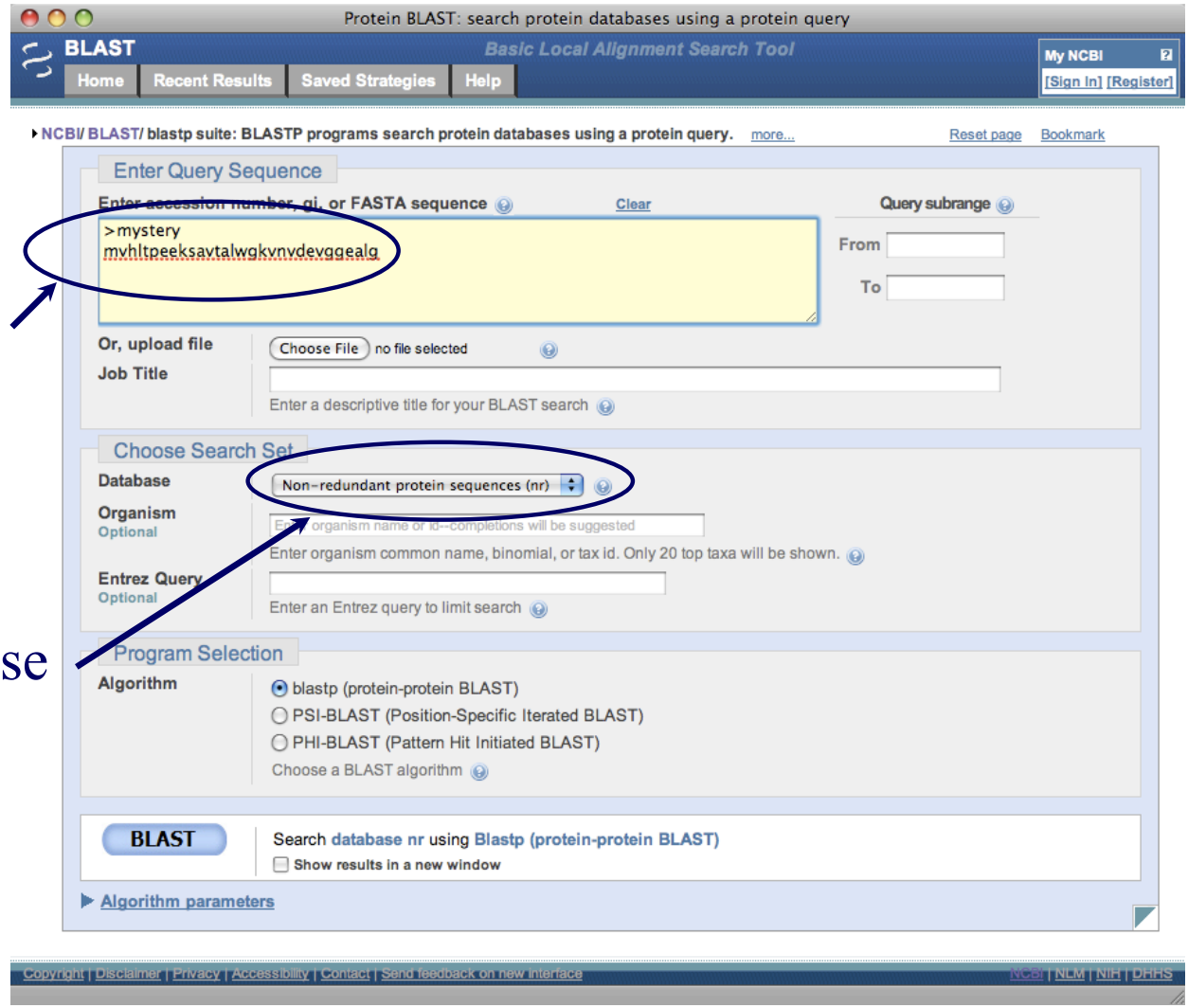
Sequence alignment

Heuristic Methods for Sequence Database Searching

Outline

- The sequence database search task
- Motivation for heuristic alignment algorithms
- The BLAST algorithm
- Variants of the BLAST algorithm



query

database

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

BLAST Results

Sequences producing significant alignments:	Score (Bits)	E 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	U G
gb|AAF00489.1|AF181989_1 hemoglobin beta subunit variant [Hom...	90.6	1e-17	U G
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	U G
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	U G
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	U G
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

Heuristic Alignment Motivation

- $O(mn)$ too slow for large databases with high query traffic
- **Heuristic algorithm**: an algorithm that isn't guaranteed to find the optimal solution, but that is efficient and finds good solutions in practice
- 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 SWISS-PROT against a query sequence:
  - say our query sequence is 362 amino-acids long
  - SWISS-PROT release 38 contained 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 > 100,000)

BLAST Overview

- **B**asic **L**ocal **A**lignment **S**earch **T**ool
- BLAST heuristically finds high scoring local alignments
- typically used to search a query sequence against a database of sequences
- key tradeoff made: sensitivity vs. speed

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

Overview of BLAST Algorithm

- 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/hits to words in list
  - extend all matches/hits 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 $T = 9$

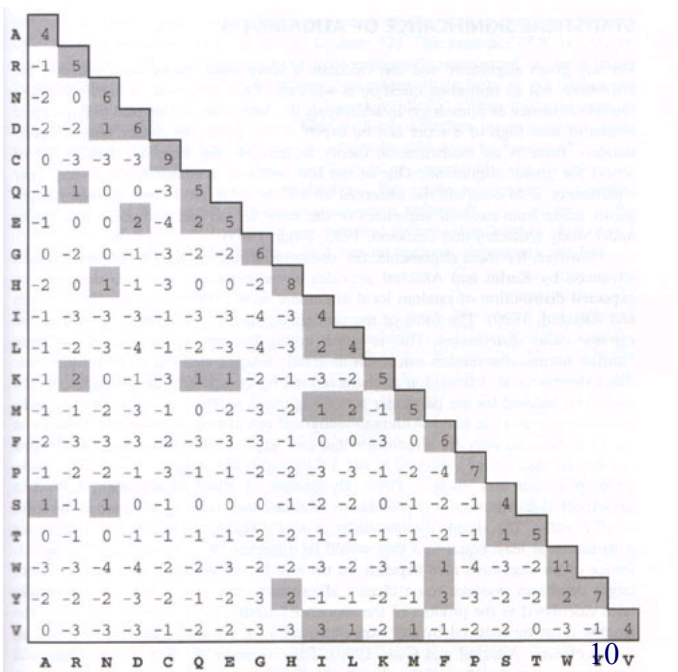
Step 1: determine all words of length w in query sequence (w-mers)

QL LN NF FS SA AG GW

Determining Query Words

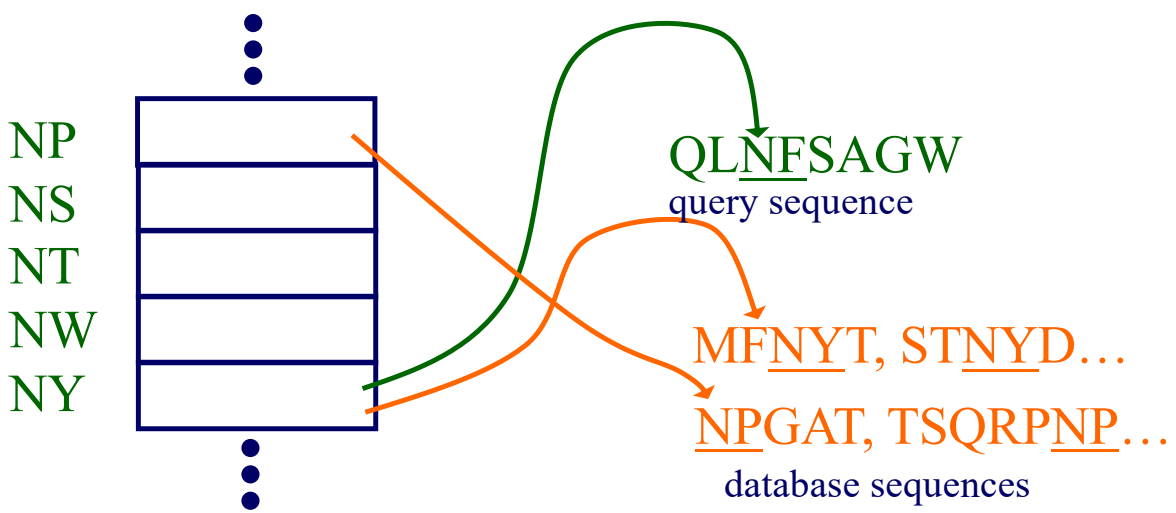
Step 2: determine all words that score at least *T* when compared to a word in the query sequence

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



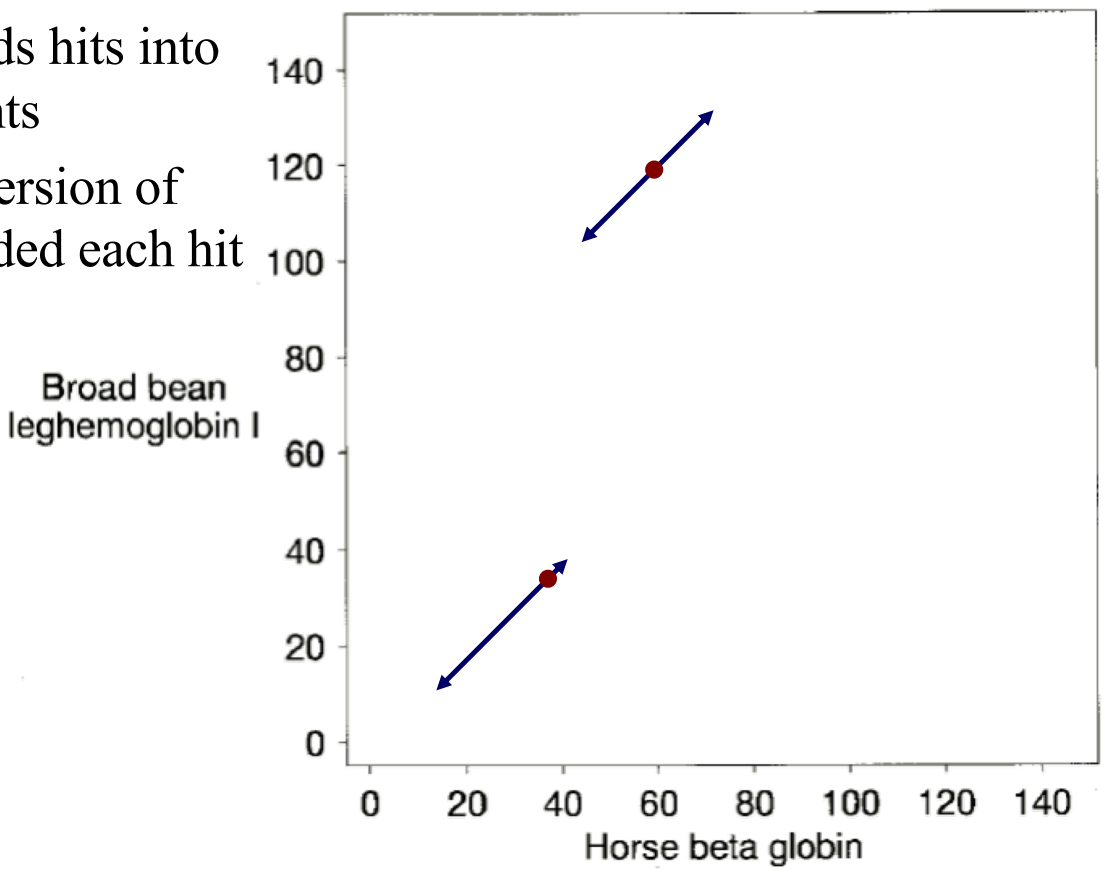
Scanning the database

- 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)



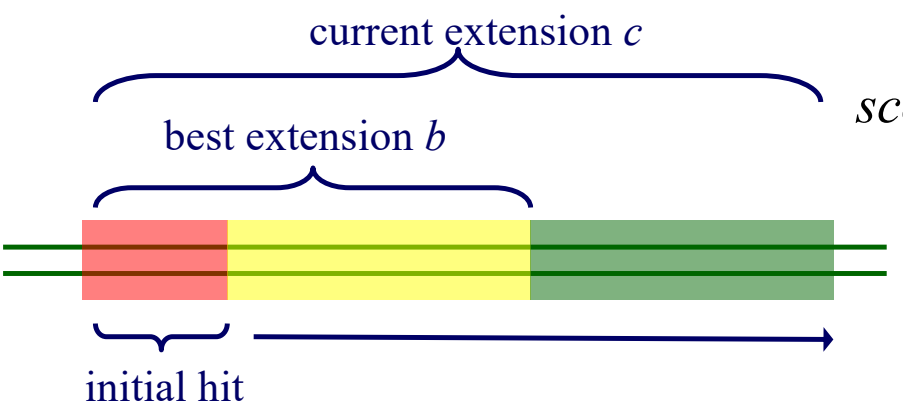
Extending Hits

- 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

How to choose w and T?

- Tradeoff between running time and sensitivity
- Sensitivity

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

- T
  - small *T*: greater sensitivity, more hits to expand
  - large *T*: lower sensitivity, fewer hits to expand
- *w*
  - Larger *w*: 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

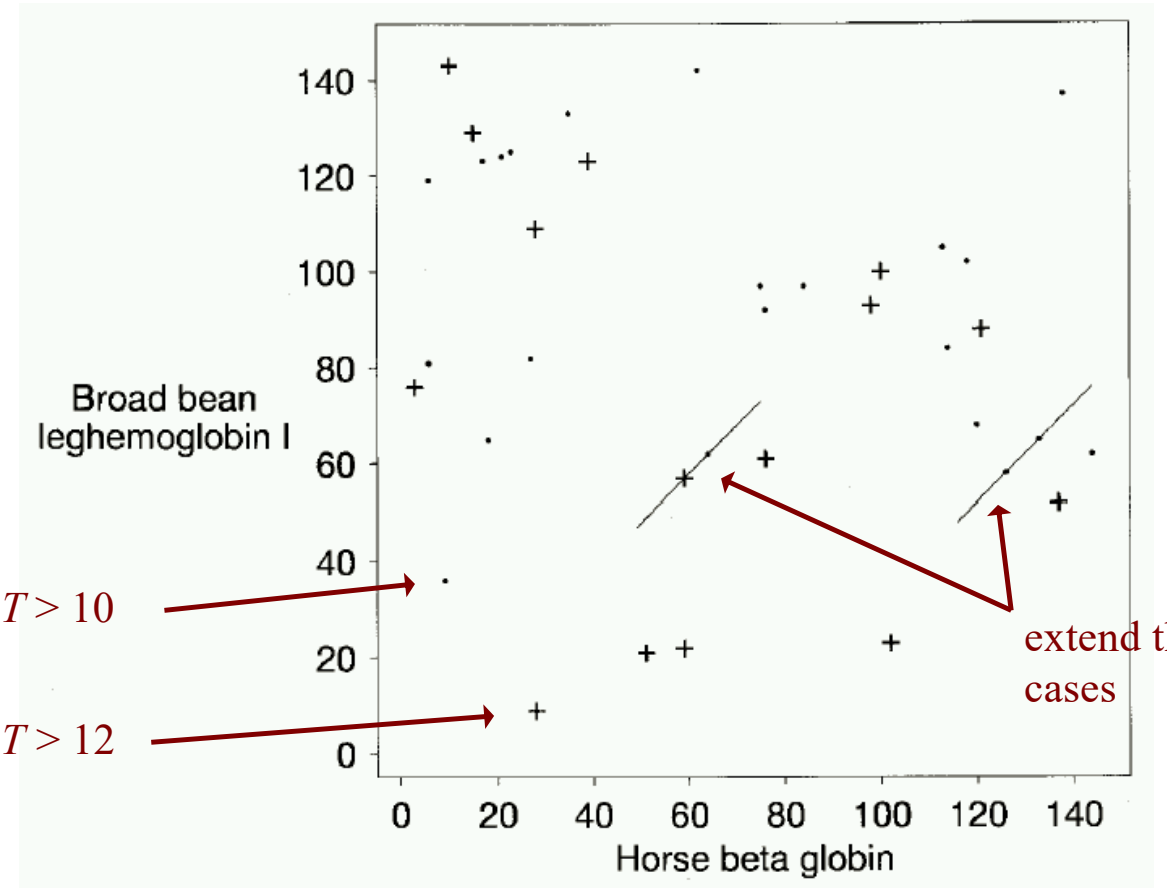


Figure from: Altschul et al. Nucleic Acids Research 25, 1997

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

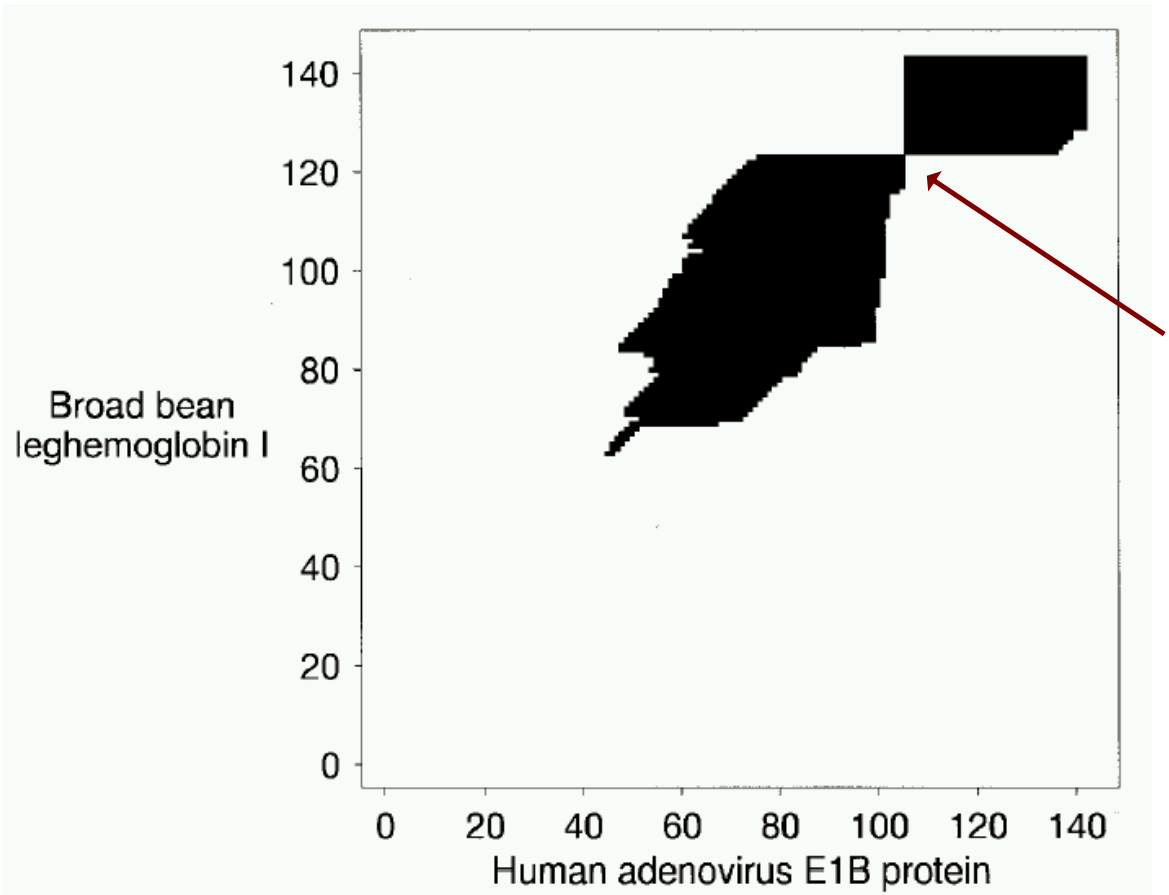


Figure from: Altschul et al. Nucleic Acids Research 25, 1997

BLAST Programs

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

Summary

- It's heuristic: may miss some good matches
- It's fast: empirically, 10 to 50 times faster than Smith-Waterman
- large impact:
  - NCBI's BLAST server handles more than 100,000 queries a day
  - most used bioinformatics program in the world