

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	3
3.2	Homology model	5
3.3	Basecalling model	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

8.4	Transition table for nanopore transducer	12
8.5	Transition table for basecalling HMM	14

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.

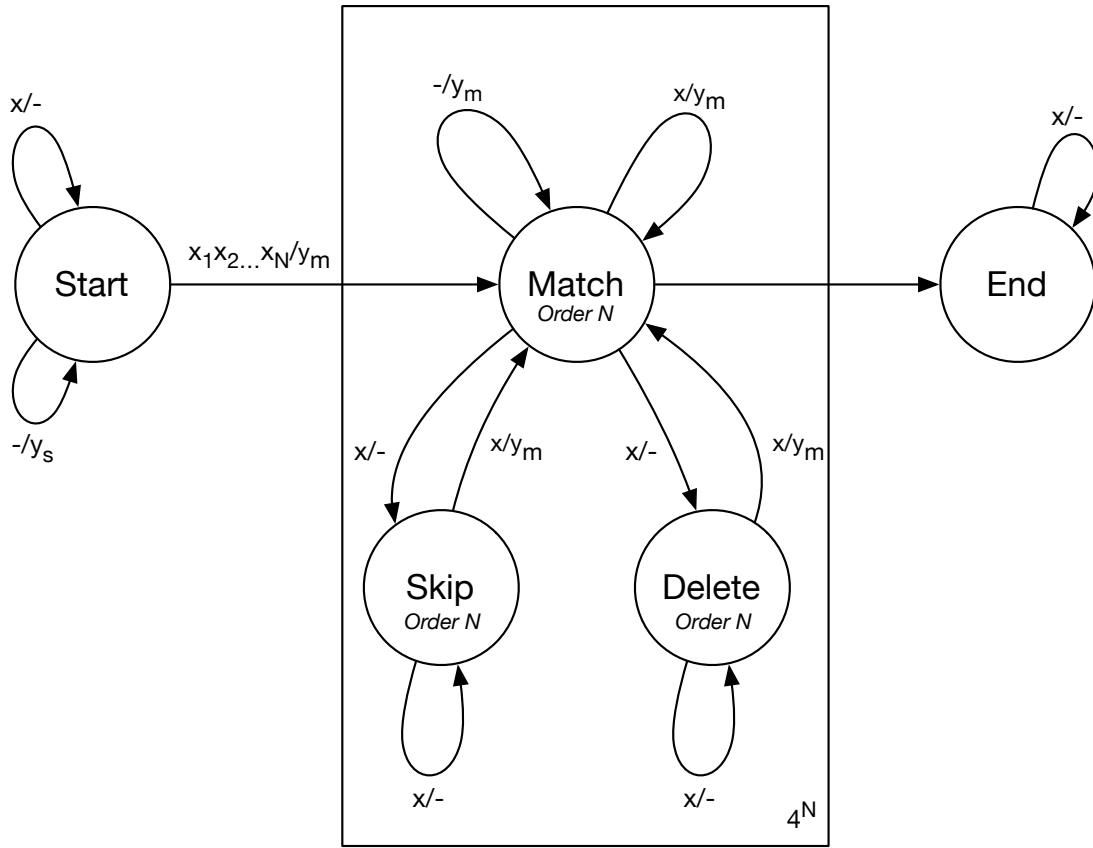
3.1 Null model

- 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

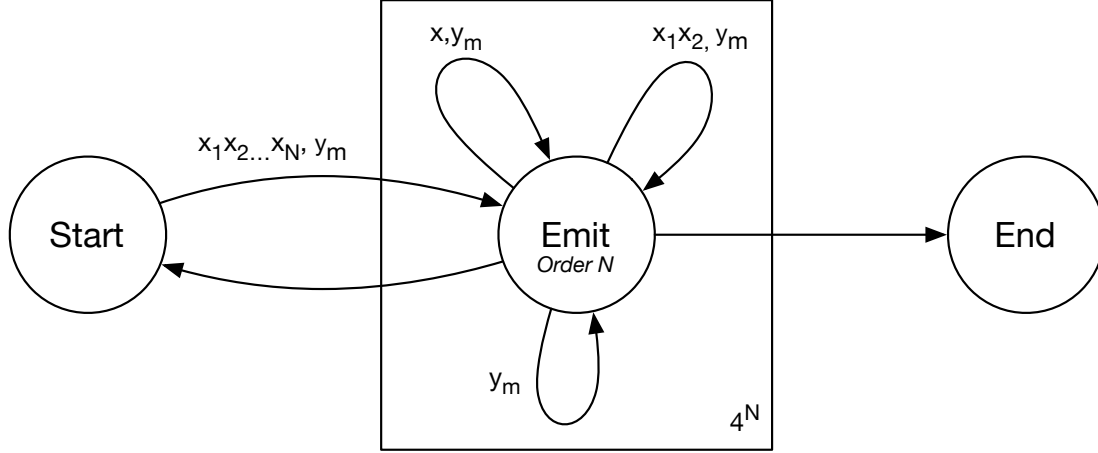


- Order- N Mealy transducer.
- Input alphabet: $\Omega = \{A, C, G, T\}$ (nucleotides)
- Output alphabet: real numbers partitioned into events, as with null model
- States: $\text{Start}, \text{End}, \{ \text{Match}_{x_1\dots x_N}, \text{Delete}_{x_1\dots x_N} : x_1\dots x_N \in \Omega^N \}$
- Parameters: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

Transducer can *skip* an individual base (no event emissions for that base), or can *delete* a run of bases (no event emissions during the run).

The transition weights for this transducer are shown in Section 8.4.

3.3 Basecalling model



- Order- N HMM.
- Unobserved transition labels: $\Omega = \{A, C, G, T\}$ (nucleotides)
- Output alphabet: real numbers partitioned into events, as with null & transducer models
- States: Start, End, $\{\text{Emit}_{x_1 \dots x_N}\}$
- Parameters: same as transducer model, plus length parameter p^{Emit} , and kmer probability distribution $q(x_1 \dots x_N)$ together with associated conditional distributions $q(x_{N+1}|x_2 \dots x_N)$ and $q(x_{N+1}, x_{N+2}|x_2 \dots x_N) = q(x_{N+1}|x_2 \dots x_N)q(x_{N+2}|x_3 \dots x_{N+1})$. Also define

$$\begin{aligned}
 p^{\text{LongDelete}} &= (p^{\text{Emit}})^N p^{\text{BeginDelete}} (p^{\text{ExtendDelete}})^{N-1} (1 - p^{\text{ExtendDelete}}) \\
 p^{\text{ShortDelete}}_{x_1 \dots x_{N+1}} &= (p^{\text{Emit}})^2 (1 - p^{\text{BeginDelete}}) p^{\text{Skip}}_{x_1 \dots x_N} (1 - p^{\text{Skip}}_{x_2 \dots x_{N+1}}) \\
 p^{\text{NoDelete}}_{x_1 \dots x_N} &= p^{\text{Emit}} (1 - p^{\text{BeginDelete}}) (1 - p^{\text{Skip}}_{x_1 \dots x_N})
 \end{aligned}$$

The transition weights for this HMM are shown in Section 8.5.

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).

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

8.4 Transition table for nanopore transducer

The following table gives the transition weights for the transducer introduced in Section 3.2.

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

8.5 Transition table for basecalling HMM

The following table gives the transition weights for the transducer introduced in Section 3.3.

Source	Destination	Weight	Unobserved	Observed
Start	Emit $_{x_1 \dots x_N}$	$q(x_1 \dots x_N)$ $\times (p_{x_1 \dots x_N}^{\text{MatchTick}})^{K_m} (1 - p_{x_1 \dots x_N}^{\text{MatchTick}})$ $\times \prod_{k=1}^{K_m} P(y_m^{(k)} \mu_{x_1 \dots x_N}^{\text{Match}}, \tau_{x_1 \dots x_N}^{\text{Match}}) dy_m^{(k)}$	$x_1 \dots x_N \in \Omega^N$	$\{y_m^{(k)} : 1 \leq k \leq K_m\},$ $K_m \sim \text{Geometric}(p_{x_1 \dots x_N}^{\text{MatchTick}}),$ $y_m^{(k)} \sim \text{Normal}(\mu_{x_1 \dots x_N}^{\text{Match}}, \tau_{x_1 \dots x_N}^{\text{Match}})$
Emit $_{x_1 \dots x_N}$	Emit $_{x_1 \dots x_N}$	$p_{x_1 \dots x_N}^{\text{MatchEvent}}$ $\times (p_{x_1 \dots x_N}^{\text{MatchTick}})^{K_m} (1 - p_{x_1 \dots x_N}^{\text{MatchTick}})$ $\times \prod_{k=1}^{K_m} P(y_m^{(k)} \mu_{x_1 \dots x_N}^{\text{Match}}, \tau_{x_1 \dots x_N}^{\text{Match}}) dy_m^{(k)}$		$\{y_m^{(k)} : 1 \leq k \leq K_m\},$ $K_m \sim \text{Geometric}(p_{x_1 \dots x_N}^{\text{MatchTick}}),$ $y_m^{(k)} \sim \text{Normal}(\mu_{x_1 \dots x_N}^{\text{Match}}, \tau_{x_1 \dots x_N}^{\text{Match}})$
Emit $_{x_1 \dots x_N}$	Emit $_{x_2 \dots x_{N+1}}$	$(1 - p_{x_1 \dots x_N}^{\text{MatchEvent}})$ $\times q(x_{N+1} x_2 \dots x_N) p_{x_2 \dots x_{N+1}}^{\text{NoDelete}}$ $\times (p_{x_2 \dots x_{N+1}}^{\text{MatchTick}})^{K_m} (1 - p_{x_2 \dots x_{N+1}}^{\text{MatchTick}})$ $\times \prod_{k=1}^{K_m} P(y_m^{(k)} \mu_{x_2 \dots x_{N+1}}^{\text{Match}}, \tau_{x_2 \dots x_{N+1}}^{\text{Match}}) dy_m^{(k)}$	$x_{N+1} \in \Omega$	$\{y_m^{(k)} : 1 \leq k \leq K_m\},$ $K_m \sim \text{Geometric}(p_{x_2 \dots x_{N+1}}^{\text{MatchTick}}),$ $y_m^{(k)} \sim \text{Normal}(\mu_{x_2 \dots x_{N+1}}^{\text{Match}}, \tau_{x_2 \dots x_{N+1}}^{\text{Match}})$
Emit $_{x_1 \dots x_N}$	Emit $_{x_3 \dots x_{N+2}}$	$(1 - p_{x_1 \dots x_N}^{\text{MatchEvent}})$ $\times q(x_{N+1}, x_{N+2} x_2 \dots x_N) p_{x_2 \dots x_{N+2}}^{\text{ShortDelete}}$ $\times (p_{x_3 \dots x_{N+2}}^{\text{MatchTick}})^{K_m} (1 - p_{x_3 \dots x_{N+2}}^{\text{MatchTick}})$ $\times \prod_{k=1}^{K_m} P(y_m^{(k)} \mu_{x_3 \dots x_{N+2}}^{\text{Match}}, \tau_{x_3 \dots x_{N+2}}^{\text{Match}}) dy_m^{(k)}$	$x_{N+1} x_{N+2} \in \Omega^2$	$\{y_m^{(k)} : 1 \leq k \leq K_m\},$ $K_m \sim \text{Geometric}(p_{x_3 \dots x_{N+2}}^{\text{MatchTick}}),$ $y_m^{(k)} \sim \text{Normal}(\mu_{x_3 \dots x_{N+2}}^{\text{Match}}, \tau_{x_3 \dots x_{N+2}}^{\text{Match}})$
Emit $_{x_1 \dots x_N}$	Start	$p^{\text{LongDelete}}$		
Emit $_{x_1 \dots x_N}$	End	$1 - p^{\text{Emit}}$		