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. 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 3×4^N states. The factor of 4^N accounts for the order-N context. For each such context, the three states are MAT, INS and DEL.

Parameters:

- Gap opening & extension probabilities λ_{go}, λ_{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

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

where L is a "label"

$$\begin{aligned} y &\sim \text{Categorical}(\mathbf{p}_L) \\ m &\sim \text{Normal}(\mu_L, \tau_L) \\ d &\sim \text{Gamma}(\alpha_L, \beta_L) \end{aligned}$$

The transition table is as follows:

Transition	From	To	Weight	Input	Output
Match	MAT	MAT	$(1-\lambda_{go})(1-\lambda_{mo})(1-\lambda_s)(1-\lambda_{ko})$	$x \in \Omega$	$(y, m, d) \sim \mathrm{CNG}(\mathrm{match}, x, c)$
Insert	MAT	INS	$\lambda_{go}/2$		
	INS	INS	λ_{gx}		
	INS	MAT	$1-\lambda_{gx}$	none	none
Delete	MAT	DEL	$\lambda_{go}/2$		
	DEL	DEL	λ_{gx}		
	DEL	MAT	$1-\lambda_{gx}$	none	none
Merge	MAT	MAT			
Split	MAT	MAT			
Skip	MAT	MAT			

Here $y \in \Omega$ where Ω is the nucleotide alphabet and $c \in \Omega^N$ is the context.

3.2 Baum-Welch algorithm

3.3 Viterbi algorithm

4 Results

5 Discussion

6 Acknowledgments

7 Figure Legends

8 Appendix

8.1 Gamma distribution

$$\begin{aligned} x &\sim \text{Gamma}(\alpha, \beta) \\ \text{E}[x] &= \alpha/\beta \\ \text{Var}[x] &= \alpha/\beta^2 \end{aligned}$$

Shape parameter α, rate parameter β.

$$P(x|\alpha,\beta) = \frac{x^{\alpha-1}\beta^{\alpha}\exp(-x\beta)}{\Gamma(\alpha)}$$

where Γ is the gamma function

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

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

8.2 Normal distribution

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

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

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