Heuristic Methods for Sequence Database Searching

BMI/CS 776

www.biostat.wisc.edu/~craven/776.html

Mark Craven

craven@biostat.wisc.edu

February 2002

Announcements

- bioinformatics talk tomorrow:

  Computation in the Imaging of Large Molecules

  Prof. George Phillips

  2/7, 4:00pm in Computer Sciences 1325
- to get on a mailing list of UW bioinformatics events: http://gacrux.biostat.wisc.edu/mailman/listinfo/bioinformatics
- reading for next week:

  Delcher et al., *Alignment of Whole Genomes*

Heuristic Alignment Motivation

- $O(mn)$ too slow for large databases with high query traffic
- heuristic methods do fast approximation to dynamic programming
  - FASTA [Pearson & Lipman, 1988]
  - BLAST [Altschul *et al*., 1990]

Heuristic Alignment Motivation

- consider the task of searching SWISS-PROT against a query sequence:
  - say our query sequence is 362 amino-acids long
  - SWISS-PROT release 38 contains 29,085,265 amino acids
  - finding local alignments via dynamic programming would entail $O(10^{10})$ matrix operations
- many servers handle thousands of such queries a day (NCBI > 50,000)

BLAST Overview

- **B**asic **L**ocal **A**lignment **S**earch **T**ool
- BLAST heuristically finds *high scoring segment pairs* (HSPs):
  - identical length segments from 2 sequences with statistically significant match scores
  - i.e. ungapped local alignments
- key tradeoff: sensitivity vs. speed

$$\text{sensitivity} = \frac{\#\,\text{significant matches detected}}{\#\,\text{significant matches in DB}}$$

BLAST Overview

- given: query sequence *q*, word length *w*, word score threshold *T*, segment score threshold *S*
  - compile a list of "words" that score at least *T* when compared to words from *q*
  - scan database for matches to words in list
  - extend all matches to seek high-scoring segment pairs
- return: segment pairs scoring at least *S*

Determining Query Words

Given:

query sequence: QLNFSAGW

word length w = 2 (typically w = 3 or 4)

word score threshold T = 8

Step 1: determine all words of length w in query sequence

QL LN NF FS SA AG GW

Determining Query Words

Step 2: determine all words that score at least T when compared to a word in the query sequence

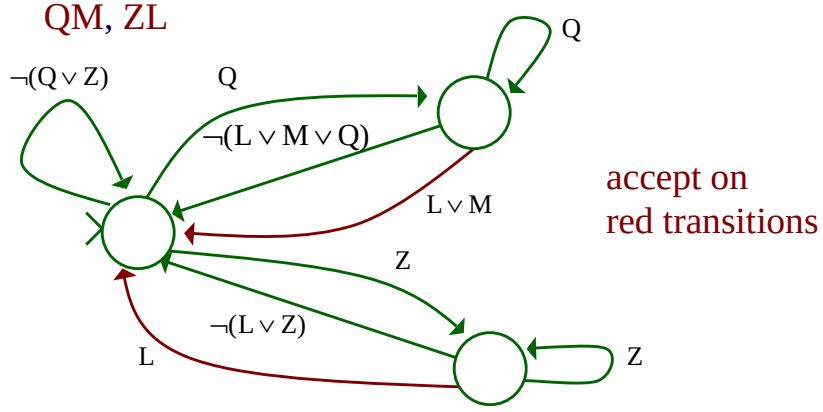
words from sequence	query words w/ T=8
QL	QL=11, QM=9, HL=8, ZL=9
LN	LN=9, LB=8
NF	NF=12, AF=8, NY=8, DF=10,…
…	
SA	none
...	

Scanning the Database

- search database for all occurrences of query words
- approach:
  - build a DFA that recognizes all query words
  - run DB sequences through DFA
  - remember hits

Scanning the Database

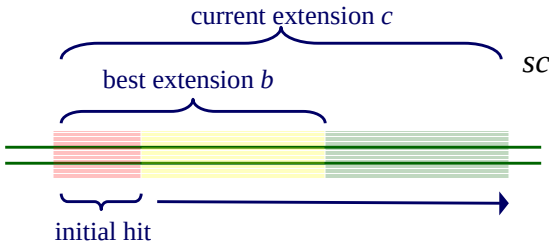
- use Mealy paradigm (accept on transitions) to save space and time
- consider a DFA to recognize the query words: QL, QM, ZL



accept on red transitions

Extending Hits

- extend hits in both directions (without allowing gaps)
- terminate extension in one direction when score falls certain distance below best score for shorter extensions



$score(c) \geq score(b) - \varepsilon$?

- return segment pairs scoring at least S

Sensitivity vs. Running Time

- the main parameter controlling the sensitivity vs. running-time trade-off is T (threshold for what becomes a query word)
  - small T: greater sensitivity, more hits to expand
  - large T: lower sensitivity, fewer hits to expand

BLAST Notes

- may fail to find all HSPs
  - may miss seeds if *T* is too stringent
  - extension is greedy
- empirically, 10 to 50 times faster than Smith-Waterman
- large impact:
  - NCBI's BLAST server handles more than 50,000 queries a day
  - most used bioinformatics program

More Recent BLAST Extensions

- the two-hit method
- gapped BLAST
- PSI-BLAST

* all are aimed at increasing sensitivity while limiting run-time

- Altschul et al., *Nucleic Acids Research* 1997

The Two-Hit Method

- extension step typically accounts for 90% of BLAST's execution time
- key idea: do extension only when there are two hits on the same diagonal within distance A of each other
- to maintain sensitivity, lower *T* parameter
  - more single hits found
  - but only small fraction have associated 2nd hit

The Two-Hit Method

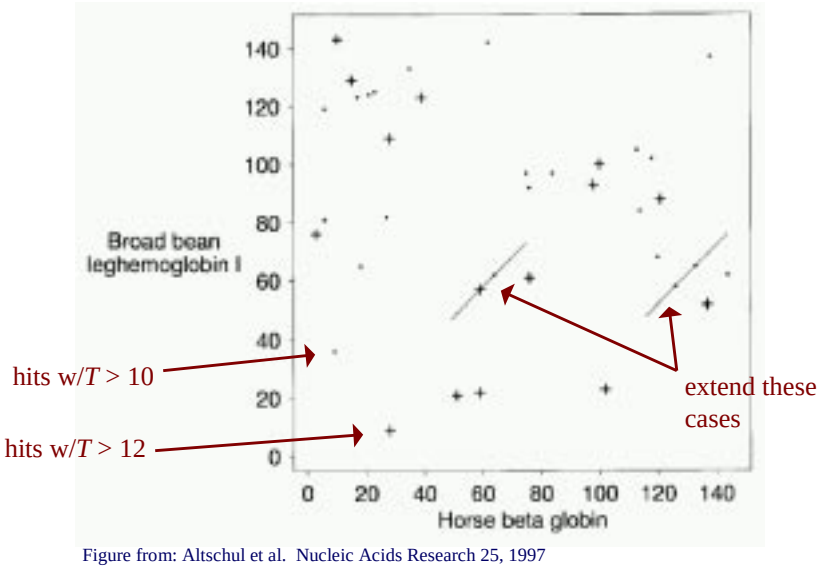


Figure from: Altschul et al. Nucleic Acids Research 25, 1997

Gapped BLAST

- trigger gapped alignment if two-hit extension has a sufficiently high score
- find length-11 segment with highest score; use central pair in this segment as seed
- run DP process both forward & backward from seed
- prune cells when local alignment score falls a certain distance below best score yet

Gapped BLAST

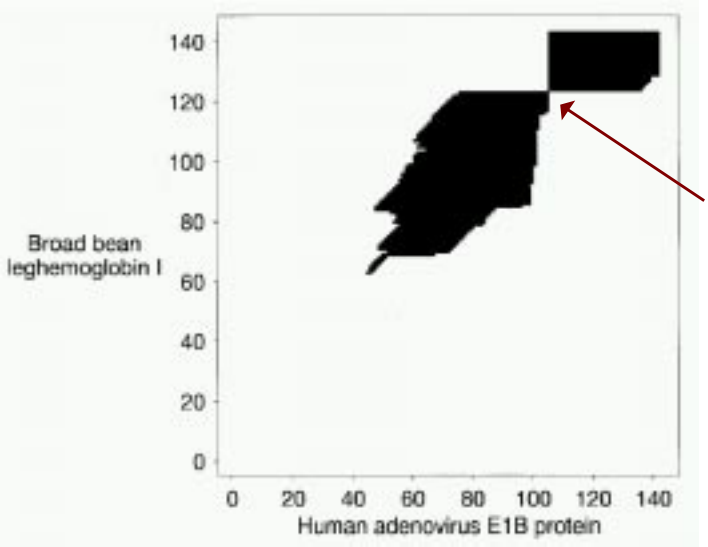
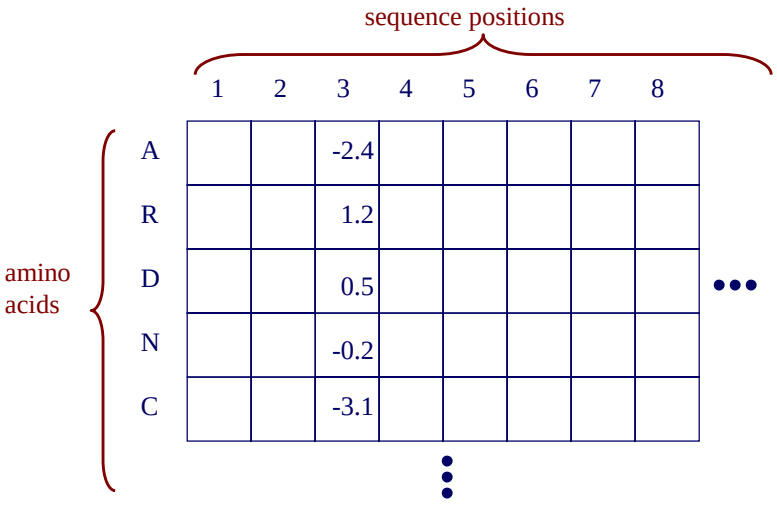


Figure from: Altschul et al. Nucleic Acids Research 25, 1997

PSI (*Position Specific Iterated*) BLAST

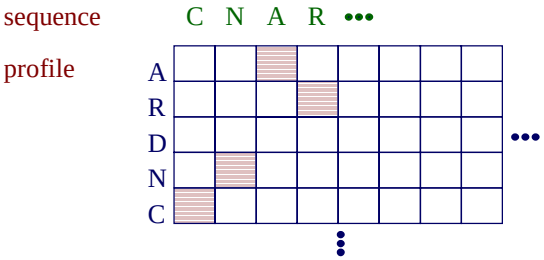
- basic idea
  - use results from BLAST query to construct a *profile matrix*
  - search database with profile instead of query sequence
  - iterate

A Profile Matrix



PSI BLAST: Searching with a Profile

- aligning profile matrix to a simple sequence
  - like aligning two sequences
  - except score for aligning a character with a matrix position is given by the matrix itself – not a substitution matrix



PSI BLAST: Constructing the Profile Matrix

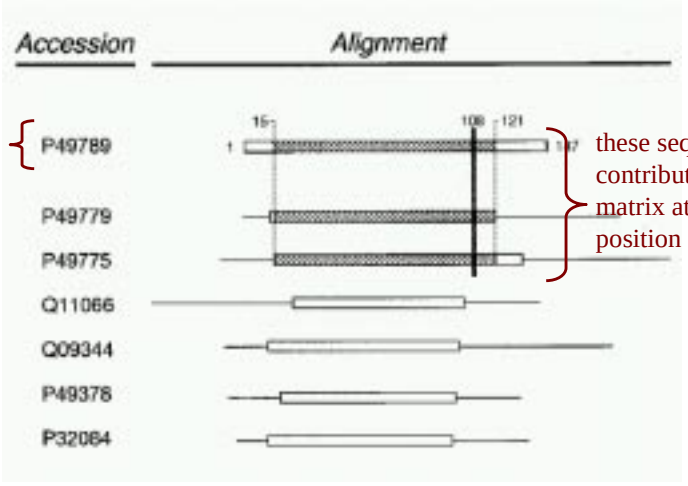


Figure from: Altschul et al. Nucleic Acids Research 25, 1997

PSI BLAST: Determining Profile Elements

- the value for a given element of the profile matrix is given by:

$$matrix(i, j) = \log\left(\frac{\Pr(a_i \mid \text{col} = j)}{\Pr(a_i)}\right)$$

- where the probability of seeing amino acid a_i in column j is estimated as:

observed frequency

pseudocount

$$\Pr(a_i \mid \text{col} = j) = \frac{\alpha\ f_{ij} + \beta\ g_{ij}}{\alpha + \beta}$$