Heuristic Methods for Sequence Database Searching

BMI/CS 576

www.biostat.wisc.edu/bmi576/

Mark Craven

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 UniProtKB/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

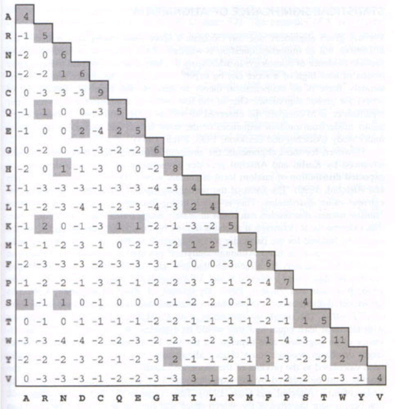
Step 1: determine all words of length *w* in query sequence

QL LN NF FS SA AG GW

Determining query words

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

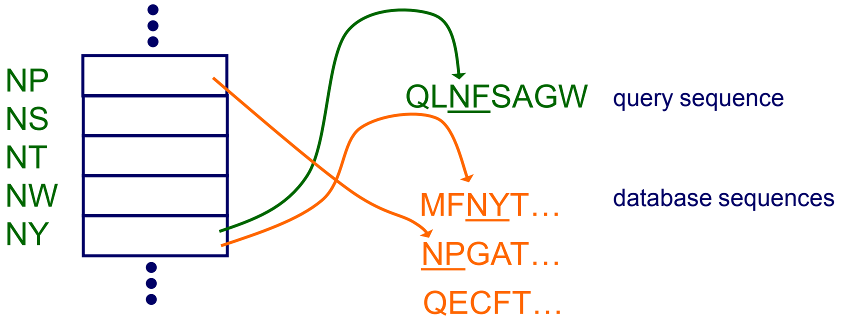
words from sequence	query words w/ T≥9
QL	QL=9
LN	LN=10
NF	NF=12, NY=9
…	
SA	none
...	



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

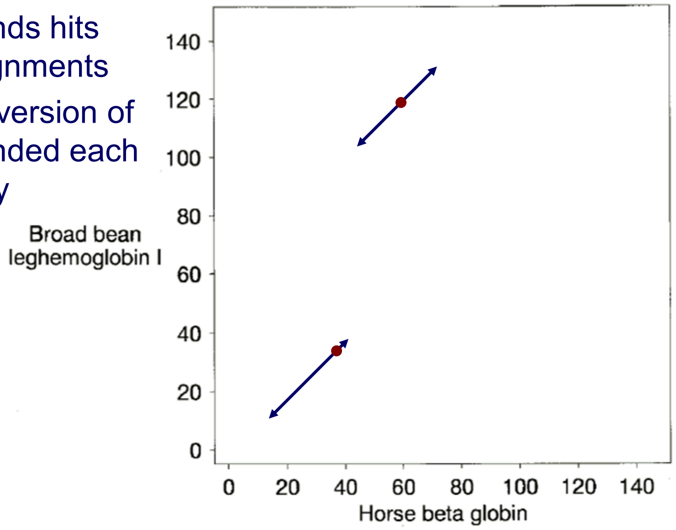
Scanning the database

- search database for all occurrences of query words
- approach:
  - index database sequences into table of words (pre-compute this)
  - index query words into table (at query time)



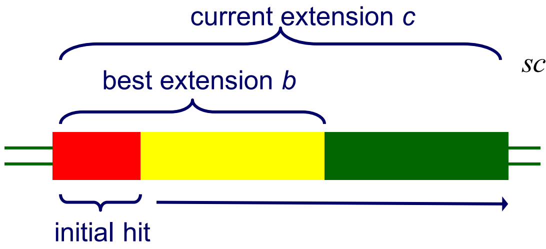
Extending hits

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



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

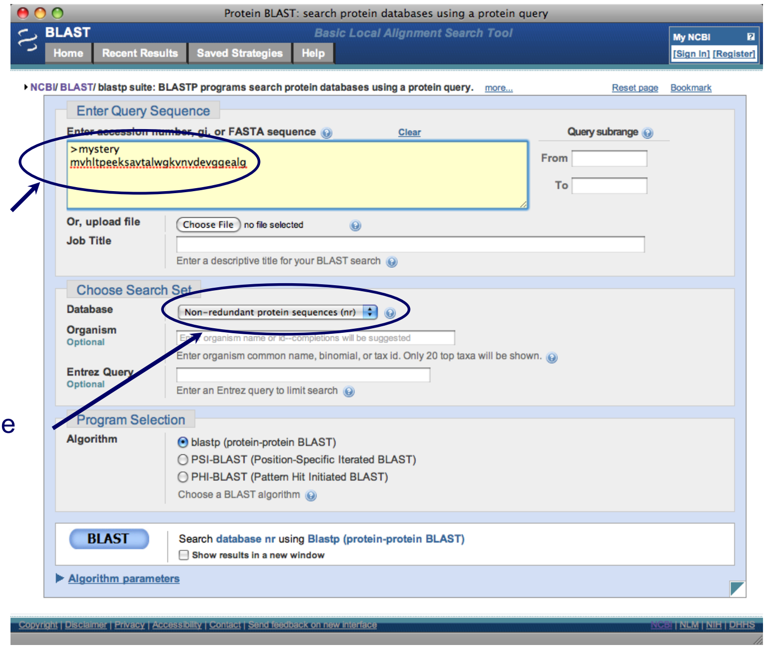


$$score\ (c) \geq score\ (b) - \varepsilon\ ?$$

- 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



Protein BLAST: search protein databases using a protein query

NCBI/ BLAST/ blastp suite: BLASTP programs search protein databases using a protein query.

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

```
>mystery
mvhltpeeksavtalwgkvnvdevggealg
```

Or, upload file: no file selected

Job Title: Enter a descriptive title for your BLAST search

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism (Optional): Enter organism name or id--completions will be suggested

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query (Optional): Enter an Entrez query to limit search

Program Selection

Algorithm

- ● blastp (protein-protein BLAST)
- ○ PSI-BLAST (Position-Specific Iterated BLAST)
- ○ PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

BLAST — Search database nr using Blastp (protein-protein BLAST)

□ Show results in a new window

query

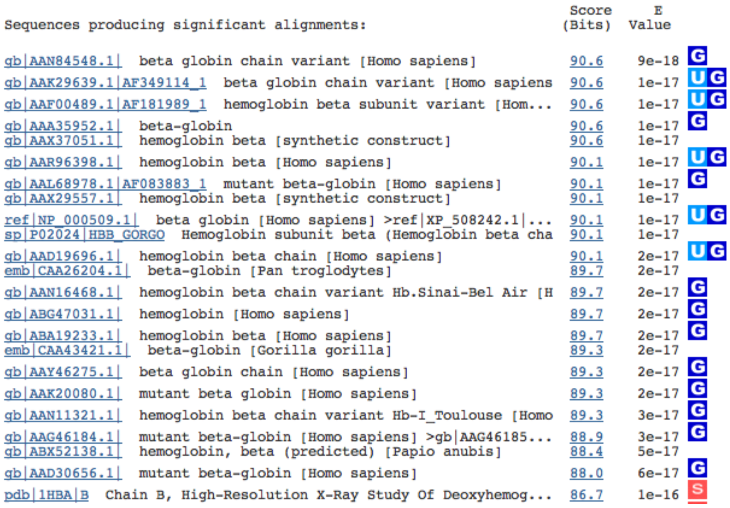
database

www.ncbi.nlm.nih.gov/BLAST/

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



```
                                                                 Score    E
Sequences producing significant alignments:                      (Bits)  Value

gb|AAN84548.1|  beta globin chain variant [Homo sapiens]          90.6    9e-18  G
gb|AAK29639.1|AF349114_1  beta globin chain variant [Homo sapiens 90.6    1e-17  UG
gb|AAF00489.1|AF181989_1  hemoglobin beta subunit variant [Hom... 90.6    1e-17  UG
gb|AAA35952.1|  beta-globin                                       90.6    1e-17  G
gb|AAX37051.1|  hemoglobin beta [synthetic construct]             90.6    1e-17
gb|AAR96398.1|  hemoglobin beta [Homo sapiens]                    90.1    1e-17  UG
gb|AAL68978.1|AF083883_1  mutant beta-globin [Homo sapiens]       90.1    1e-17  G
gb|AAX29557.1|  hemoglobin beta [synthetic construct]             90.1    1e-17
ref|NP_000509.1|  beta globin [Homo sapiens] >ref|XP_508242.1|... 90.1    1e-17  UG
sp|P02024|HBB_GORGO  Hemoglobin subunit beta (Hemoglobin beta cha 90.1    1e-17
gb|AAD19696.1|  hemoglobin beta chain [Homo sapiens]              90.1    2e-17  UG
emb|CAA26204.1|  beta-globin [Pan troglodytes]                    89.7    2e-17
gb|AAN16468.1|  hemoglobin beta chain variant Hb.Sinai-Bel Air [H 89.7    2e-17  G
gb|ABG47031.1|  hemoglobin [Homo sapiens]                         89.7    2e-17  G
gb|ABA19233.1|  hemoglobin beta [Homo sapiens]                    89.7    2e-17  G
emb|CAA43421.1|  beta-globin [Gorilla gorilla]                    89.3    2e-17
gb|AAY46275.1|  beta globin chain [Homo sapiens]                  89.3    2e-17  G
gb|AAK20080.1|  mutant beta globin [Homo sapiens]                 89.3    2e-17  G
gb|AAN11321.1|  hemoglobin beta chain variant Hb-I_Toulouse [Homo 89.3    3e-17  G
gb|AAG46184.1|  mutant beta-globin [Homo sapiens] >gb|AAG46185... 88.9    3e-17  G
gb|ABX52138.1|  hemoglobin, beta (predicted) [Papio anubis]       88.4    5e-17
gb|AAD30656.1|  mutant beta-globin [Homo sapiens]                 88.0    6e-17  G
pdb|1HBA|B  Chain B, High-Resolution X-Ray Study Of Deoxyhemog... 86.7    1e-16  S
```

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