CS481: Bioinformatics Algorithms

Can Alkan
EA509
calkan@cs.bilkent.edu.tr

http://www.cs.bilkent.edu.tr/~calkan/teaching/cs481/
Heuristic Similarity Searches

- Genomes are huge: Smith-Waterman quadratic alignment algorithms are too slow
- Alignment of two sequences usually has short identical or highly similar fragments
- Many heuristic methods (i.e., FASTA) are based on the same idea of *filtration*
  - Find short exact matches, and use them as seeds for potential match extension
  - “Filter” out positions with no extendable matches
PatternHunter: faster and even more sensitive

- BLAST: matches short consecutive sequences (consecutive seed)
  - Length = $k$
  - Example ($k = 11$):
    
    11111111111

    Each 1 represents a “match”

- PatternHunter: matches short non-consecutive sequences (spaced seed)

- Increases sensitivity by locating homologies that would otherwise be missed

- Example (a spaced seed of length 18 w/ 11 “matches”):

  111010010100110111

  Each 0 represents a “don’t care”, so there can be a match or a mismatch
Spaced seeds

Example of a hit using a spaced seed:

```
GAGTACTCAACACCAACATTAGTGGCAATGGAAT...
|||       |||||       |||||       |||||       |||||       |||
GAATACTCAACAGCAACACTAATGGCAGCAGAAAT...
111010010100110111
```
Why is PH better?

- **BLAST**: redundant hits
  - This results in > 1 hit and creates clusters of redundant hits

- **PatternHunter**: This results in very few redundant hits
Why is PH better?

BLAST may also miss a hit

GAGTACTCAACACACCACATTTAGTGGGCAATGGAAAT

GAATACTCAACAGCAACATCAATGGGCAGCAGAAAT

9 matches

In this example, despite a clear homology, there is no sequence of continuous matches longer than length 9. BLAST uses a length 11 and because of this, BLAST does not recognize this as a hit!

Resolving this would require reducing the seed length to 9, which would have a damaging effect on speed.
Advantage of Gapped Seeds

11 positions
11 positions
10 positions
Why is PH better?

- Higher hit probability
- Lower expected number of random hits
Use of Multiple Seeds

Basic Searching Algorithm
1. Select a group of spaced seed models
2. For each hit of each model, conduct extension to find a homology.
Another method: BLAT

- BLAT (BLAST-Like Alignment Tool)
- Same idea as BLAST - locate short sequence hits and extend
BLAT vs. BLAST: Differences

- BLAT builds an index of the database and scans linearly through the query sequence, whereas BLAST builds an index of the query sequence and then scans linearly through the database.

- Index is stored in RAM which is memory intensive, but results in faster searches.
BLAT: Fast DNA Alignments

Steps:
1. Break DNA into 500 base chunks.
2. Use an index to find regions in genome similar to each chunk of DNA.
3. Do a detailed alignment between genomic regions and DNA chunk.
4. Use dynamic programming to stitch together detailed alignments of chunks into detailed alignment of whole.
BLAT: Indexing

- An index is built that contains the positions of each $k$-mer in the genome
- Each $k$-mer in the query sequence is compared to each $k$-mer in the index
- A list of ‘hits’ is generated - positions in DNA and in genome that match for $k$ bases
Indexing: An Example

Here is an example with $k = 3$:

Genome: cacaattatcacgaccgc

3-mers (non-overlapping): cac aat tat cac gac cgc

Index: aat 3  gac 12
        cac 0,9  tat 6
        cgc 15

Multiple instances map to single index

cDNA (query sequence): aatttctcac

3-mers (overlapping): aat att ttc tct ctc tca cac

0 1 2 3 4 5 6

Hits: aat 0,3
      cac 6,0
      cac 6,9

clump: cacAATtatCACgaccgc

Position of 3-mer in query, genome
However...

- BLAT was designed to find sequences of 95% and greater similarity of length >40; may miss more divergent or shorter sequence alignments
PatternHunter and BLAT vs. BLAST

- PatternHunter is 5-100 times faster than Blastn, depending on data size, at the same sensitivity.
- BLAT is several times faster than BLAST, but best results are limited to closely related sequences.