Ryan Lorig-Roach
3/17/17

Nanopore Modelling Methods (abridged)

The progression of individual molecules through the nanopore is stochastic, and as such, it demands probabilistic modeling of the signal it produces. As a means of achieving this, a workflow has been established which removes the time-dependence of the signal output, and makes use of a Hidden Markov Model (HMM) to account for randomness in the signal as well as the randomness in the order of ligation of our polymer with accessory sequences. With these factors accounted for, it can be asserted how likely the presence of an experimental homopolymer is.

Methods

Preparing and running the control library:

Synthetic single stranded thymine homopolymers of 40nt length were ordered and used as a control to test the efficacy of our library preparation protocol and to aid in training the HMM that will be used to detect homopolymers. To give the polyT a highly recognizable label, it was annealed to a "barcode" with a polyA overhang. Polymerase filled in the remaining length of polyA to fully complement the 40nt of thymine, and 2 more sequencing adapters were ligated to the ends of the construct: a hairpin and a lead adapter (see appendix for diagram). The hairpin allows for both complements of the read to be threaded through the nanopore sequentially, while the lead adapter anchors the construct to the flowcell membrane and initiates entry into the pore.

An R9 MinION flowcell was used to analyze the library. This flowcell makes use of a mutated biological pore, which is maximally occupied by 6 base pairs of DNA at any given time during a translocation event. The sequence-dependent blockage of the pore produces a change in current through the opening. This level change is later identified in the recorded signal by a segmenting algorithm (developed by Kevin Karplus). Ideally, each segment can then be considered to be emitted from a single kmer of length k=6.

Construction of the HMM:

For the barcode and homopolymer, their nucleotide sequence was used to find the expected mean current for each kmer using previous data. Unknown sequences such as the proprietary hairpin and lead adapters had to be manually curated as "profiles", with no nucleotide identity attached to them. In either case, a list of currents can be produced which serves as the sequence 'x' of emissions which the HMM deals with.

Given a dictionary of meta-states (sequences) and their transitions to one another, their linear HMM is constructed (fig. 1B), and the overall HMM (fig. 1C) is then configured, with all bridge sequences dynamically created in the process (fig. 1D). If one or both of the flanking nucleotide sequences of a bridge are not known, a gradient between the flanking currents is produced, and their emission distributions are relaxed (with respect to standard deviation) to allow flexibility. Otherwise, if both flanks have known nucleotide sequences, a table of predetermined values for expected kmer currents is used to bridge the sequences.

Since the current recorded in the Nanopore is dependent on the movement of a single molecule, much of the HMM's configuration is dedicated to modeling the random rate and direction of molecular movement. Segmentation handles the finer noise seen in the sample (fig. 1A), and some of

the change in translocation rate, but larger scale changes in the net motion of the DNA often result in insertions. The configuration of the model can be justified as follows:

- Silent “skip” states are included to take into account the possibility that the translocation will move too quickly for detection.
- “Drop” states model real events in which the pore is temporarily (or permanently) blocked, producing an emission of little to no current.
- The “blip” state handles noise that did not get averaged out during segmentation. Over-segmentation was preferred over incomplete segmentation, and thus needed to be accounted for with blip states.
- Self-transitions in “emit” states also accommodate for over-segmentation, but have the alternate purpose of representing stalls which halt the progress of DNA’s translocation (especially prevalent at the hairpin).
- “Emit” states are allowed reverse transitions in the event that the DNA molecule randomly shifts backwards
- Reverse transitions were not allowed between junctions where one linear HMM branched into another. Logically, if the current state is within a bridge, this is only possible if the next meta-state is necessarily the other side of the bridge, otherwise the bridge state and its emissions should not be observed.
- Initial transitions between meta-states like the homopolymer and barcode were determined empirically.

The model was trained with the Baum-Welch algorithm on a minimal training set of 14 reads to determine whether alignment on 7 additional test reads was satisfactory. Running additional libraries became impossible as a result of a shortage of proprietary materials used for “2D” library preparation, so the volume of training data was a significant limiting factor during analysis.

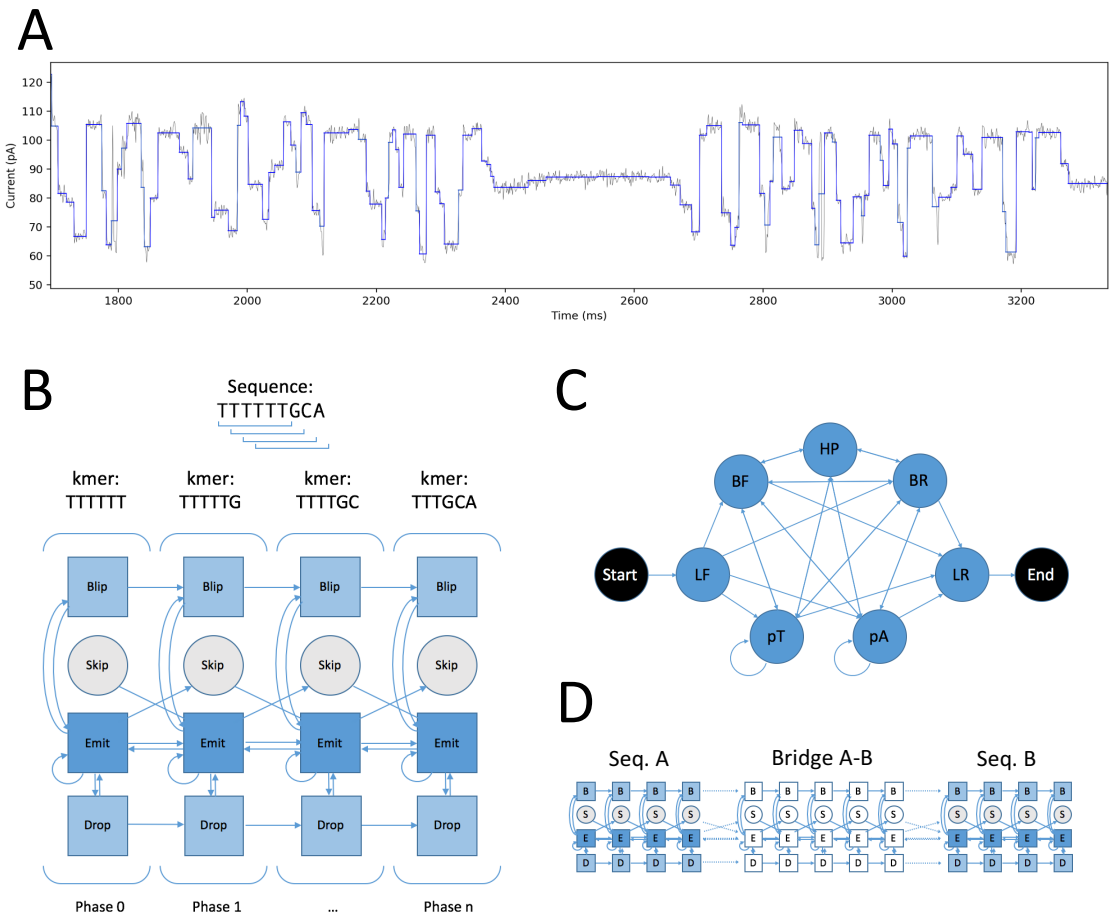

Figure 1: **(A)** An example of segmentation of an event in which a homopolymer is flanked by two reverse barcodes. The original signal is shown in black, with its segmented counterpart in blue. **(B)** The linear HMM that models each expected sequence, with the “emit” state producing kmer-specific currents. **(C)** State machine representing the possible events. Each circle is a “meta-state” corresponding to a linear HMM which models one of the fragments that make up the construct: forward lead adapter (LF), reverse lead adapter (LR), polyT (pT), polyA (pA), forward barcode (BF), reverse barcode (BR), and hairpin (HP). A short final sequence (not shown) was added between the end state and the reverse adapter to accommodate a rare termination in which several extra levels were observed**(D)** Transitions between linear HMMs also include additional states that contain the set of kmers that bridge any two sequences.

A null model was created, consisting of a single emission state with a normal emission distribution centered around 100pA, and standard deviation of 40pA. The emit-to-end transition for this state was adjusted so the expected total sequence length matched that of the positive reads. The log likelihood ratio of 7 positive and 7 negative test sequences was calculated using the null model and the construct model. The prior probability for the model is currently unknown.

Acknowledgements:

HMM Guidance:
Miten Jain
Ariah Mackey
Hugh Olsen

Provided code:
Kevin Karplus – segmenter (AKA speedyStatSplit)
Jacob Schreiber and Adam Novak – Yet Another Hidden Markov Model (YAHMM) package
Bryan Thornlow – Fast5 file parsing

Appendix:

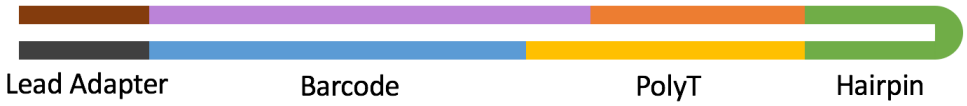

Figure 1: *The design for the sequencing construct is shown. The lead adapter and hairpin are necessary for all 2D reads, while the barcode (with overhang) is suited for the annealing and identification of the homopolymer.*