

# Heuristic Methods for Sequence Database Searching

BMI/CS 576

[www.biostat.wisc.edu/bmi576/](http://www.biostat.wisc.edu/bmi576/)

Mark Craven

[craven@biostat.wisc.edu](mailto:craven@biostat.wisc.edu)

Fall 2011

## 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; Altschul *et al.*, *Nucleic Acids Research* 1997]

## Heuristic alignment motivation

- consider the task of searching UnitProtKB/Swiss-Prot against a query sequence:
  - say our query sequence is 362 amino-acids long
  - most recent release of DB contains 188,719,038 amino acids
  - finding local alignments via dynamic programming would entail  $O(10^{11})$  matrix operations
- many servers handle thousands of such queries a day (NCBI > 500,000)

## Heuristic alignment

- heuristic algorithm: a problem-solving method which isn't guaranteed to find the optimal solution, but which is efficient and finds good solutions
- key heuristics in BLAST
  - look for seeds of high scoring alignments
  - use dynamic programming selectively
- key tradeoff made: sensitivity vs. speed
$$\text{sensitivity} = \frac{\# \text{ significant matches detected}}{\# \text{ significant matches in DB}}$$

# Overview of BLAST (Basic Alignment Search Tool)

- given: query sequence  $q$ , word length  $w$ , word score threshold  $T$ , segment score threshold  $S$ 
  - compile a list of “words” (of length  $w$ ) 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 alignments
- return: alignments scoring at least  $S$

## Determining query words

Given:

query sequence: **QLNFSAGW**

word length  $w = 2$  (default for protein usually  $w = 3$ )

word score threshold  $T = 9$

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

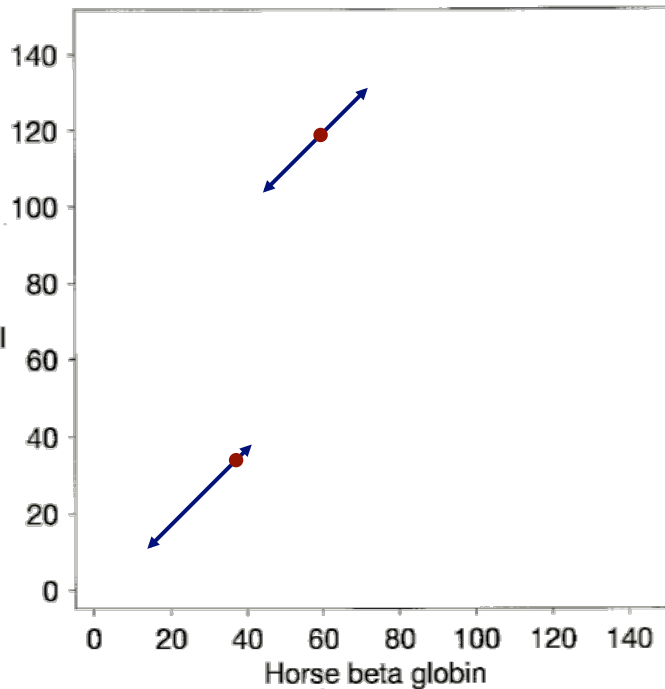
**QL LN NF FS SA AG GW**



## Extending hits

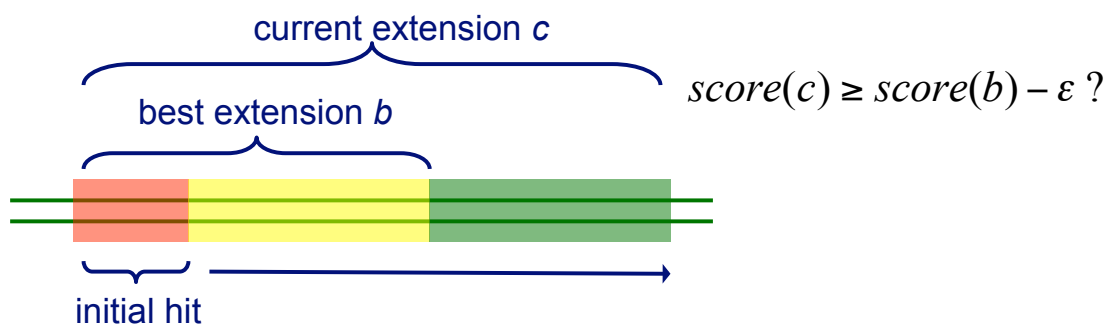
- BLAST extends hits into local alignments
- The original version of BLAST extended each hit separately

Broad bean  
leghemoglobin I



## Extending hits in original BLAST

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



- 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

## 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

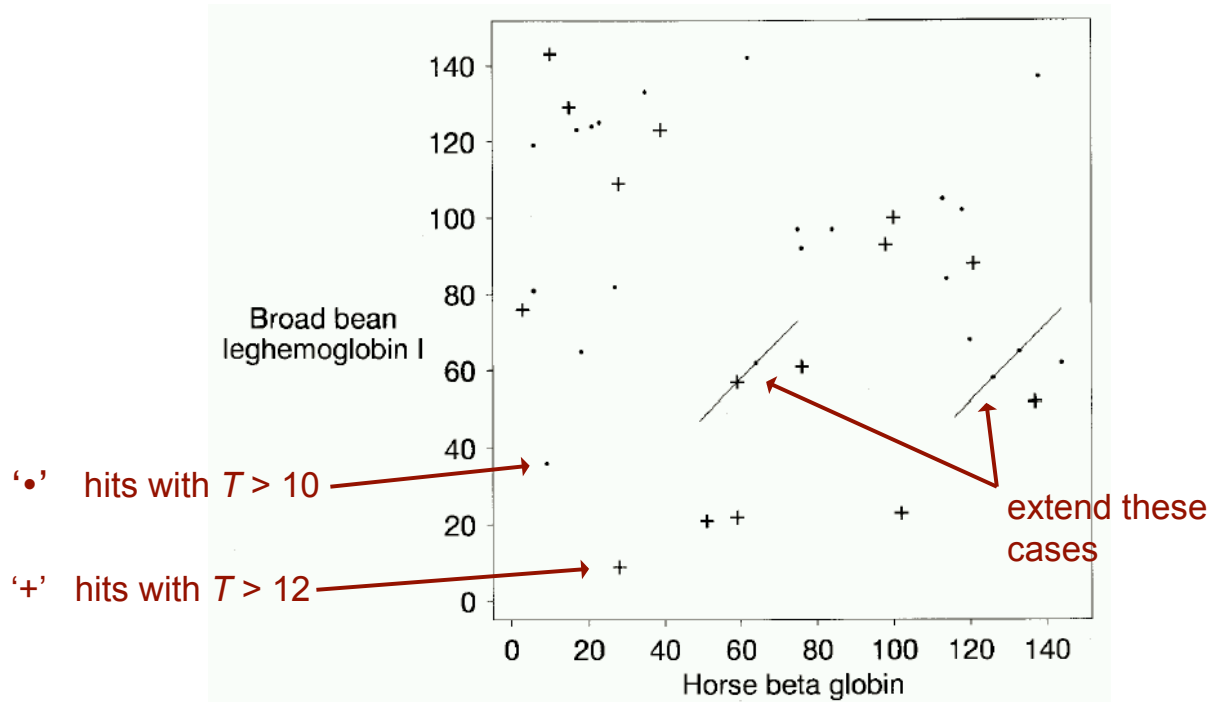


Figure from: Altschul et al. *Nucleic Acids Research* 25, 1997

## 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

filled cells show alignment pairings considered

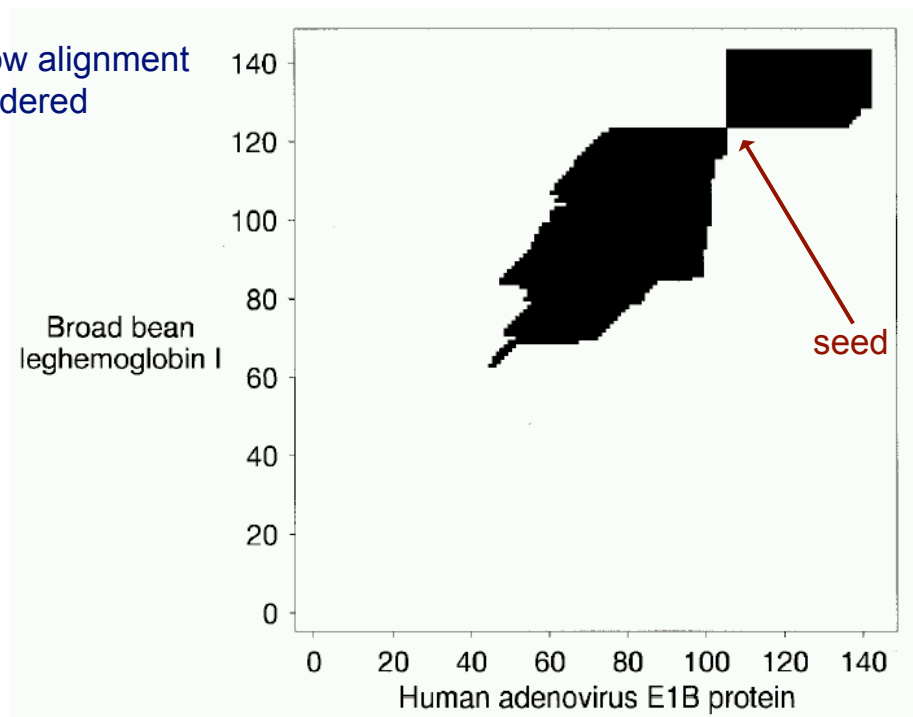


Figure from: Altschul et al. *Nucleic Acids Research* 25, 1997

## 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

sequence positions

	1	2	3	4	5	6	7	8		
amino acids	A			-2.4						
	R			1.2						
	D			0.5						...
	N			-0.2						
	C			-3.1						
					⋮					

## 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

sequence	C	N	A	R	...					
profile	A									
	R									
	D									...
	N									
	C									
					⋮					

# PSI BLAST: constructing the profile matrix

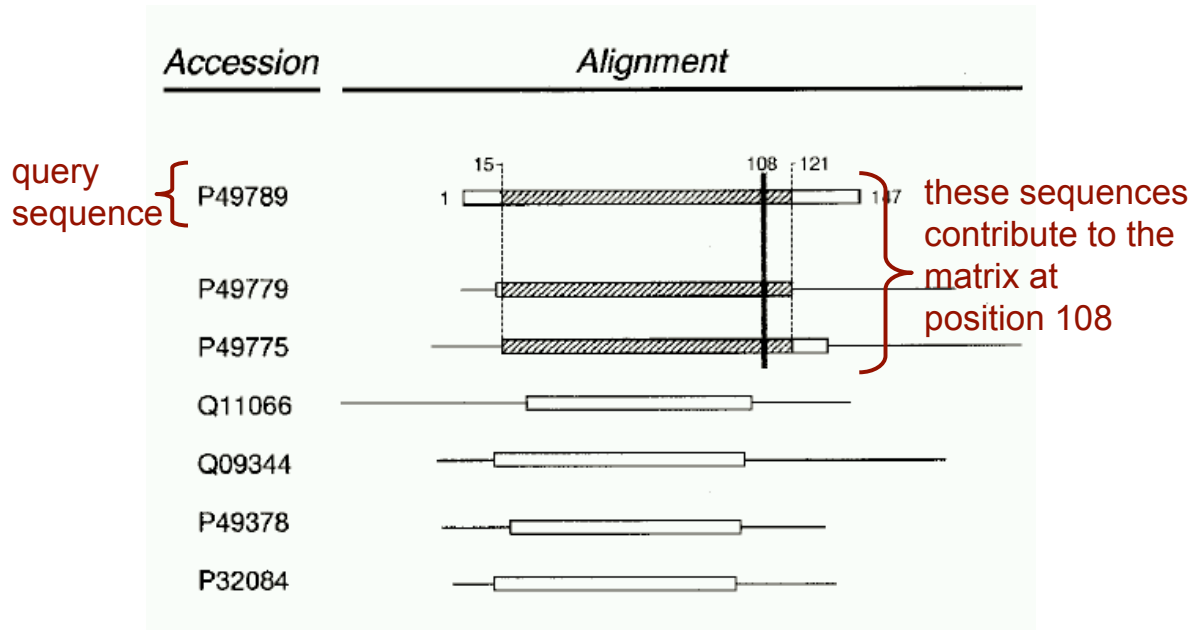


Figure from: Altschul et al. Nucleic Acids Research 25, 1997

The screenshot shows the NCBI BLAST web interface. The "Enter Query Sequence" section has a text input field containing the query: `>mystery mwhltpeeksavtalwqkxnvdevgqealq`. A blue circle highlights this text, with a blue arrow pointing to the word "query" on the left. The "Choose Search Set" section has a dropdown menu for "Database" set to "Non-redundant protein sequences (nr)", which is also circled in blue. A blue arrow points from the word "database" on the left to this dropdown. The "Program Selection" section shows "blastp (protein-protein BLAST)" selected. At the bottom, there is a "BLAST" button and a checkbox for "Show results in a new window".

# BLAST programs

Program	Query	Database
BLASTP	Protein	Protein
BLASTN	DNA	DNA
BLASTX	Translated DNA	Protein
TBLASTN	Protein	Translated DNA
TBLASTX	Translated DNA	Translated DNA

# BLAST results

Sequences producing significant alignments:	Score (Bits)	E Value	
<a href="#">gb AAN84548.1</a> beta globin chain variant [Homo sapiens]	<a href="#">90.6</a>	9e-18	<b>G</b>
<a href="#">gb AAK29639.1 AF349114_1</a> beta globin chain variant [Homo sapiens]	<a href="#">90.6</a>	1e-17	<b>UG</b>
<a href="#">gb AAF00489.1 AF181989_1</a> hemoglobin beta subunit variant [Homo sapiens]	<a href="#">90.6</a>	1e-17	<b>UG</b>
<a href="#">gb AAA35952.1</a> beta-globin	<a href="#">90.6</a>	1e-17	<b>G</b>
<a href="#">gb AAK37051.1</a> hemoglobin beta [synthetic construct]	<a href="#">90.6</a>	1e-17	
<a href="#">gb AAR96398.1</a> hemoglobin beta [Homo sapiens]	<a href="#">90.1</a>	1e-17	<b>UG</b>
<a href="#">gb AAL68978.1 AF083883_1</a> mutant beta-globin [Homo sapiens]	<a href="#">90.1</a>	1e-17	<b>G</b>
<a href="#">gb AAK29557.1</a> hemoglobin beta [synthetic construct]	<a href="#">90.1</a>	1e-17	
<a href="#">ref NP_000509.1</a> beta globin [Homo sapiens] > <a href="#">ref XP_508242.1</a> ...	<a href="#">90.1</a>	1e-17	<b>UG</b>
<a href="#">sp P02024 HBB_GORGO</a> Hemoglobin subunit beta (Hemoglobin beta chain)	<a href="#">90.1</a>	1e-17	
<a href="#">gb AAD19696.1</a> hemoglobin beta chain [Homo sapiens]	<a href="#">90.1</a>	2e-17	<b>UG</b>
<a href="#">emb CAA26204.1</a> beta-globin [Pan troglodytes]	<a href="#">89.7</a>	2e-17	
<a href="#">gb AAN16468.1</a> hemoglobin beta chain variant Hb.Sinai-Bel Air [Homo sapiens]	<a href="#">89.7</a>	2e-17	<b>G</b>
<a href="#">gb ABG47031.1</a> hemoglobin [Homo sapiens]	<a href="#">89.7</a>	2e-17	
<a href="#">gb ABA19233.1</a> hemoglobin beta [Homo sapiens]	<a href="#">89.7</a>	2e-17	<b>G</b>
<a href="#">emb CAA43421.1</a> beta-globin [Gorilla gorilla]	<a href="#">89.3</a>	2e-17	
<a href="#">gb AAY46275.1</a> beta globin chain [Homo sapiens]	<a href="#">89.3</a>	2e-17	<b>G</b>
<a href="#">gb AAK20080.1</a> mutant beta globin [Homo sapiens]	<a href="#">89.3</a>	2e-17	<b>G</b>
<a href="#">gb AAN11321.1</a> hemoglobin beta chain variant Hb-I_Toulouse [Homo sapiens]	<a href="#">89.3</a>	3e-17	<b>G</b>
<a href="#">gb AAG46184.1</a> mutant beta-globin [Homo sapiens] > <a href="#">gb AAG46185.1</a> ...	<a href="#">88.9</a>	3e-17	<b>G</b>
<a href="#">gb ABX52138.1</a> hemoglobin, beta (predicted) [Papio anubis]	<a href="#">88.4</a>	5e-17	
<a href="#">gb AAD30656.1</a> mutant beta-globin [Homo sapiens]	<a href="#">88.0</a>	6e-17	<b>G</b>
<a href="#">pdb 1HBA B</a> Chain B, High-Resolution X-Ray Study Of Deoxyhemoglobin	<a href="#">86.7</a>	1e-16	<b>S</b>

# BLAST comments

- it's heuristic: may miss some good matches
- it's fast: empirically, 10 to 50 times faster than Smith-Waterman
- PSI-BLAST can detect more distant relationships among protein sequences, but the process of generalizing the query can also lead it astray
- large impact:
  - NCBI's BLAST server handles more than 500,000 queries a day
  - most used bioinformatics program in the world