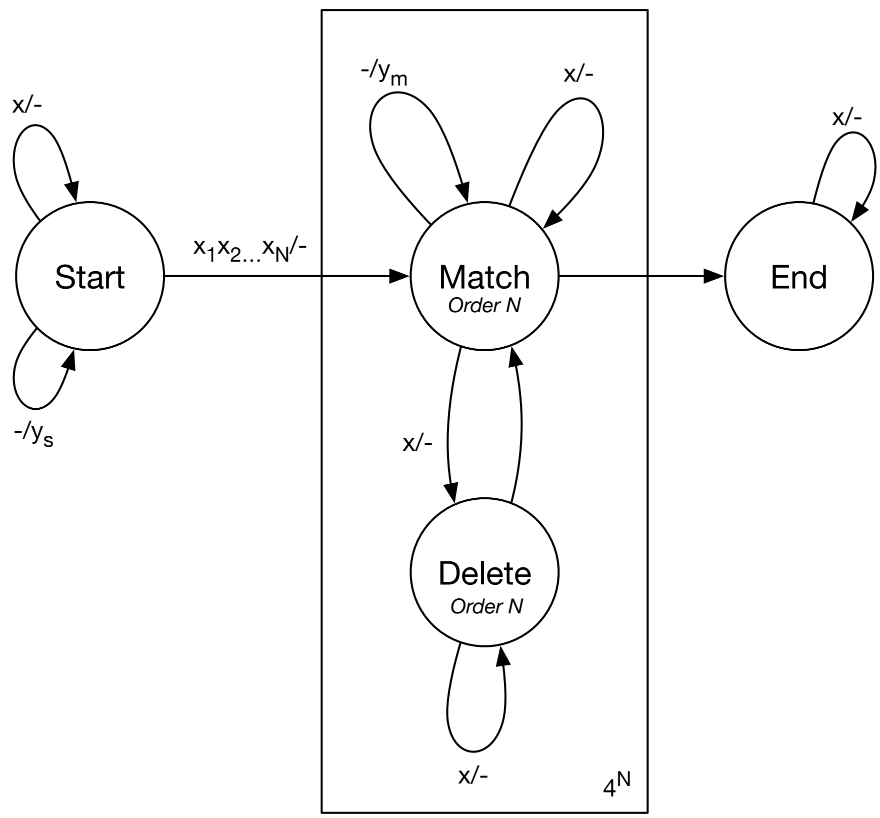
Strategy...

Data sets...

3 Methods

Model & inference algorithms.

3.1 Model



- Order-N Mealy transducer.
- Input alphabet: $\Omega = \{A, C, G, T\}$ (nucleotides)
- Output alphabet: $\Re$ (real numbers signifying current levels)
- States: Start, $\text{Match}_{x_1 \ldots x_N}$, $\text{Delete}_{x_1 \ldots x_N}$, End
- Parameters: $p^{\text{StartEmit}}, p^{\text{BeginDelete}}, p^{\text{ExtendDelete}}, \mu^{\text{Start}}, \tau^{\text{Start}}$, $\{p^{\text{MatchEmit}}_{x_1 \ldots x_N}, \mu^{\text{Match}}_{x_1 \ldots x_N}, \tau^{\text{Match}}_{x_1 \ldots x_N} : x_1 \ldots x_N \in \Omega^N\}$

Transitions:

Source	Destination	Weight	Absorbs	Emits
Start	Start	$p^{\text{StartEmit}}$ $\times P(y_s \mid \mu^{\text{Start}}, \tau^{\text{Start}})$		$y_s \sim \text{Normal}(\mu^{\text{Start}}, \tau^{\text{Start}})$
Start	Start	1	$x \in \Omega$	
Start	$\text{Match}_{x_1 \ldots x_N}$	$1 - p^{\text{StartEmit}}$	$x_1 \ldots x_N \in \Omega^N$	
$\text{Match}_{x_1 \ldots x_N}$	$\text{Match}_{x_1 \ldots x_N}$	$p^{\text{MatchEmit}}_{x_1 \ldots x_N}$ $\times P(y_m \mid \mu^{\text{Match}}_{x_1 \ldots x_N}, \tau^{\text{Match}}_{x_1 \ldots x_N})$		$y_m \sim \text{Normal}(\mu^{\text{Match}}_{x_1 \ldots x_N}, \tau^{\text{Match}}_{x_1 \ldots x_N})$
$\text{Match}_{x_1 \ldots x_N}$	$\text{Match}_{x_2 \ldots x_{N+1}}$	$(1 - p^{\text{MatchEmit}}_{x_1 \ldots x_N})$ $\times(1 - p^{\text{BeginDelete}})$	$x_{N+1} \in \Omega$	
$\text{Match}_{x_1 \ldots x_N}$	$\text{Delete}_{x_2 \ldots x_{N+1}}$	$(1 - p^{\text{MatchEmit}}_{x_1 \ldots x_N})$ $\times p^{\text{BeginDelete}}$	$x_{N+1} \in \Omega$	
$\text{Match}_{x_1 \ldots x_N}$	End	$1 - p^{\text{MatchEmit}}_{x_1 \ldots x_N}$		
$\text{Delete}_{x_1 \ldots x_N}$	$\text{Delete}_{x_2 \ldots x_{N+1}}$	$p^{\text{ExtendDelete}}$	$x_{N+1} \in \Omega$	
$\text{Delete}_{x_1 \ldots x_N}$	$\text{Match}_{x_1 \ldots x_N}$	$1 - p^{\text{ExtendDelete}}$		
End	End	1	$x \in \Omega$	

3.2 Baum-Welch algorithm

As usual.

3.3 Viterbi algorithm

As usual.

4 Results

5 Discussion

6 Acknowledgments

7 Figure Legends

8 Appendix

8.1 Exponential distribution

$$\begin{aligned} x &\sim \text{Exponential}(\kappa) \\ P(x|\kappa) &= \kappa \exp(-\kappa x) \\ \mathrm{E}[x] &= \kappa^{-1} \\ \mathrm{Var}[x] &= \kappa^{-2} \end{aligned}$$

Rate parameter κ.

8.2 Gamma distribution

$$\begin{aligned} x &\sim \text{Gamma}(\alpha, \beta) \\ P(x|\alpha, \beta) &= \frac{x^{\alpha-1}\beta^{\alpha}\exp(-x\beta)}{\Gamma(\alpha)} \\ \mathrm{E}[x] &= \alpha/\beta \\ \mathrm{Var}[x] &= \alpha/\beta^2 \end{aligned}$$

Shape parameter α, rate parameter β. $\Gamma()$ is the gamma function

$$\Gamma(\alpha) = \int_0^\infty z^{\alpha-1}\exp(-z)dz$$

Note $\Gamma(n) = (n-1)!$ for positive integer n.

8.3 Normal distribution

$$x \sim \text{Normal}(\mu, \tau)$$

Mean μ, precision τ (precision is reciprocal of variance).

$$P(x|\mu,\tau) = \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{\tau}{2}(x-\mu)^2\right)$$