

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	2
2.3	Implementation	3
2.4	Evaluation	3
3	Methods	3
3.1	Model	3
3.2	Baum-Welch algorithm	5
3.3	Viterbi algorithm	5
4	Results	6
5	Discussion	6
6	Acknowledgments	7
7	Figure Legends	8
8	Appendix	9
8.1	Exponential distribution	9
8.2	Gamma distribution	9

8.3 Normal distribution	9
-----------------------------------	---

1 Abstract

State machine algorithms for aligning Nanopore reads. Initial goal is simple reusable code for aligning a nanopore read to a reference sequence. No attempt at optimization yet.

2 Specification

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 transducer.
- Input: nucleotide
- Output: nucleotide, segment mean, duration
- Emissions:
 - categorical (base k -mer)
 - mixture of Normal/gamma (mean/duration)
- Transitions:
 - *Match*: emit single segment, absorb 1 base
 - *Insert*: affine gap insertion of bases: emits segments, absorbs no bases
 - *Delete*: affine gap deletion of bases: emits no segments, absorbs bases
 - *Merge*: emit single segment, absorb 2 or 3 bases

- *Split*: emit single segment, absorb 0 bases
- *Skip*: emit single segment, absorb $2 \dots K$ bases (large K , low extension penalty)

This can be achieved by a Mealy transducer with 4×4^N states. The factor of 4^N accounts for the order- N context. For each such context, the four states are MAT, INS, DEL and SKP.

Parameters:

- Gap opening & extension probabilities $\lambda_{go}, \lambda_{gx}$
- Merge probability λ_{mo} , probability that it's a 3-merge is λ_{mx}
- Split probability λ_s
- Skip probability λ_{ko} , skip extension probability λ_{kx}

In general the emissions are of the form

$$(y, m, d) \sim \text{CNE}(L)$$

where L is a “label” indexing the appropriate emission distribution

$$y \sim \text{Categorical}(\mathbf{p}_L)$$

$$m \sim \text{Normal}(\mu_L, \tau_L)$$

$$d \sim \text{Exponential}(\kappa_L)$$

where Ω is the nucleotide alphabet, $y \in \Omega$ is the nucleotide as decoded by the basecaller, $m \in \mathbb{R}$ is the mean segment current, and $d \in \mathbb{R}^+$ is the segment duration.

The transition table is as follows:

Transition	From	To	Weight	Input	Output
Match	MAT	MAT	$(1 - \lambda_{go})(1 - \lambda_{mo})(1 - \lambda_s)(1 - \lambda_{ko})$	$x \in \Omega$	$(y, m, d) \sim \text{CNE}(\text{match}, x, c)$
Insert	MAT	INS	$\lambda_{go}/2$	none	$(y, m, d) \sim \text{CNE}(\text{insert})$
	INS	INS	λ_{gx}	none	$(y, m, d) \sim \text{CNE}(\text{insert})$
	INS	MAT	$1 - \lambda_{gx}$	none	none
Delete	MAT	DEL	$\lambda_{go}/2$	$x \in \Omega$	none
	DEL	DEL	λ_{gx}	$x \in \Omega$	none
	DEL	MAT	$1 - \lambda_{gx}$	none	none
Merge	MAT	MAT	$\lambda_{mo}(1 - \lambda_{mx})$	$x \in \Omega^2$	$(y, m, d) \sim \text{CNE}(\text{merge2}, x)$
	MAT	MAT	$\lambda_{mo}\lambda_{mx}$	$x \in \Omega^3$	$(y, m, d) \sim \text{CNE}(\text{merge3}, x)$
Split	MAT	MAT	λ_s	none	$(y, m, d) \sim \text{CNE}(\text{split}, x)$
Skip	MAT	SKP	λ_{ko}	$x \in \Omega$	none
	SKP	SKP	λ_{kx}	$x \in \Omega$	none
	SKP	MAT	$1 - \lambda_{kx}$	none	none

Here $c \in \Omega^N$ is the input context.

3.2 Baum-Welch algorithm

3.3 Viterbi algorithm

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).

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