

# Nanopore automata

Ian Holmes<sup>1,2,\*</sup>

**1** Lawrence Berkeley National Laboratory, Berkeley, CA, USA

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

## Contents

<b>1</b>	<b>Abstract</b>	<b>2</b>
<b>2</b>	<b>Specification</b>	<b>2</b>
2.1	Parameterization algorithm . . . . .	2
2.2	Reference search algorithm . . . . .	3
2.3	Implementation . . . . .	3
2.4	Evaluation . . . . .	3
<b>3</b>	<b>Methods</b>	<b>3</b>
3.1	Null model . . . . .	4
3.2	Homology model . . . . .	5
3.3	Baum-Welch algorithm . . . . .	6
3.4	Viterbi algorithm . . . . .	6
<b>4</b>	<b>Results</b>	<b>7</b>
<b>5</b>	<b>Discussion</b>	<b>7</b>
<b>6</b>	<b>Acknowledgments</b>	<b>8</b>
<b>7</b>	<b>Figure Legends</b>	<b>9</b>
<b>8</b>	<b>Appendix</b>	<b>10</b>
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

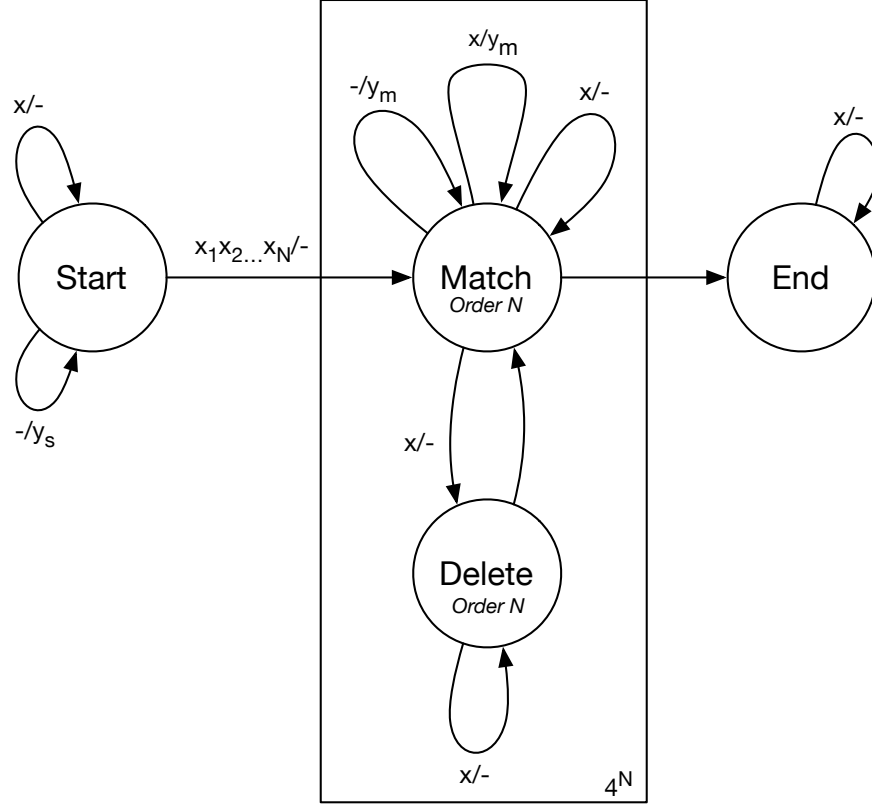
Model & inference algorithms.

### 3.1 Null model

- Output alphabet:  $\Re$  (real numbers signifying current levels)
- $K$  events. Sequence is  $y_1 \dots y_K$
- Parameters:  $p^{\text{NullEmit}}, \mu^{\text{Null}}, \tau^{\text{Null}}$
- Gaussian emissions:  $y_n \sim \text{Normal}(\mu^{\text{Null}}, \tau^{\text{Null}})$
- Probability is

$$P(y_1 \dots y_K) dy_1 \dots dy_K = (1 - p^{\text{NullEmit}}) \prod_{n=1}^K p^{\text{NullEmit}} P(y_n | \mu^{\text{Null}}, \tau^{\text{Null}}) dy_n$$

### 3.2 Homology model



- Order- $N$  Mealy transducer.
- Input alphabet:  $\Omega = \{A, C, G, T\}$  (nucleotides)
- Output alphabet:  $\Re$  (real numbers signifying current levels)
- 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:  $p^{\text{StartEmit}}, p^{\text{Skip}}, p^{\text{BeginDelete}}, p^{\text{ExtendDelete}},$   
 $\{ p_{x_1 \dots x_N}^{\text{MatchEmit}}, \mu_{x_1 \dots x_N}^{\text{Match}}, \tau_{x_1 \dots x_N}^{\text{Match}} : x_1 \dots x_N \in \Omega^N \}$

Transitions:

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

### 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  $\alpha$ , rate parameter  $\beta$ .  $\Gamma()$  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  $\mu$ , precision  $\tau$  (precision is reciprocal of variance).

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