Special course in Computer Science: Advanced Text Algorithms

Lecture 9: Database Sequence Searching

Eugen Czeizler
Department of IT, Abo Akademi
http://combio.abo.fi/teaching/textalg/

(slides originally by I. Petre, E. Czeizler, V. Rogojin)

Database Sequence Searching

- Bio-sequence databases represent one of the most important applications of string algorithms in computational biology
  - These databases contain huge collections of DNA, RNA or protein sequences

- Biological discoveries based on sequence similarity are now routine.

Relevance of DB Sequence Searching

- One of the first such discoveries was the connection between *oncogenes* and growth-regulating proteins.

- *Simian sarcoma is a retrovirus that causes cancer in* monkeys

- It's *oncogene called v-sis was sequenced in 1983* (Doolittle et. al)

- Comparing the amino acid sequence encoded by *v-sis* against published protein sequences revealed significant similarity with a *growth factor called PDGF*

Relevance of DB Sequence Searching

- The genome of bacterium *Haemophilus influezae* was sequenced in 1995 (Fleischmann et al.)
  - 1,743 assumed coding regions were translated into amino acid sequences, and searched for similarity in the Swiss-Prot database
  - 1,007 of them matched s.t. the biochemical function could be deduced for each of them

Heuristic Database Searching

- Exact similarity computation between a query string and database sequences takes $\Theta(nm)$ time using dynamic programming approach

- With current technology aligning a query against the entire database is not feasible, even with special purpose hardware
  - there are specialized chips for sequence alignment, and services that search databases on a 4,000-processor computer

Heuristic Search Applications

- Sequences similar to a new one are normally searched with fast heuristic methods, before (or instead of) exact similarity computation

- Dominant search applications: **FASTA** and **BLAST**
  - these are actually suites of programs tuned for different problem domains (DNA, protein, DNA translated to protein etc.)
  - exclude large parts of the DB from more careful and time-consuming examination
  - do not permit precise analyzes of speed or accuracy

Heuristic Methods: FASTA and BLAST

FASTA

- Short for “fast-all”
- First fast sequence searching algorithm for comparing a query sequence against a database.

BLAST

- Basic Local Alignment Search Technique
- It is an improvement of FASTA with respect to search speed, ease of use, statistical rigor.

Common Ideas

- A search is performed by finding good local alignments between the query sequence and the DB sequences
  - good alignments usually include short identical or highly similar fragments
- Thus, exact or highly similar occurrences of query subwords are first located
- Then these are extended to longer alignments of sufficiently high similarity

FASTA

- FASTA is one of the first widely used programs for searching protein and DNA sequence databases (Lipman & Pearson, 1985, 1988)
  - Current version (36) was released in March 2010

- The method first looks for exact matches between words in query and test sequence
  - Let P be a query sequence, and T any database sequence (each considered in turn)
  - A user-specified parameter ktup is used for locating exact occurrences of ktup-length substrings (k-tuples) of P in T
  - For DNA the ktup is usually 6
  - For proteins the ktup is usually 2

- These local alignments are then extended to longer alignments

FASTA on the Web

- Many websites offer **FASTA** searches

- Institut de Génétique Humaine, Montpellier France, GeneStream server
  http://www2.igh.cnrs.fr/bin/fasta-guess.cgi
- European Bioinformatics Institute, Cambridge, UK
  http://www.ebi.ac.uk/htbin/fasta.py?request
- EMBL, Heidelberg, Germany
  http://www.embl-heidelberg.de/cgi/fasta-wrapper-free
- Munich Information Center for Protein Sequences (MIPS)
at Max-Planck-Institut, Germany
  http://speedy.mips.biochem.mpg.de/mips/programs/fasta.html
- Institute Pasteur, France
  http://central.pasteur.fr/seqanal/interfaces/fasta.html
- National Cancer Center of Japan
  http://bioinfo.ncc.go.jp

BLAST

- **BLAST** (Altschul, Gish, Miller, Myers & Lipman, 1990) is the dominant search program for bio-sequence databases
  - Current version is 2.2.24 (August 2010), http://www.ncbi.nlm.nih.gov/BLAST/
- Based on the same assumption as **FASTA** that good alignments contain short lengths of exact matches
- Both **BLAST** and **FASTA** search for local sequence similarity
  - although they have exactly the same goals, though they use somewhat different algorithms and statistical approaches.

BLAST

- In **BLAST** we have the confluence of three lines of research:
  - Lipman et al. to improve hot-spot selectivity
  - sub-linear expected-time approximate matching of Myers
  - probability estimates for the statistical significance of reported matches by Karlin, Altschul & Dembo

- **BLAST** benefits
  - Speed
  - Statistical rigor
  - More sensitive
  - User friendly

BLAST

- Input:
  - Query sequence P
  - Database of sequences DB
  - Minimal score S
- Output:
  - Sequences from DB (T), such that P and T have scores > S

BLAST Functionality and Concepts

- **BLAST** searches for local regions of high similarity without gaps between the query sequence and each database sequence
  - It does not require identical words like **FASTA**
  - Intuition: similar regions of equal length between proteins suggest functional similarity; Insertions/deletions tend to change the shape and thus the function of a protein

BLAST Concepts

- **BLAST** considers substrings of a given lenght k called words (the k-tuples in **FASTA**)
  - For proteins it searches for 3 amino acids-long subsequences (i.e., k=3), and for DNA sequences it searches for 11 bases-long subsequences (i.e., k=11)

- **BLAST** locates words in the DB sequence (T) and words in the query sequence (P) having an alignment without gaps with score above a fixed threshold t

- These hot-spots (called hits) are then extended into segment-pairs with score above C, if possible
  - Hits are located applying *exact set matching*

BLAST Concepts

- Given strings P (query) and T (DB sequence), a segment pair is a pair of substrings P′ of P and T′ of T aligned without spaces
  - |P′|=|T′|

- A segment pair is locally maximal if it cannot be extended or shortened at either end without decreasing its score

- A segment pair of maximal score is called a maximal segment pair (MSP)

- **BLAST** tries to find and report all DB sequences that have an MSP with P above a cutoff score C

Locating and Examining Hits

- For searching protein sequences, **BLAST** constructs for each substring α of length k of P a t-neighborhood: all strings of length k having similarity ≥ t with α

- Then a database sequence T is scanned to locate words from all the t-neighborhoods derived from P; these are called hits

- **BLAST** tries to extend each hit into a locally maximal segment pair with score above C;

Efficiency of BLAST

- **BLAST** avoids quadratic-time dynamic programming
  - told to be > 50 faster than Smith-Waterman local alignment algorithm

- Word length k and neighborhood similarity bound t need to be selected to minimize
  - the probability of missing an MSP of score above C
  - the size of t-neighborhoods, and
  - the frequency of hits

- For proteins, values like k =3–5 and t = 17 (with a version of PAM matrices) are reported as good compromises

Blast variants

- BLAST contains 5 programs, which differ in the types of sequences they compare and at what level.

Program	Query seq type	DB seq type	Alignment level
blastn	Nucleotide	Nucleotide	Nucleotide
blastp	Protein	Protein	Protein
blastx	Nucleotide	Protein	Protein
tblastn	Protein	Nucleotide	Protein
tblastx	Nucleotide	Nucleotide	Protein

Using BLAST for protein sequences

- **BLASTP** compares a protein sequence against a protein database
  - It is used for instance when one wants to find the function of a given protein sequence
  - It identifies common regions between proteins or it identifies related protein sequences

- **TBLASTN** compares a protein sequence with a nucleotide database
  - It is used for instance when one wants to identify new genes encoding a given protein
  - It compares the query protein sequence with DNA sequences which are first translated into their 6 possible reading frames
  - It maps a protein to genomic DNA

Using BLAST for DNA sequences

- **BLASTN** compares a DNA sequence against a DNA database
  - It is used for instance when one is interested in screening repetitive elements, cross-species sequence exploration, annotating genomic DNA, etc

- **TBLASTX** compares a DNA sequence translated into a protein against a DNA database which is also translated into proteins
  - It is used for instance when one wants to do a cross-species gene prediction at the genome or transcription level (EST map) or search for proteins which are not yet in databases

Using BLAST for DNA sequences

- **BLASTX** compares a DNA sequence translated into a protein against a protein database
  - It is used for instance when one wants to find protein-encoding genes in genomic DNA, or when one wants to see whether a given DNA sequence corresponds to a known protein.

Using BLAST for DNA sequences

- Using **BLAST** for DNA sequences involves similar operations as in the case of protein sequences but it does not work as well.

- It is faster and more accurate to use **BLAST** for protein sequences
  - If the reading frame of a given DNA sequence is known then it is better to translate the DNA sequence into a protein sequence and then use the search for protein sequences

Conclusions

- **BLAST** is the most used tool for searching sequence databases
  - It is fast and very reliable
  - It does not find necessarily the best global alignment, but usually it finds the best matching subsequences in the database
  - It is user friendly, quite easy to use with the default parameters
  - There is a solid statistical framework for interpreting the results, i.e., evaluating the scores