Heuristic Methods for Sequence Database Searching

Jiří Kléma

Department of Computer Science,
Czech Technical University in Prague

Lecture based on Mark Craven's class at University of Wisconsin

http://cw.felk.cvut.cz/wiki/courses/b4m36bin/start

Heuristic alignment motivation

- $\mathcal{O}(mn)$ too slow for large databases with high query traffic,
- heuristic methods do fast approximation to dynamic programming
  - FASTA [Pearson and Lipman, 1988],
  - BLAST [Altschul et al., 1990; Altschul et al., Nucleic Acids Research 1997],
- consider the task of searching UniProtKB/Swiss-Prot against a query sequence
  - say our query sequence is 362 amino-acids long,
  - the release 2021 of DB contains 203,340,877 amino acids (grows approx by 1% annually),
  - finding local alignments via dynamic programming would entail $\mathcal{O}(10^{11})$ matrix operations,
- many servers handle thousands of such queries a day (NCBI $>$ 500,000).

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,
- key heuristics in BLAST
  - look for seeds of high scoring alignments,
  - use dynamic programming selectively,
- key tradeoff made: sensitivity vs. speed

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

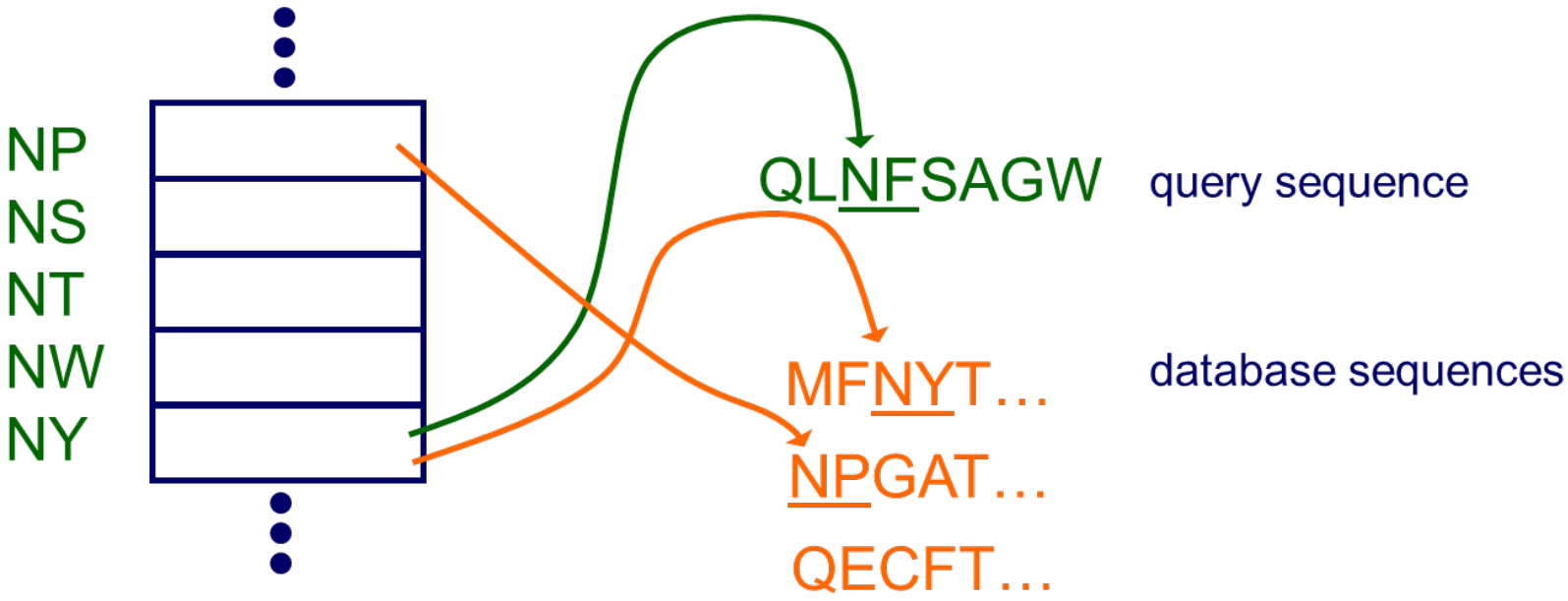
Determining query words

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

words from sequence	query words with $T \geq 9$	substitution scores (BLOSUM62)
QL	**QL**=9	s(Q,Q)=5, s(L,L)=4
LN	**LN**=10	s(L,L)=4, s(N,N)=6
NF	**NF**=12, **NY**=9	s(N,N)=6, s(F,F)=6, s(F,Y)=3
...	...	...
SA	none	

Scanning the database

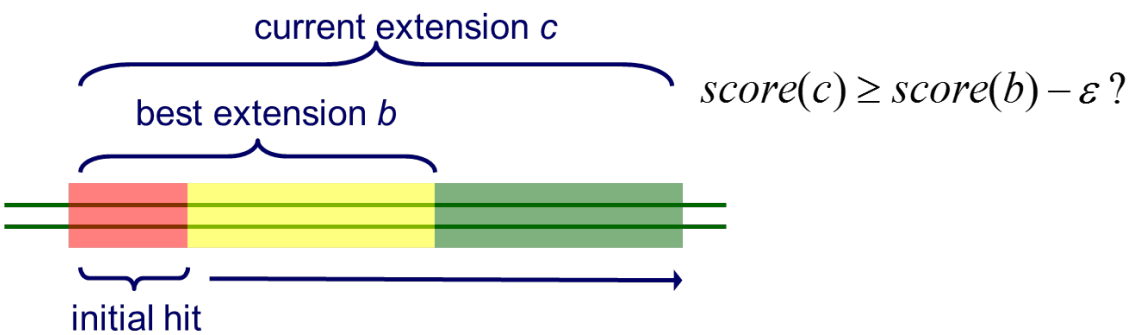
- Search database for all occurrences of query words
  - index database sequences into table of words (pre-compute this),
  - index query words into table (at query time).



Marc Craven, BMI/CS 576, www.biostat.wisc.edu/bmi576.

Extending hits

- BLAST extends hits into local alignments,
- the original version of BLAST extended each hit separately
- 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.

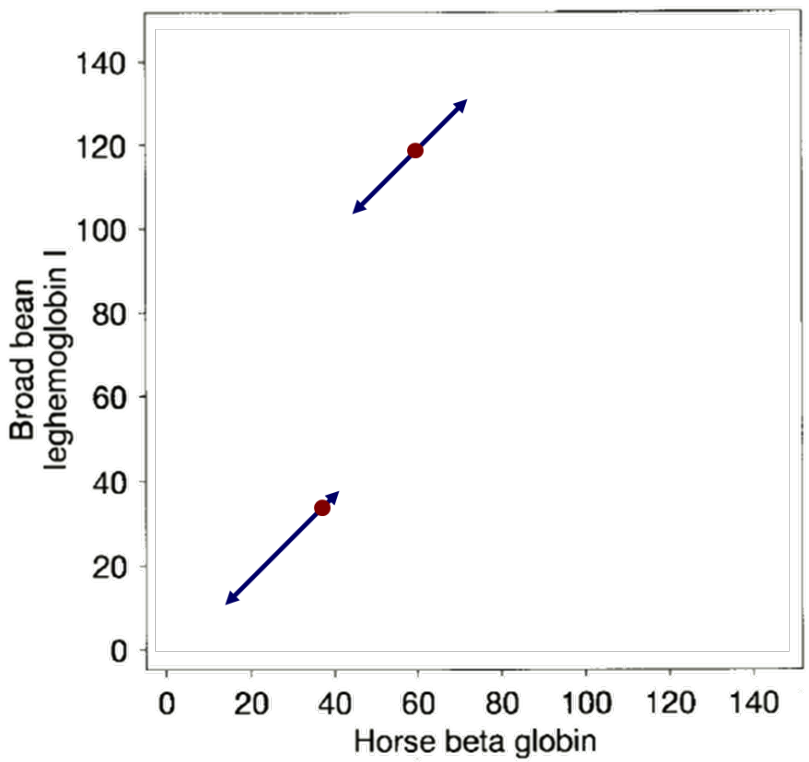


Marc Craven, BMI/CS 576, www.biostat.wisc.edu/bmi576.

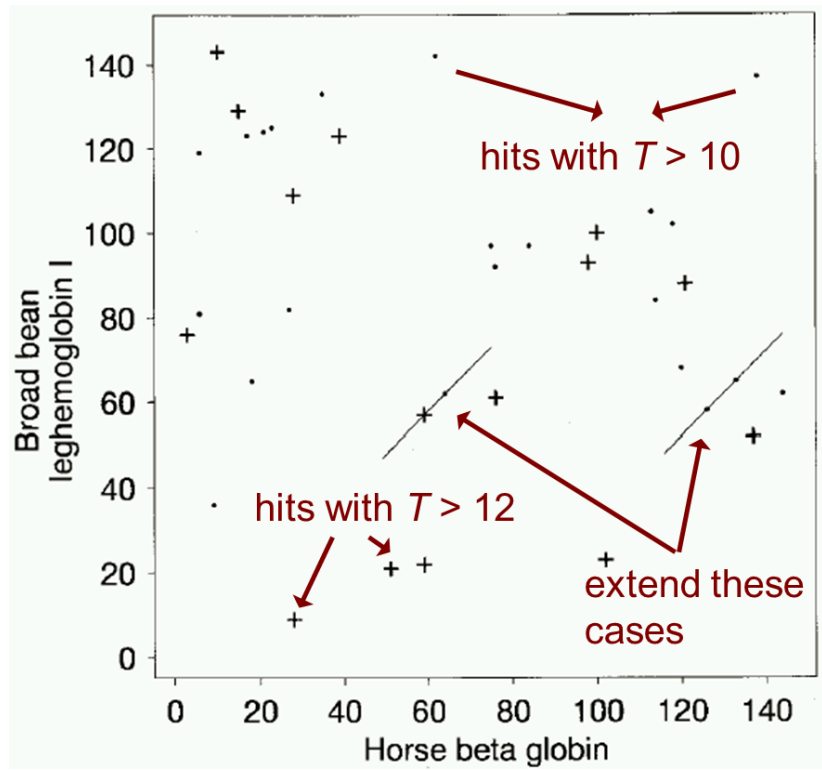
The two-hit method

- 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,
- extension step typically accounts for 90% of BLAST's execution time,
- the two-hit method
  - 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.

Extending hits in original and two-hit BLAST



original BLAST

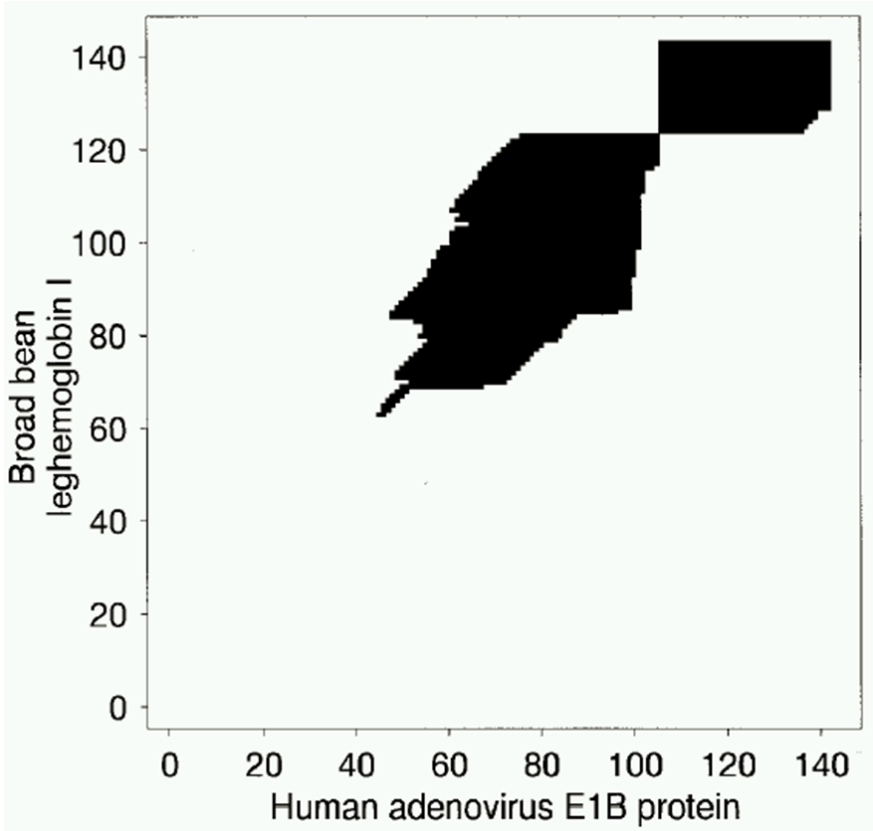


two-hit method

Marc Craven, BMI/CS 576, www.biostat.wisc.edu/bmi576.

Gapped BLAST

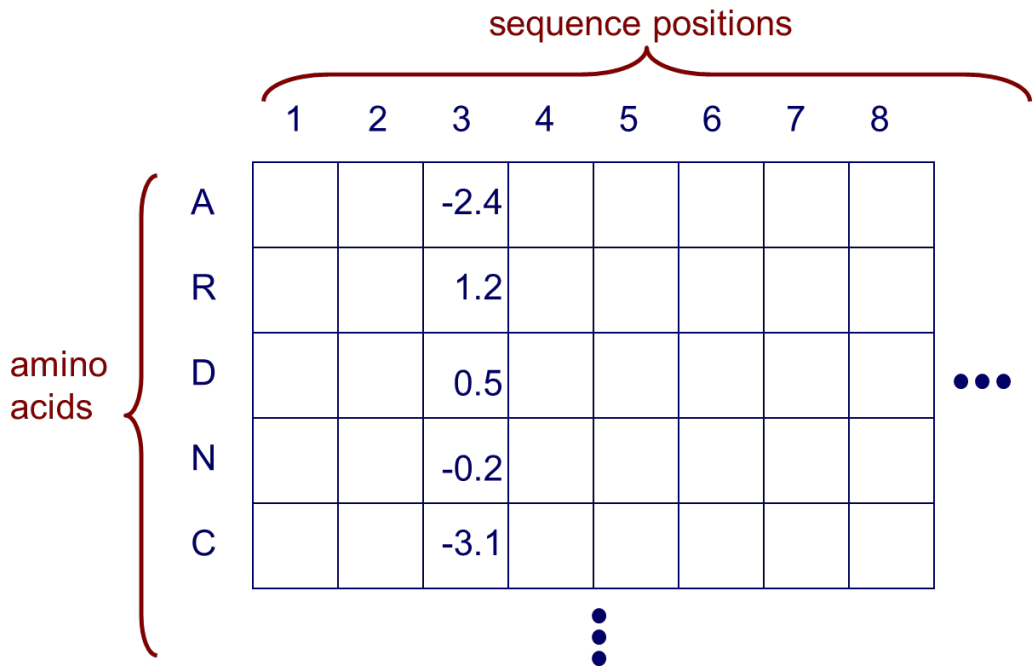
- trigger gapped alignment if two-hit extension has a sufficiently high score,
- find length-11 segment with highest score; use central pair in this segment as seed,
- run DP process both forward & backward from seed,
- prune cells when local alignment score falls a certain distance below best score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



Altschul et al. Nucleic Acids Research 25, 1997

PSI (Position Specific Iterated) BLAST

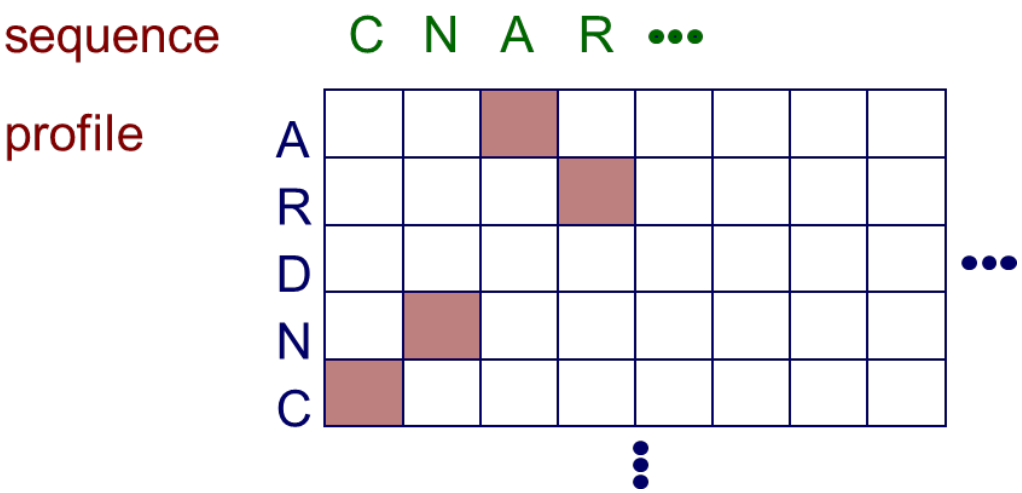
- basic idea
  - use results from BLAST query to construct **a profile matrix**,
  - search database with profile instead of query sequence,
  - iterate.



Marc Craven, BMI/CS 576, www.biostat.wisc.edu/bmi576.

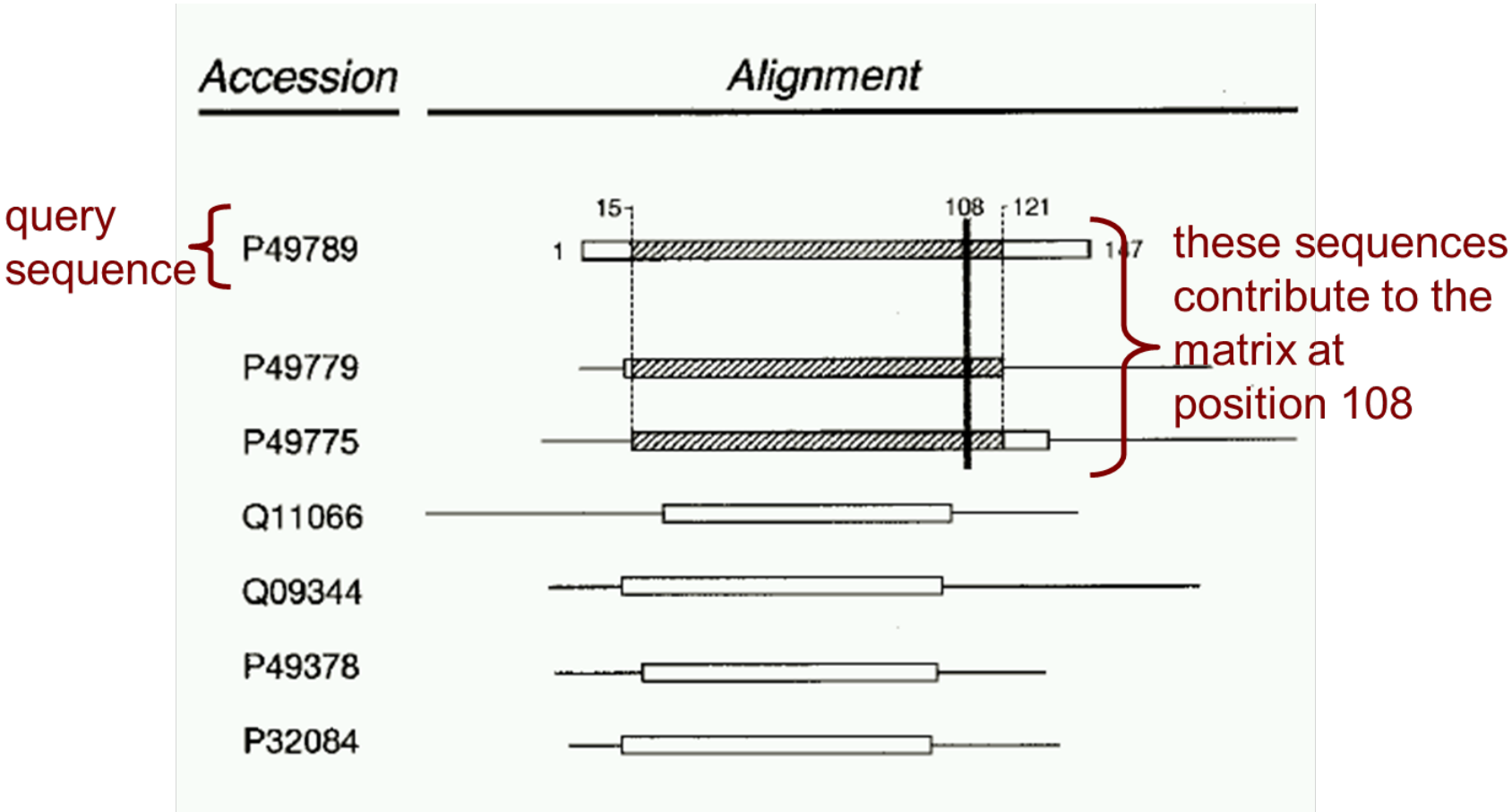
PSI BLAST – searching with a profile

- 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.



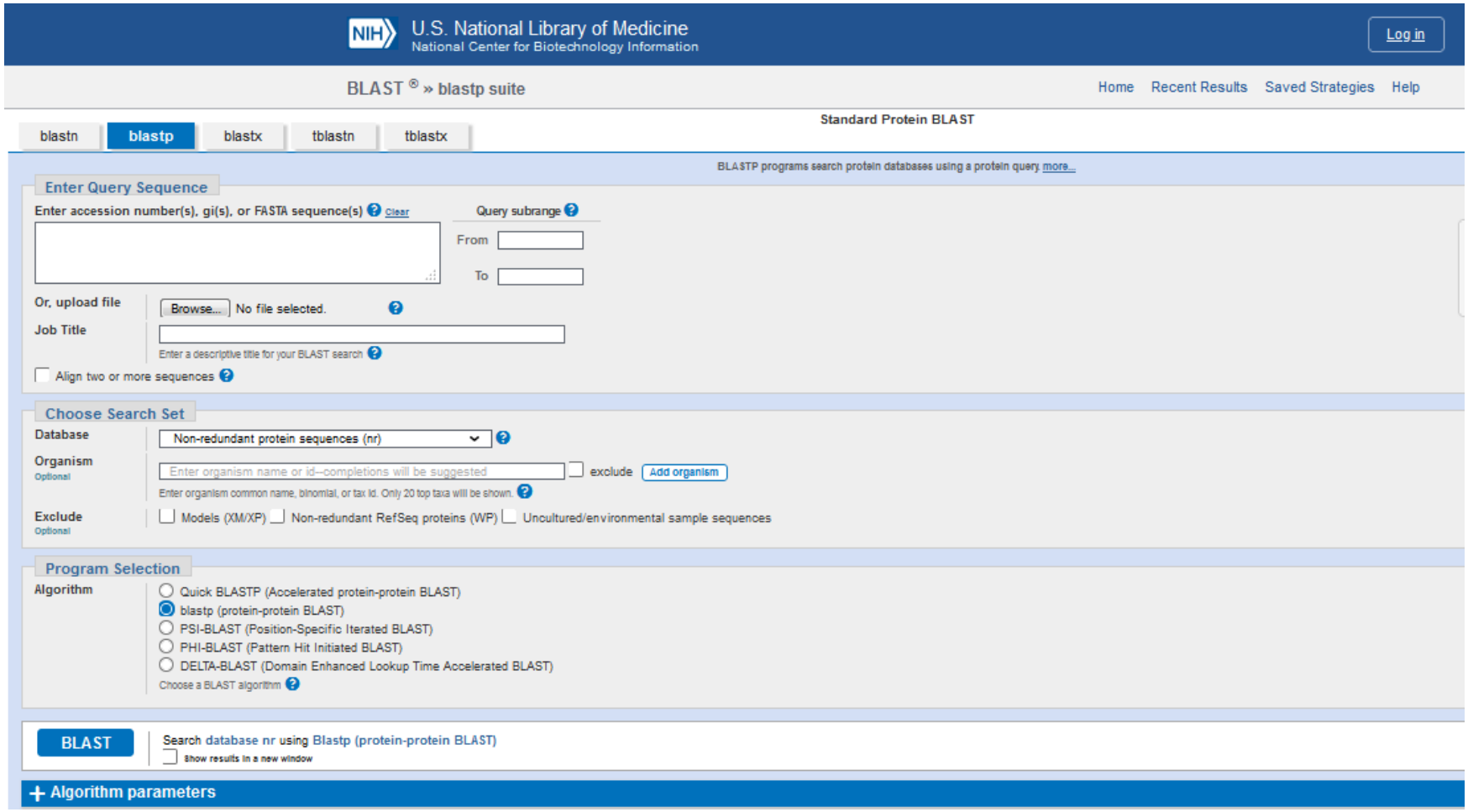
Marc Craven, BMI/CS 576, www.biostat.wisc.edu/bmi576.

PSI BLAST: constructing the profile matrix



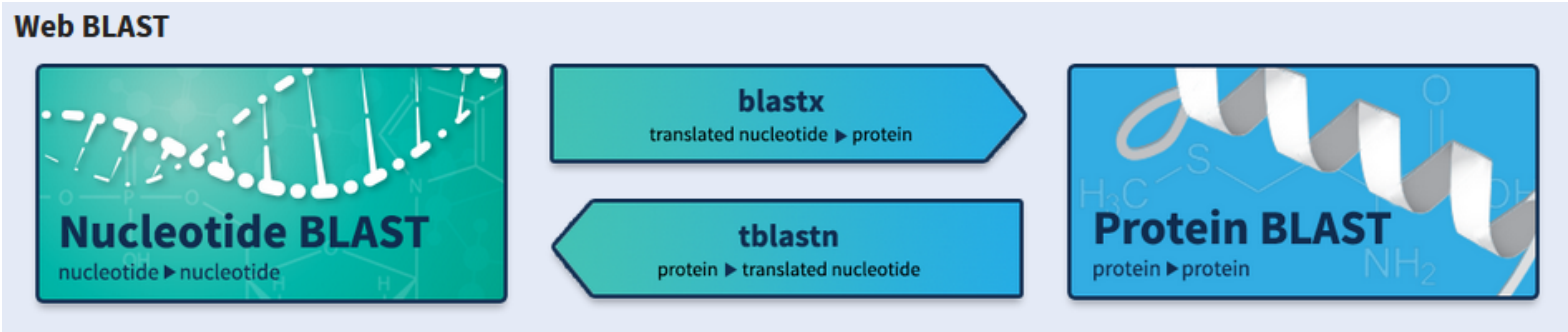
Altschul et al. Nucleic Acids Research 25, 1997.

NCBI BLAST



https://blast.ncbi.nlm.nih.gov/Blast.cgi

BLAST programs



Sequences producing significant alignments Download New Select columns Show 100

select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignment New MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
beta-globin [Homo sapiens]	Homo sapiens	58.3	58.3	90%	3e-09	100.00%	30	AAP74754.1
beta globin [Homo sapiens]	Homo sapiens	58.3	58.3	90%	3e-09	100.00%	30	AAC97959.1
beta globin variant [Homo sapiens]	Homo sapiens	58.3	58.3	90%	4e-09	100.00%	31	AAP44006.1
mutant beta-globin [Homo sapiens]	Homo sapiens	58.3	58.3	90%	4e-09	100.00%	31	AAG46182.1
hemoglobin beta [Homo sapiens]	Homo sapiens	58.3	58.3	90%	4e-09	100.00%	33	AFR11469.1
beta-globin thalassemia [Homo sapiens]	Homo sapiens	58.3	58.3	90%	6e-09	100.00%	36	AAA16335.1
beta-globin [Homo sapiens]	Homo sapiens	58.3	58.3	90%	6e-09	100.00%	37	AAA88069.1
truncated beta globin [Homo sapiens]	Homo sapiens	58.3	58.3	90%	7e-09	100.00%	39	ACF16769.1
beta globin [Homo sapiens]	Homo sapiens	58.3	58.3	90%	8e-09	100.00%	41	ACZ67952.1
beta globin [Homo sapiens]	Homo sapiens	58.3	58.3	90%	9e-09	100.00%	42	AAB60348.1
beta-globin [Homo sapiens]	Homo sapiens	58.3	58.3	90%	2e-08	100.00%	55	AWD38994.1
beta globin [Homo sapiens]	Homo sapiens	58.3	58.3	90%	2e-08	100.00%	57	AAC97372.1

https://blast.ncbi.nlm.nih.gov/Blast.cgi

BLAST summary

- it is heuristic: may miss some good matches,
- it is fast: empirically, 10 to 50 times faster than dynamic programming (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
  - most used bioinformatics program in the world,
- some recent changes/extensions
  - SmartBLAST – an original query immediatelly followed by MSA and phylogenetic analysis.