MIRA Internals

Michael Cusack
5/17/10
BME235

Purpose

- **M**imicking **I**ntelligent **R**ead **A**ssembly
  - "strategies used by human experts"
- **Difficult genomes**
  - Lots of **repeats** or other sequence aberrations
- **Hybrid** Assembly
  - Combing **several data types**
  - Using all available data

Data (That Can Be) Used

1. the **initial trace data**, representing the gel electrophoresis signal
2. the called nucleic acid **sequence** (*required*)
3. **position specific confidence values** for the called bases of the nucleic acid sequence
4. a stretch in each sequence marked as **HCR**
5. **general properties** like direction of the clone read and name of the sequencing template etc.
6. **special sequence properties** in different regions of a read (like sequencing vector, known standard repeat sequence and known SNP sites etc.) that have been tagged or marked.

Read Scanning (Fast Error Tolerant Pair-wise Comparisons)

Both are less sensitive then Smith-Waterman, but much faster.

DNA-Shift-AND

- O(c*n), c=# allowed errors
- Takes words from start, middle and end of read1 and searches each in read2
- Must find 2 of 3 to establish relationship

ZEBRA

- Transcribe, Divide, Reorganize, Concentrate and Conquer strategy
- Hashes each octet of bases (16-bit int) and creates hash index table

More Thorough Comparison to Establish Type of Relationship

- Once initial relationships are established, MIRA uses a **modified Smith-Waterman** algorithm to perform local alignment of overlap
- Uses **banding**
- Uses information generated from DNA-SAND/ZEBRA

Building Graph

- Overlap alignment + complementary data (orientation, overlap region, score, etc) is called an **aligned dual sequences** (ADS) and kept in memory if passes S-W
- Good alternatives also kept
- ADS's create **weighted** (by score) **overlap graph**(s)
- Each unconnected graph is a possible contig

Iterative Process

- Start with **highest quality**
  - Each read is split into a **high confidence region (HCR)** and a **low confidence region (LCR)** by **quality clipping**
  - Only HCR bases are used to build **initial contigs**
  - LCR bases are used **cautiously**

Creating Contigs

Pathfinder

- Finds best nodes (those with highest scoring overlaps in HCRs)
  - **Anchors**
- Extends in such a way that the **uncertainties of the consensus bases are lowest**
- Uses a **n, m-step recursive look-ahead** algorithm to detect repeats

Contig Builder

- Once a path is decided each contig must be **compiled** and **approved**
- If a read along path is **overall too different** from existing **consensus** despite high scoring overlap, it is **rejected** and the **pathfinder is run again** from that point

Independent Observations

According to the author:
(from *http://www.freelists.org/post/mira_talk/How-does-Mira-determine-quality-scores,2*)

One **central pillar** of the **quality calculation** in MIRA is the rule that **independent observations** of a base confirm this base better that non-independent observations. When a base was **read from both directions**, one can assume independence of observations: it's not the whole truth, but close enough. As a side note: **observing a base with different sequencing technologies** also constitutes independent observations.

Repeats

- Can be told when there are **known repeated elements.**
  - Such as ALU repeats in humans.
- When these regions/reads are detected much **stricter control mechanisms** can be applied.
- When there is a **discrepancy** in a read matching a repeated element, **signal processing of the trace** is used to determine if the error is **explainable.**
- If **percentage of unexplainable errors** is greater than threshold (default: 1%), reads are **rejected from consensus** and **returned to assembly graph**.