

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	2
2	Specification	2
2.1	Parameterization algorithm	2
2.2	Reference search algorithm	3
2.3	Implementation	3
2.4	Evaluation	3
3	Methods	3
3.1	Null model	4
3.2	Homology model	5
3.3	Baum-Welch algorithm	6
3.4	Viterbi algorithm	6
4	Results	7
5	Discussion	7
6	Acknowledgments	8
7	Figure Legends	9
8	Appendix	10
8.1	Exponential distribution	10

8.2	Gamma distribution	10
8.3	Normal distribution	10

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, suitable for long-read assemblers
- 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, for improved assembly
- transducer-based versions of Rahman & Pachter’s CGAL

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

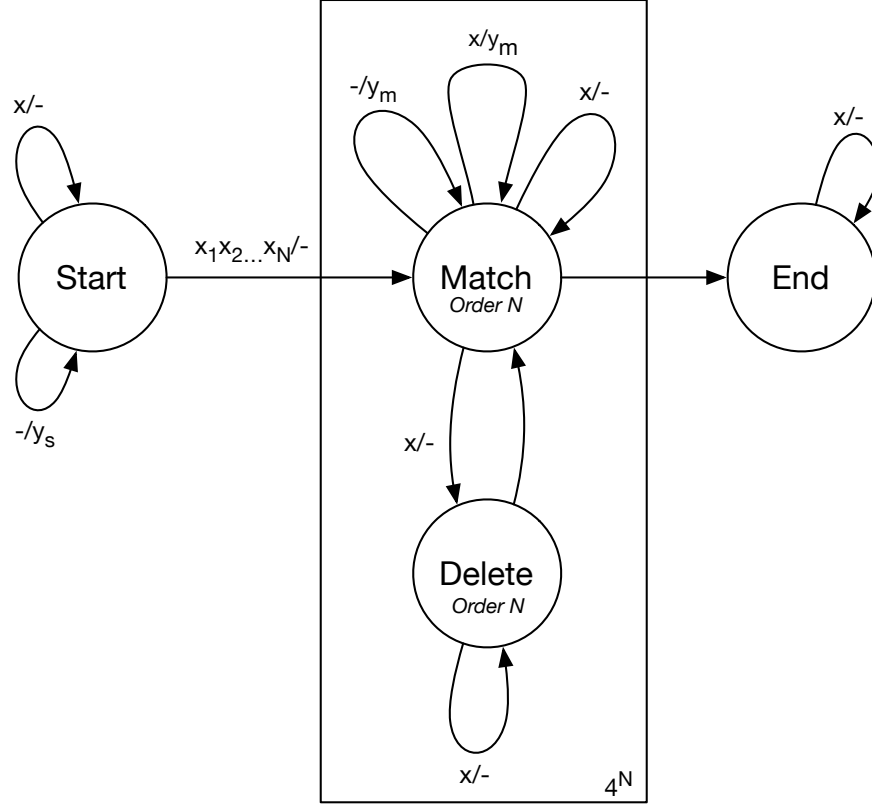
Model & inference algorithms.

3.1 Null model

- Output alphabet: \Re (real numbers signifying current readings)
- K current samples (*ticks*). Sequence is $y_1 \dots y_K$
 - partitioned into a sequence of L *events*: $E_1 \dots E_L$
- Parameters: $p^{\text{NullEvent}}, p^{\text{NullTick}}, \mu^{\text{Null}}, \tau^{\text{Null}}$
- Gaussian emissions: $y_n \sim \text{Normal}(\mu^{\text{Null}}, \tau^{\text{Null}})$
- Probability is

$$\begin{aligned}
 & P(E_1 \dots E_L, y_1 \dots y_K) dy_1 \dots dy_K \\
 &= (1 - p^{\text{NullEvent}}) \prod_{\text{events}: E_l} p^{\text{NullEvent}} (1 - p^{\text{NullTick}}) \prod_{\text{ticks}: y_k \in E_l} p^{\text{NullTick}} P(y_k | \mu^{\text{Null}}, \tau^{\text{Null}}) dy_k
 \end{aligned}$$

3.2 Homology model



- Order- N Mealy transducer.
- Input alphabet: $\Omega = \{A, C, G, T\}$ (nucleotides)
- Output alphabet: real numbers partitioned into events, as with null model
- States: Start, End, $\{ \text{Match}_{x_1 \dots x_N}, \text{Delete}_{x_1 \dots x_N} : x_1 \dots x_N \in \Omega^N \}$
- Parameters: $p^{\text{StartEvent}}, p^{\text{Skip}}, p^{\text{BeginDelete}}, p^{\text{ExtendDelete}},$
 $\{ p_{x_1 \dots x_N}^{\text{MatchEvent}}, p_{x_1 \dots x_N}^{\text{MatchTick}}, \mu_{x_1 \dots x_N}^{\text{Match}}, \tau_{x_1 \dots x_N}^{\text{Match}} : x_1 \dots x_N \in \Omega^N \}$

Transducer can *skip* an individual base (on average one less event emission for that base), or can *delete* a run of bases (zero probability of event emissions

during the run).

Transitions:

Source	Destination	Weight	Absorbs	Emits
Start	Start	$p^{\text{StartEvent}}$ $\times P(y_s \mu^{\text{Start}}, \tau^{\text{Start}}) dy_s$		$\{y_s^{(k)} : 1 \leq k \leq K_s\},$ $K_s \sim \text{Geometric}(p^{\text{NullTick}}),$ $y_s^{(k)} \sim \text{Normal}(\mu^{\text{Null}}, \tau^{\text{Null}})$
Start	Start	1	$x \in \Omega$	
Start	Match $_{x_1 \dots x_N}$	$1 - p^{\text{StartEvent}}$	$x_1 \dots x_N \in \Omega^N$	
Match $_{x_1 \dots x_N}$	Match $_{x_1 \dots x_N}$	$p^{\text{MatchEvent}}_{x_1 \dots x_N}$ $\times P(y_m \mu^{\text{Match}}_{x_1 \dots x_N}, \tau^{\text{Match}}_{x_1 \dots x_N}) dy_m$		$\{y_m^{(k)} : 1 \leq k \leq K_m\},$ $K_m \sim \text{Geometric}(p^{\text{MatchTick}}_{x_1 \dots x_N}),$ $y_m^{(k)} \sim \text{Normal}(\mu^{\text{Match}}_{x_1 \dots x_N}, \tau^{\text{Match}}_{x_1 \dots x_N})$
Match $_{x_1 \dots x_N}$	Match $_{x_2 \dots x_{N+1}}$	$(1 - p^{\text{MatchEvent}}_{x_1 \dots x_N})$ $\times (1 - p^{\text{BeginDelete}}) p^{\text{Skip}}$	$x_{N+1} \in \Omega$	
Match $_{x_1 \dots x_N}$	Match $_{x_2 \dots x_{N+1}}$	$(1 - p^{\text{MatchEvent}}_{x_1 \dots x_N})$ $\times (1 - p^{\text{BeginDelete}})(1 - p^{\text{Skip}})$ $\times P(y_m \mu^{\text{Match}}_{x_1 \dots x_N}, \tau^{\text{Match}}_{x_1 \dots x_N}) dy_m$	$x_{N+1} \in \Omega$	$\{y_m^{(k)} : 1 \leq k \leq K_m\},$ $K_m \sim \text{Geometric}(p^{\text{MatchTick}}_{x_1 \dots x_N}),$ $y_m^{(k)} \sim \text{Normal}(\mu^{\text{Match}}_{x_1 \dots x_N}, \tau^{\text{Match}}_{x_1 \dots x_N})$
Match $_{x_1 \dots x_N}$	Delete $_{x_2 \dots x_{N+1}}$	$(1 - p^{\text{MatchEvent}}_{x_1 \dots x_N})$ $\times p^{\text{BeginDelete}}$	$x_{N+1} \in \Omega$	
Match $_{x_1 \dots x_N}$	End	$1 - p^{\text{MatchEvent}}_{x_1 \dots x_N}$		
Delete $_{x_1 \dots x_N}$	Delete $_{x_2 \dots x_{N+1}}$	$p^{\text{ExtendDelete}}$	$x_{N+1} \in \Omega$	
Delete $_{x_1 \dots x_N}$	Match $_{x_1 \dots x_N}$	$1 - p^{\text{ExtendDelete}}$		
End	End	1	$x \in \Omega$	

3.3 Baum-Welch algorithm

As usual.

3.4 Viterbi algorithm

As usual.

4 Results

5 Discussion

6 Acknowledgments

7 Figure Legends

8 Appendix

8.1 Exponential distribution

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

Rate parameter κ .

8.2 Gamma distribution

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

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).

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