Algorithms for Computational Biology

Zsuzsanna Lipták

Masters in Molecular and Medical Biotechnology
a.a. 2015/16, fall term

Database search with BLAST (summary)

Database search

- Until now: compare two sequences
  - how similar/different are they? (score/value)
  - where are the similarities/differences? (alignment)

Database search

- Until now: compare two sequences
  - how similar/different are they? (score/value)
  - where are the similarities/differences? (alignment)

- Now: compare one sequence to a database (i.e. to many sequences)

Database search

Goal:
Identifying sequences in the DB which have high local similarity with the query.

- We know how to do this: Smith-Waterman DP-algorithm.
- But: too slow!

Say all sequences have length n (query t and all DB seq's), and there are r sequences in the DB.

- exact solution (Smith-Waterman): $O(r \cdot n^2)$

Example

- UniProt/SwissProt (protein database): 548 454 sequences, 195 409 447 aa's (avg. length 350 aa's) version 29/04/15
- NCBI Genbank (nucleotide database): 182 188 746 sequences, 189 739 230 107 nucleotides (avg. length 1041 nucl.) April 2015, no WGS

Say all sequences have length n (query t and all DB seq's), and there are r sequences in the DB.

- exact solution (Smith-Waterman): $O(r \cdot n^2)$

Example

- UniProt/SwissProt (protein database): 548 454 sequences, 195 409 447 aa's (avg. length 350 aa's) version 29/04/15
- NCBI Genbank (nucleotide database): 182 188 746 sequences, 189 739 230 107 nucleotides (avg. length 1041 nucl.) April 2015, no WGS

So we would get something like $350 \cdot 350 \cdot 548454 = 67\,185\,615\,000 =$ about 67 billion ($67 \cdot 10^9$) steps, which takes 18 hours on a computer that performs 1 million operations per second (for UniProt), and 197 434 482 454 026 ($\approx 1.9 \cdot 10^{12}$), about 6 years, for Genbank. And still about 1 hour on a computer performing 1 billion operations per second.

Say all sequences have length n (query t and all DB seq's), and there are r sequences in the DB.

- exact solution (Smith-Waterman): $O(r \cdot n^2)$

Example

- UniProt/SwissProt (protein database): 548 454 sequences, 195 409 447 aa's (avg. length 350 aa's) version 29/04/15
- NCBI Genbank (nucleotide database): 182 188 746 sequences, 189 739 230 107 nucleotides (avg. length 1041 nucl.) April 2015, no WGS

So we would get something like $350 \cdot 350 \cdot 548454 = 67\,185\,615\,000 =$ about 67 billion ($67 \cdot 10^9$) steps, which takes 18 hours on a computer that performs 1 million operations per second (for UniProt), and 197 434 482 454 026 ($\approx 1.9 \cdot 10^{12}$), about 6 years, for Genbank. And still about 1 hour on a computer performing 1 billion operations per second.

And this is for one query only!

BLAST: Basic Local Alignment Search Tool

- Altschul et al. 1990, 1997
- looks for sequences in a database with high local similarity to query
- heuristic algorithm
- solid mathematical foundations (Karlin-Altschul statistics)
- extremely successful, now the database search tool ("to blast a sequence against a database")
- NCBI[1] Blast at: `http://blast.ncbi.nlm.nih.gov/Blast.cgi`

[1] NCBI = National Center for Biotechnology Information

Basic idea

Basic idea

If there is a good local alignment between two sequences, then this local alignment is likely to contain two short substrings with high score when aligned without gaps.

Basic steps of BLAST

1. create list of high-scoring words with query
2. scan DB for these words (called seeds)
3. extend seeds in both directions to form good local alignment (these are called MSPs = maximum segment pairs)

BLAST then gives a significance score to the MSPs and only retains them if above a certain threshold.

BLAST2

Some innovations of BLAST2 (Altschul 1997)

- start with two seeds instead of one, not too far apart
- gapped alignments
- extension of statistical theory to HSPs (high-scoring segment pairs)

The NCBI BLAST website

- Different versions of BLAST, depending on the task (nucl-nucl: blastn, megablast, . . . , prot-prot: blastp, psi-blast, nucl-prot: blastx, prot-nucl: tblastn, . . .)
- Different databases (nucl vs. prot, different organisms, different types of db, different levels of assembly, . . .)
- Very good explanations and help pages!
- If you haven't done it yet, then you should try it and play around! E.g. download a sequence from Genbank or Swissprot, modify it and blast it!