Minimap2: fast pairwise alignment for long DNA sequences (Li 2017)

Presenter: Taher Mun

Motivation

- Short-read aligners work great, for short reads
- Don’t work as well with long reads (SMRT, ONT):
  - Long reads are noisy
  - Seed+extend with BWT not great for reads that capture SVs

Existing software

- GraphMap
    - Indexes with seeds
    - Uses graph-based approach for chaining seeds
    - Produces full alignment
- Mashmap
    - Indexes with minimizers
    - Maps by comparing minhash sketches
- Minimap
    - Indexes with minimizers
    - Chains seeds through single linkage clustering
    - Does not produce full alignment (only maps)

Minimap (Li 2015) in more detail

https://github.com/lh3/minimap#algorithm-overview

1. Index reference by collecting minimizers into hash table
2. For each query, get minimizers, looked them up in hash table, record positions on ref and query
3. Cluster matches within ~500bp windows, merge windows as necessary.
4. Create chains from each cluster by solving LIS problem for reference minimizers. Break chain when gap is too large
5. Output the start and end of the chain if it has enough matches and is long enough
6. Repeat from 1 is more target matches exist.

Minimap2

- Next iteration of minimap
- Adds more accurate chaining
- Adds DP step to produce CIGAR string

Algorithm

https://github.com/lh3/minimap2#algorithm-overview

1. Index reference minimizers in a hash table
2. Minimizers in query that match a non-frequent minimizer in reference are used as *seeds*
3. Sort seeds by position in the reference. Chain them with dynamic programming. Each chain represents a potential mapping.
4. Set of primary mappings created by going down list of chains by descending chain score and adding them if they do not overlap with an existing chain in the set (otherwise they are secondary mappings)
5. Retain all primary mappings, along with top secondary mappings that reach a score threshold
6. (optional) filter out bad seeds, extend from left-most seed by global alignment. Split when score drops by 400. Then also extend from right-most seed.
7. Chains+alignments output in SAM file format.

Chaining

- Use DP to create chains, according to this formula

$$f(i) = \max\left\{ \max_{i>j\geq 1} \{f(j) + d(j,i) - \gamma(j,i)\}, w_i \right\} \tag{1}$$

- f(i) is chaining score. d(j,i) is matches between anchors, y is gap penalty
- Heuristic - start from i-1, only do *h* iterations for each anchor (h=50)
- Overlapping and small chains are filtered out
    - ^ the time saver

Alignment

- Uses DP algorithm from https://github.com/ocxtal/minialign
- “Banded alignment” - don’t fill whole DP matrix
- Optimization possible due to score cap from gap cost, max matching score
- Breaks global alignment if score drops between chains
    - Local alignment here, might find inversions

Results

See figure a, b in paper

On simulated reads:

- Minimap has consistently higher recall
- Mapping quality consistent with repetitiveness/error
- >30x faster than existing aligners

Conclusion

- Minimap2 is pretty fast, deals well with the noise
- Can be improved by using better banded DP heuristics
- Could be used for whole-genome alignment (eg. instead of nucmer)

References

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv:1303.3997 [q-Bio].

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics *32*, 2103–2110.

Li, H. (2017). Minimap2: fast pairwise alignment for long DNA sequences. ArXiv:1708.01492 [q-Bio].

Sovic, I., Sikic, M., Wilm, A., Fenlon, S.N., Chen, S., and Nagarajan, N. (2015). Fast and sensitive mapping of error-prone nanopore sequencing reads with GraphMap.

Suzuki, H. (2017). minialign: fast and accurate alignment tool for PacBio and Nanopore long reads.