Data structure: Lookup Table

Application: BLAST

The Look-up Table Data Structure

Definitions:

- A *k-mer* is a string of length *k*.
- A *lookup table* is a table of size $|\Sigma|^k$ that stores the location of all *k-mers* in the input sequence(s)

1 2 3 4 5 6 7 8

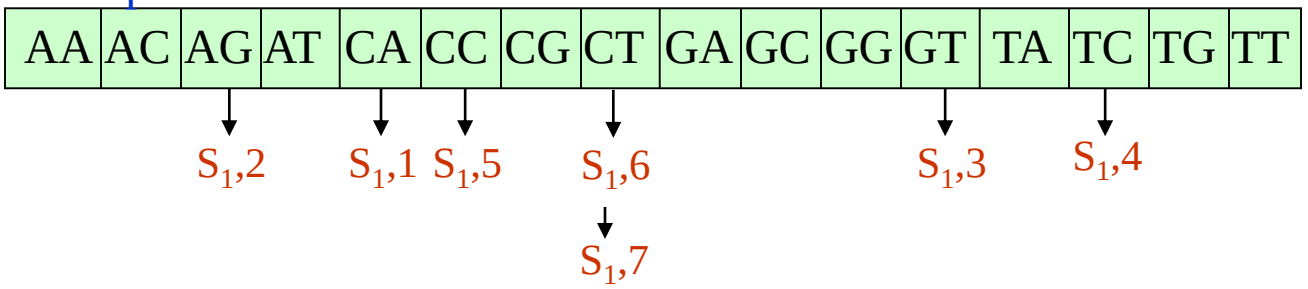
S_1: C A G T C C T C

Σ={A,C,G,T}

k = 2

→ 4^2 (=16) entries in lookup table

Lookup table:



Lookup Table - Application

- How to build lookup tables?
  - (see lecture notes)

- BLAST : basic local alignment search tool

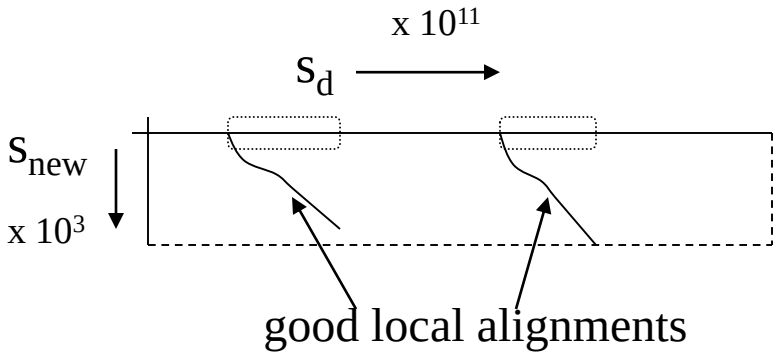
Need for a Fast Alignment Method

- What to do with a newly found gene candidate, s_{new}?
- Locate "similar" genes in GenBank

One-to-many

One Approach: (database search)

1. Concatenate all sequences in our genomic database into one sequence, say s_d
2. Compute the local alignment between s_{new} and s_d
3. Report all "significant" local alignments



Run-time: $O(|s_d|.|s_{new}|)$

Very long query time !!

Basic Local Alignment Search Tool (BLAST)

- Altschul *et al.* (1990) developed a program called BLAST to quickly query large sequence databases

- **Input:**
  - Query sequence q and a sequence database D

- **Output:**
  - List of all significant local alignment hits ranked in increasing order of *E-value* (aka *p-value*, which is the probability that a random sequence scores more than q against D).

The BLAST algorithm

0. Preprocess: Build a *lookup table* of size $|\Sigma|^k$ for all k-length words in D

1 2 3 4 5 6 7

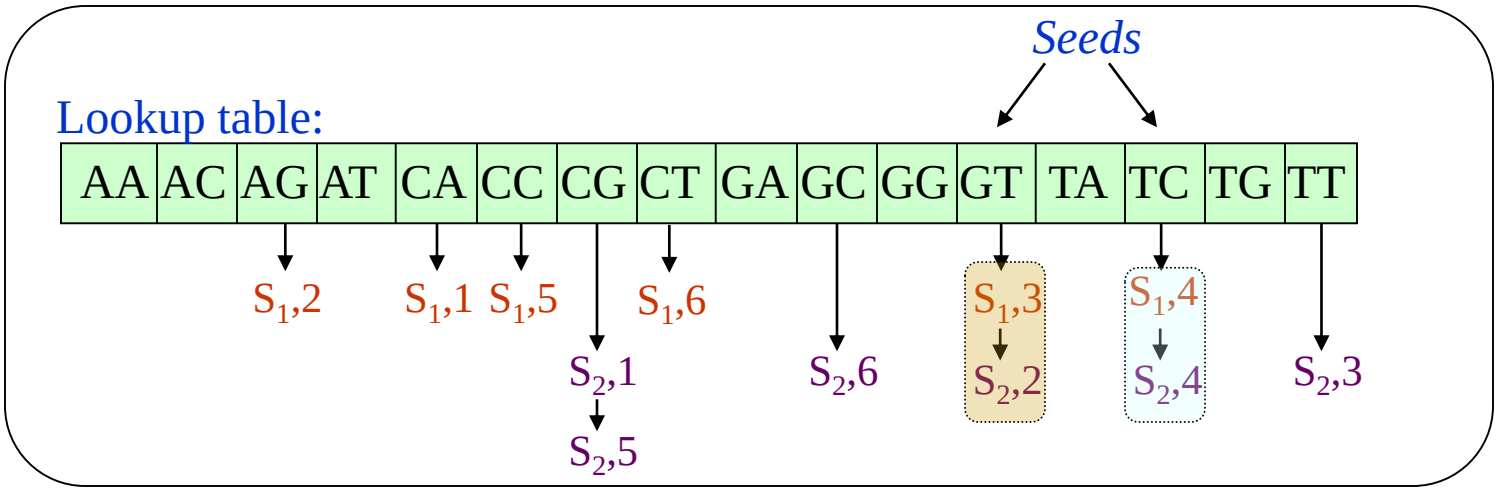
S_1: C A G T C C T

S_2: C G T T C G C

Σ={A,C,G,T}

k = 2

→ 4^2 (=16) entries in lookup table



Preprocessing is a one time activity

BLAST Algorithm …

1. Identify Seeds: Find all *k*-mers in *q* that are also in D using the lookup table

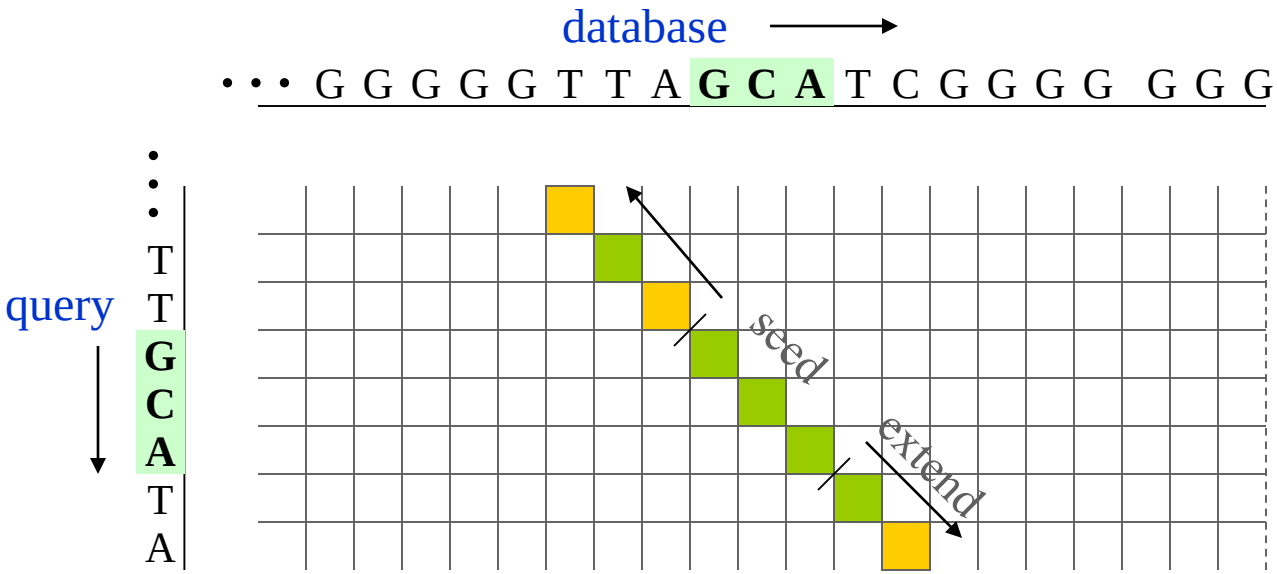
2. Extend seeds: Extend each seed on either side until the aggregate alignment score falls below a threshold
   - Ungapped: Extend by only either matches or mismatches
   - Gapped: Extend by matches, mismatches or a limited number of insertion/deletion gaps

3. Record all local alignments that score more than a certain statistical threshold

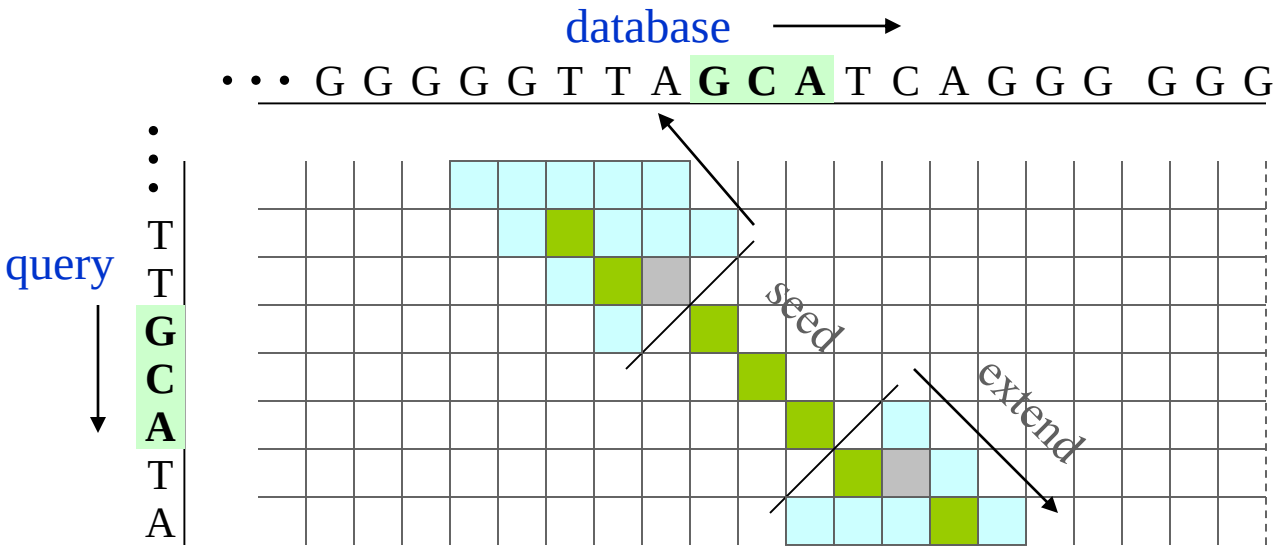
4. Rank and report all local alignments in non-decreasing order of *E-value*

Illustration of BLAST Algorithm

Compute only a band of diagonals around the seed



Ungapped Extension



Gapped Extension (over a band of diagonals)

Different Types of BLAST Programs

Program	Query	Database
blastn	nucleotide	nucleotide
blastp	protein/peptide	protein/peptide
blastx	nucleotide	protein/peptide
tblastn	protein/peptide	nucleotide
tblastx	nucleotide	nucleotide

http://www.ncbi.nlm.nih.gov/blast

Selected Bibliography for Alignment Topics

Papers

- S. Needleman and C. Wunsch (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J. Molecular Biology*, 48:443-453.
- D. Hirschberg (1975). A linear space algorithm for computing maximal common subsequences. *Communications of the ACM*, 18(6):341-343.
- T. Smith and M. Waterman (1981). Overlapping genes and information theory, *J. Theoretical Biology*, 91:379-380.
- O. Gotoh (1982). An improved algorithm for matching biological sequences. *J. Molecular Biology*, 162(3):705-708.
- J. Fickett (1984). Fast optimal alignment. Nucleic Acids Research, 12(1):175-179.
- M.S. Gelfand *et al.* (1996). Gene recognition via spliced alignment. *Proc. National Academy of Sciences*, 93(17):9061-9066.
- A. Delcher *et al.* (1999). Alignment of whole genomes. *Nucleic Acids Research*, 27(11):2369-2376.
- X. Huang and K. Chao (2003). A generalized global alignment algorithm. *Bioinformatics*, 19(2):228-233.
- S. Rajko and S. Aluru (2004). Space and time optimal parallel sequence alignments. *IEEE Transactions on Parallel and Distributed Systems*, 15(12):1070-1081.

Books

- D. Gusfield (1997). Algorithms on strings, trees and sequences: Computer Science and Computational Biology. Cambridge University Press, Cambridge, London.
- J. Setubal and J. Meidanis (1997). Introduction to computational molecular biology. PWS Publishing Company, Boston, MA.
- B. Jackson and S. Aluru (2005). Chapter: “Pairwise sequence alignment” in Handbook of computational molecular biology, Ed. S. Aluru, Chapman & Hall/CRC Press.

Selected Bibliography for BLAST Related Topics

Serial BLAST

- S. Altschul *et al.* (1990). Basic Local Alignment Search Tool, *J. Molecular Biology*, 215:403-410.
- W. Gish and D.J. States (1993). Identification of protein coding regions by database similarity search. *Nature Genetics*. 3:266-272.
- T.L. Madden *et al.* (1996). Applications of network BLAST server. *Meth. Enzymol*. 266:131-141.
- S. Altschul, *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389-3402.
- Z. Zhang *et al.* (2000). A greedy algorithm for aligning DNA sequences. *J. Computational Biology*, 7(1-2):203-214.

HPC BLAST

- T. Rognes (2001). ParAlign: A parallel sequence alignment algorithm for rapid and sensitive database searches, Nucleic Acids Research, 29:1647-1652.
- R. Bjornson *et al. (*2002). TurboBLAST®: A parallel implementation of BLAST built on the TurboHub, *Proc. International Parallel and Distributed Processing Symposium.*
- A. Darling, L. Carey and W.C. Feng (2003). The design, implementation, and evaluation of mpiBLAST, *Proc. ClusterWorld.*
- D. Mathog (2003). Parallel BLAST on split databases, *Bioinformatics*, 19:1865-1866.
- J. Wang and Q. Mu (2003). Soap-HT-BLAST: High-throughput BLAST based on web services, *Bioinformatics*, 19:1863-1864.
- H. Lin *et al.* (2005). Efficient data access for parallel BLAST, *Proc. International Parallel and Distributed Processing Symposium.*
- K. Muriki, K. Underwood and R. Sass (2005). RC-BLAST: Towards a portable, cost-effective open source hardware implementation, *Proc. HiCOMB 2005.*
- M. Salisbury (2005). Parallel BLAST: Chopping the database, *Genome Technology*, pp 21-22.

NCBI BLAST - Web Resources

- NCBI BLAST Webpage:
  http://www.ncbi.nlm.nih.gov/BLAST/

- For a comprehensive list of BLAST related references:
  http://www.ncbi.nlm.nih.gov/blast/blast_references.shtml