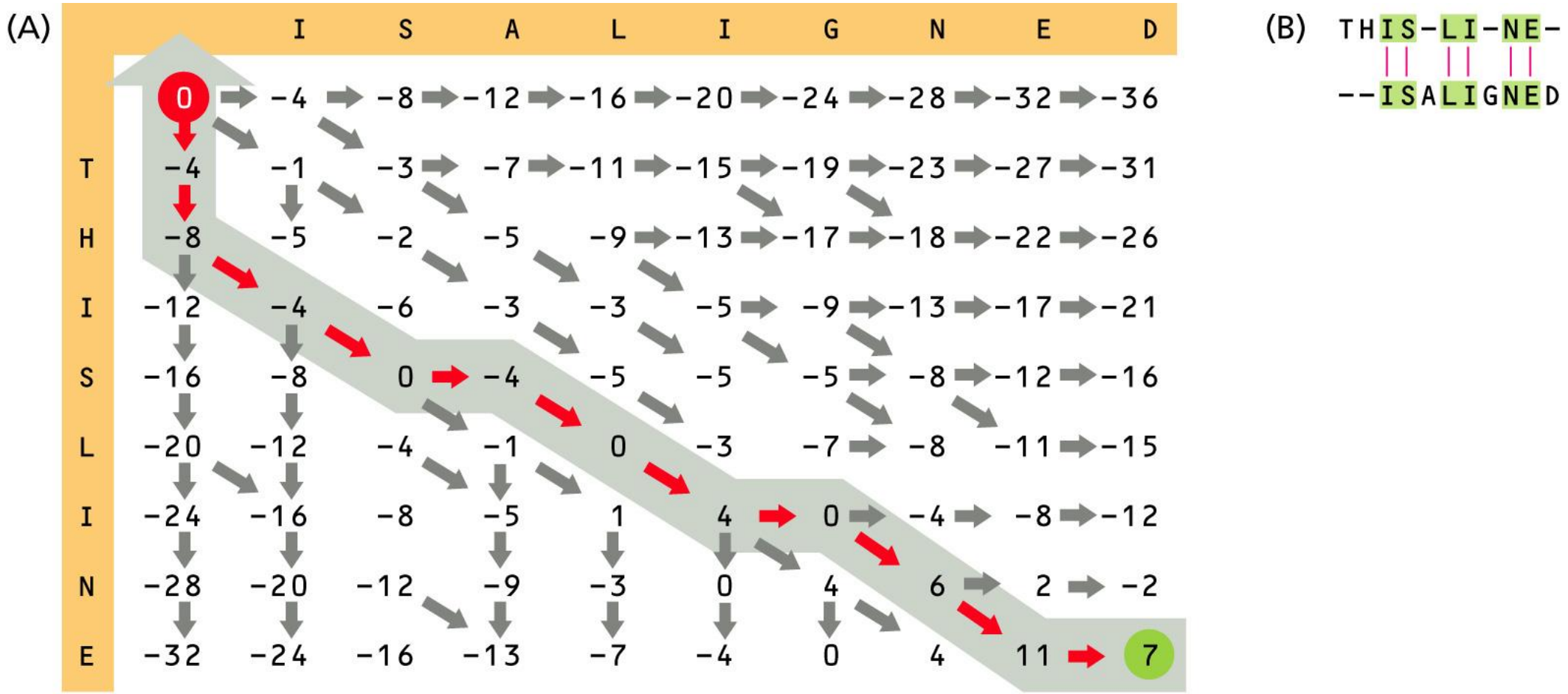
# Sequence Similarity Search: BLAST and related tools

by Ahmet Sacan

Images are from "Understanding Bioinformatics" and Wikipedia.

# Dynamic Programming Time Complexity: O(mn)



# Sequence Similarity Search: Multiple Local Alignment

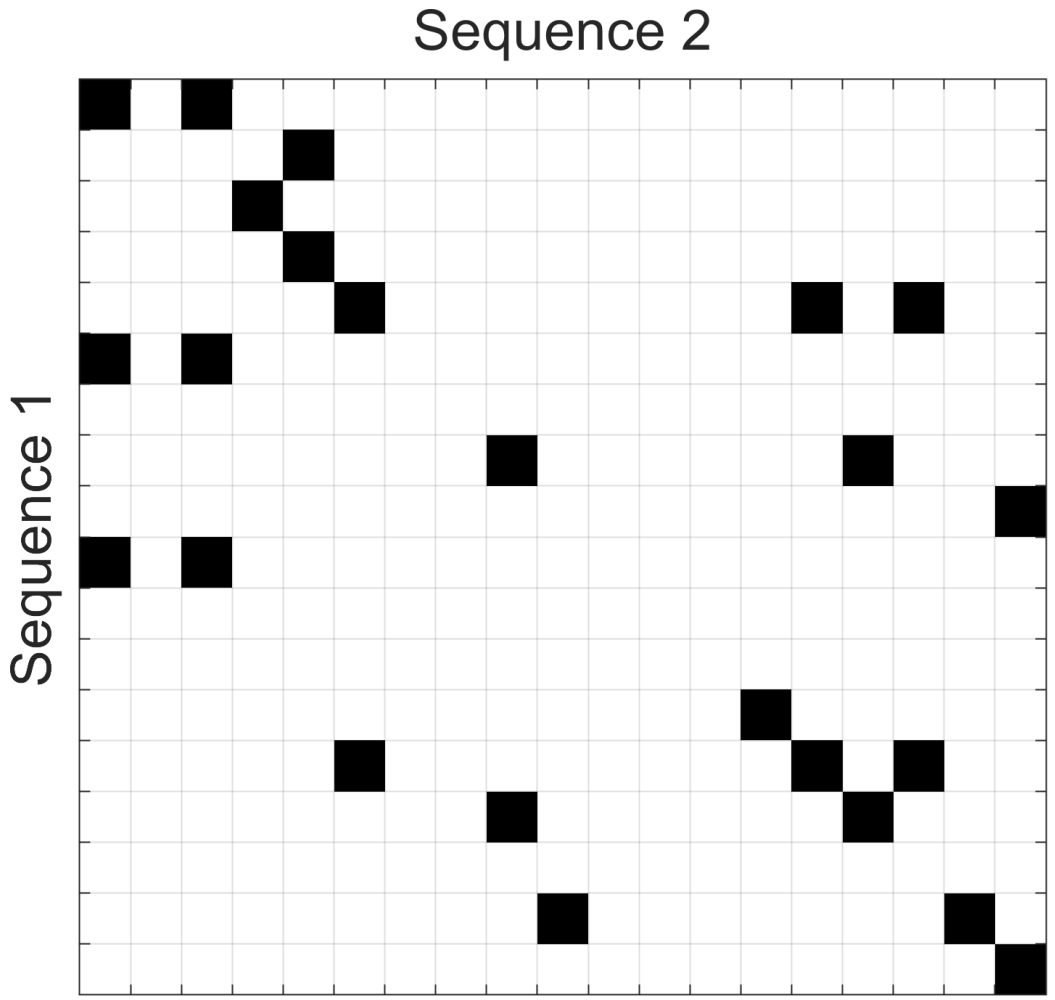
- Query: a "small" sequence.
- Database: a "large" sequence.
- Goal: Find within the database, subsequences that are **similar** to the query.

# BLAST observation/assumption

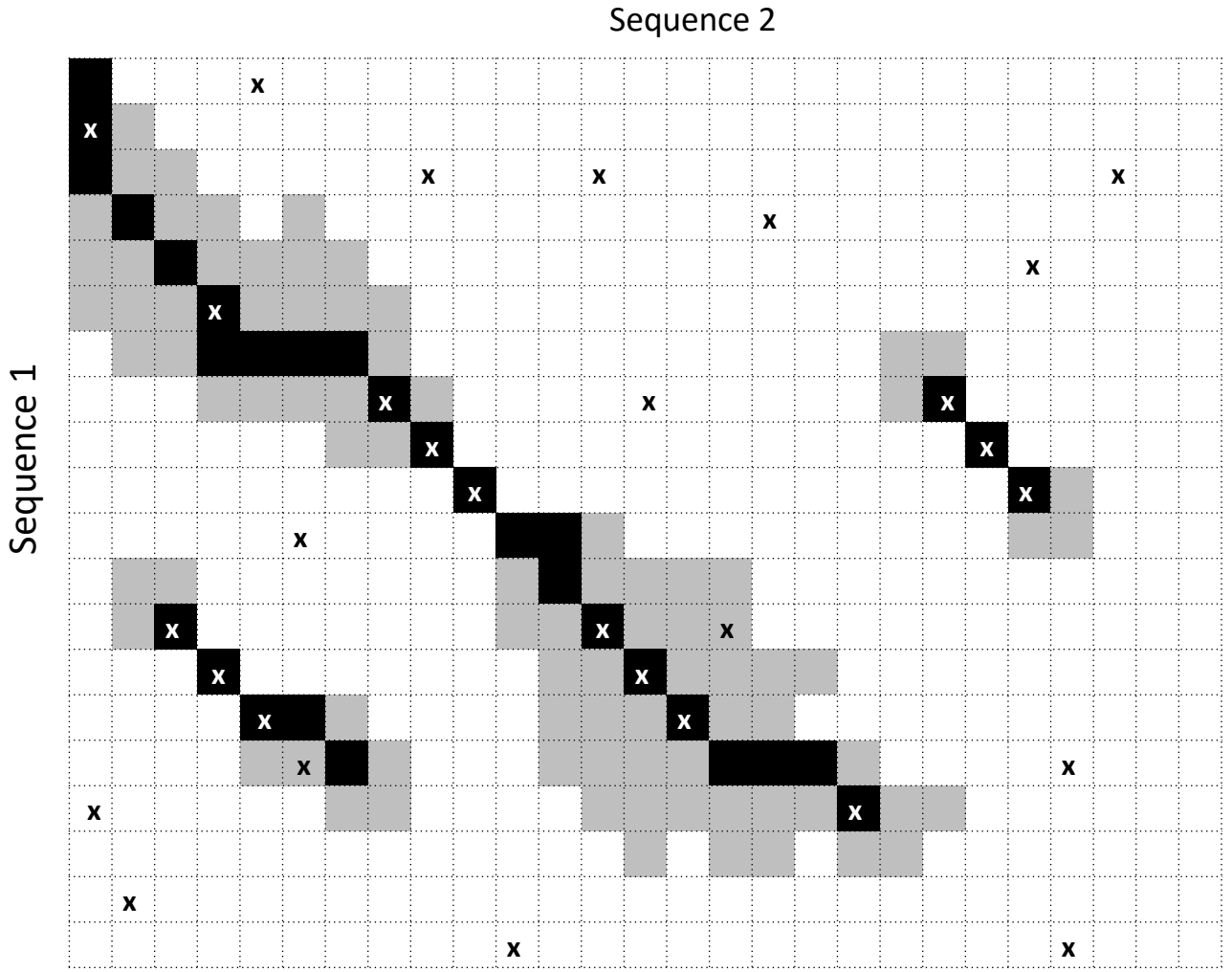
- Similar sequences share short exact matches.

```
T-HRHMTKEFTGLDMEPAF
|  |||  |    |||  ||
TITRHMN-EAWSIDMEMAF
```

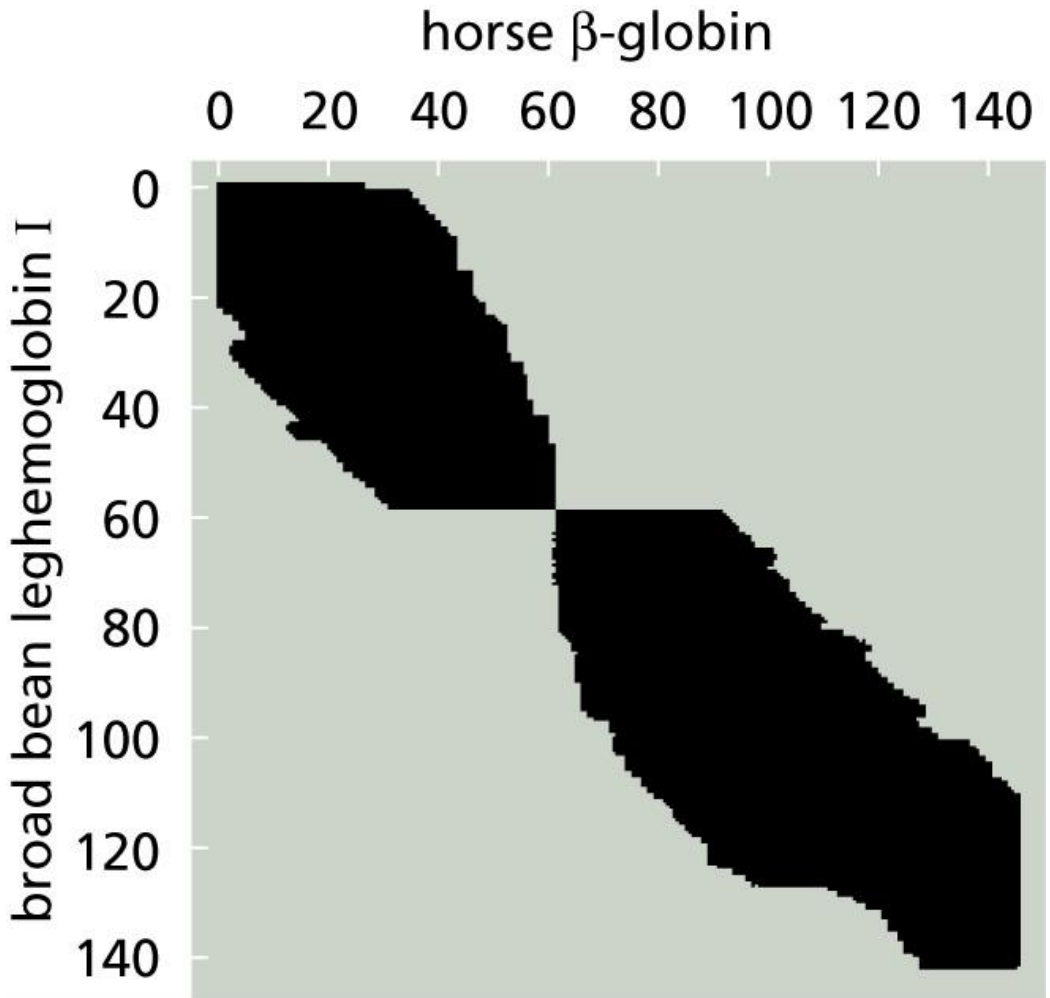
- Method:
  - Quickly identify these short exact "seed" matches.
  - Extend only around these seeds.



# BLAST: find seeds & extend



# X-drop method: O(m+n)



# Finding location of "seeds" (short exact matches)

- e.g., find all positions of **ACC** in:

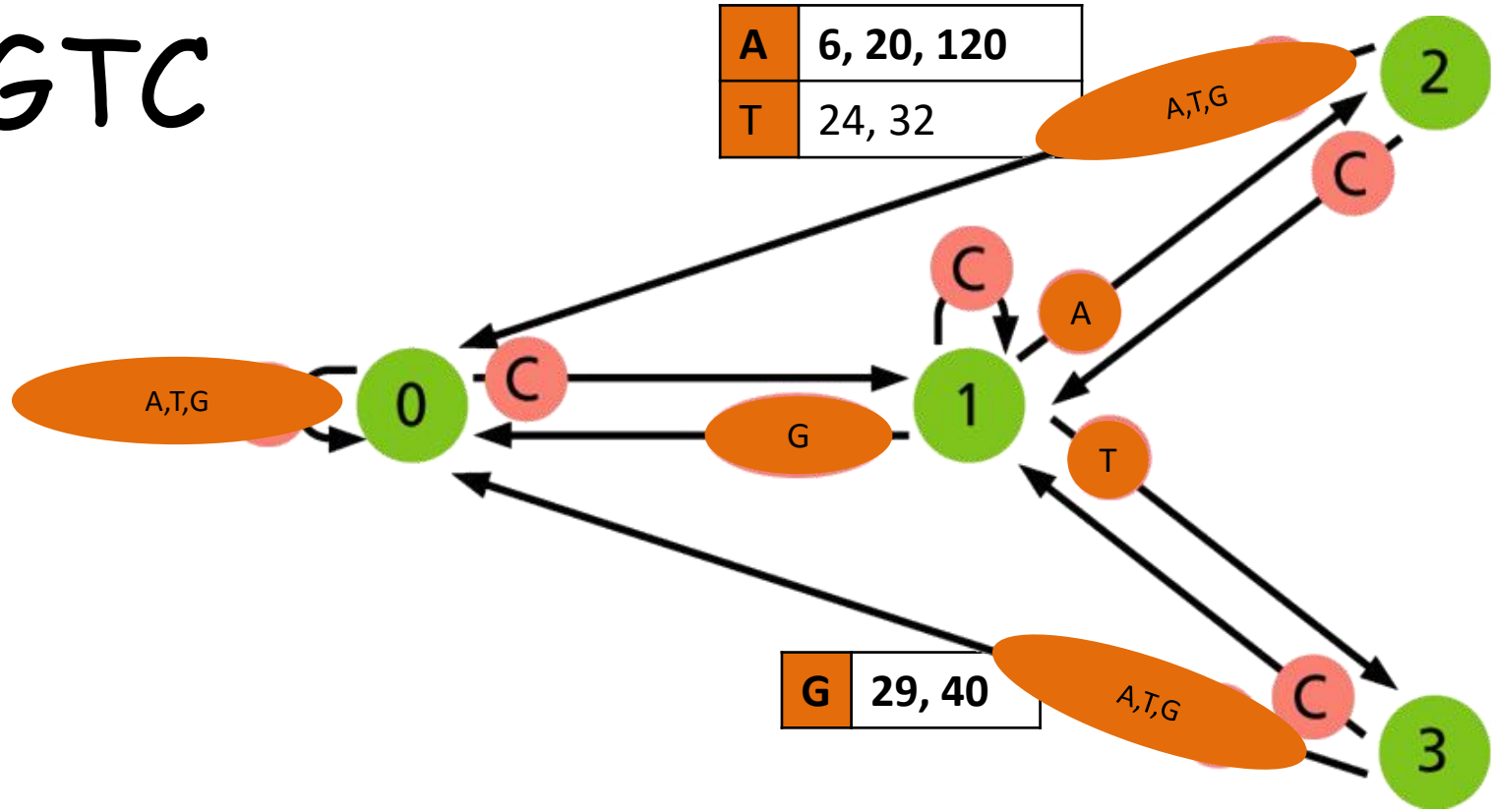
CACTGCGAAGCGGCTTCTTCAGAGCACGGGCTGGAACTGGCAGG
C**ACC**GCGAGCCCCTAGC**ACC**CGACAAGCTGAGTGTGCAGGACGA
GTCCCC**ACC**ACAC

- Pre-process: Build positions of all triplets.
  - AAA: []
  - AAC: 35
  - AAG: 8 69
  - AAT: []
  - ACA: 67 98
  - ACC: 46 62 95
  - ACG: 26 85
  - ACT: 2 36
  - AGA: 21
  - ...

# There are fast methods to identify seeds

- BLAST builds a Finite State Automata from the database
- Example Query Sequence:

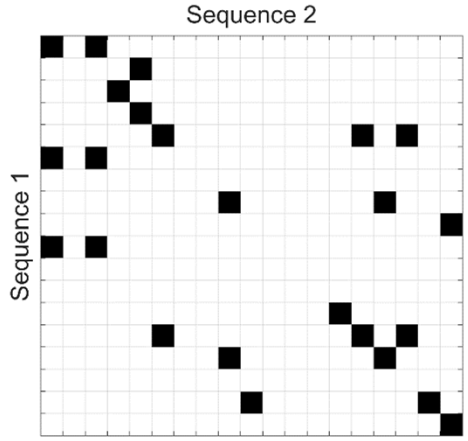
  –AGCCATGGTC




- FASTA uses hashing

# k-mer size

- Query: TTCCPSIVARSNFNVCRLPGTPEAICATYTGCIIIPGATCP

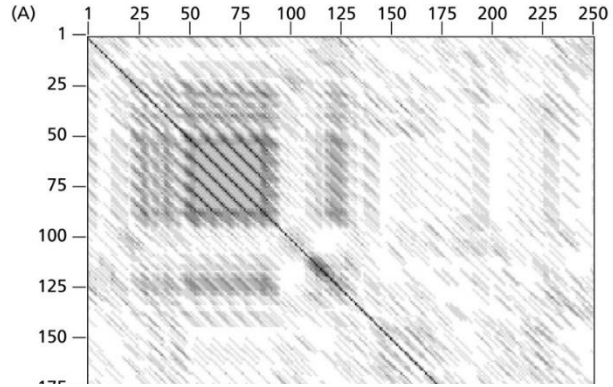



- k=1: T, T, C, C, P, S, I, V …
- k=10: TTCCPSIVAR, TTCCPSIVARS, …

- Blast Defaults: proteins: k=3, DNA: k=11

# BLAST: Additional improvements

- Low-complexity regions can cause false-positives and are filtered out
- In proteins, use additional seeds that are similar to the query protein's seeds.
  - query: **TTC**CPSIVAR
  - TTC neighbors: SSC, TSC, STC
- Don't extend each k-mer hit. Only extend if two k-mer hits are found in close proximity.




(B)

```
>sp[P04156]PRIO HUMAN  MAJOR PRION PROTEIN PRECURSOR (PRP) (PRP27-30) (PRP33
          Length = 253

Score =  312 bits (792), Expect = 5e-85
Identities = 154/236 (65%), Positives = 154/236 (65%)

Query: 64   MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYXXXXXXXXXX 123
            MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRY
Sbjct: 1    MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQP 60

Query: 124 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXTHSQWNKPSKPKTNMKHMXXXXXXXX 183
            XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXTHSQWNKPSKPKTNMKHM
Sbjct: 1    HGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTHSQWNKPSKPKTNMKHMAGAAAAGA 120

Query: 184 XXXXXXXXXXXXXXXRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCV 243
                        RPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCV
Sbjct: 121 VVGGLGGYMLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCV 180

Query: 244 NITIKQHXXXXXXXXXXXXXDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFS 299
            NITIKQH               DVKMMERVVEQMCITQYERESQAYYQRGSSMVLFS
Sbjct: 181 NITIKQHTVTTTTKGENFTETDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFS 236
```

# BLAST Summary

- Preprocessing:
  - Build an index
  - Identify neighbor k-mers for protein sequences
- Given a query
  - Remove low complexity regions
  - Extract k-mers
  - Locate k-mers in the database (efficiently, using the index)
  - Extend from these seed locations using X-drop method