PerM: efficient mapping of short sequencing reads with periodic full sensitive spaced seed

Esha Desai
USC ID: 6993245898

Why PerM?

- Need for highly efficient methods to align short DNA read
- PerM (Periodic Seed Mapping) is a mapping software that uses periodic spaced seeds which significantly improves mapping efficiency for large reference genomes when compared with state-of-the-art programs.
- Allows entire genomes to be loaded to memory, while multiple processors simultaneously map reads to the reference.
- Data structure in PerM requires only 4.5 bytes per base to index the human genome
- Full sensitivity for up to three mismatches
- High sensitivity for four and five mismatches
- Minimizes the number of random hits per query
- Hence significantly speeding up the running time.

Methods used by PerM

- PerM's periodic seeds allow to increase mapping efficiency and sensitivity using the following methods:

1. Seed notation
2. Motivation for Periodic Design
3. Periodic seeds: generalization, indexing and extendability
4. Seed-search algorithm
5. Implementation and bitwise encoding

The experimental results, the periodic spaced seeds used in PerM outperform the seeds used in MAQ in terms of mapping speed and sensitivity for both Illumina and SOLiD data

Question

What are the possible solutions to improve performance with PerM?

- For full sensitivity to many mismatches on a short read, single periodic seeds may prove incapable of providing efficient mapping performance.
- Showed that the optimization of multiple seeds cannot be easier than the Golomb Ruler Design problem, considered likely to be NP-hard.
- The costly step of hashing to multiple index tables may be necessary to increase seed weight and eliminate a bottleneck in the checking step.
- Proposal of three methods to design high weight multiple seeds: a constrained exhaustive search, a reduction to the integer programming problem and a ‘tuples-grouping’ algorithm.