

Bioinformatics 1: lecture 6

Followup for Lecture 5?

Statistics for pairwise alignments

Database searching using FASTA

Database searching using BLAST

You have seen....

Dynamic programming:

Global alignment

Global/local alignment (no end gaps. 3 ways to do it.)

Local alignment

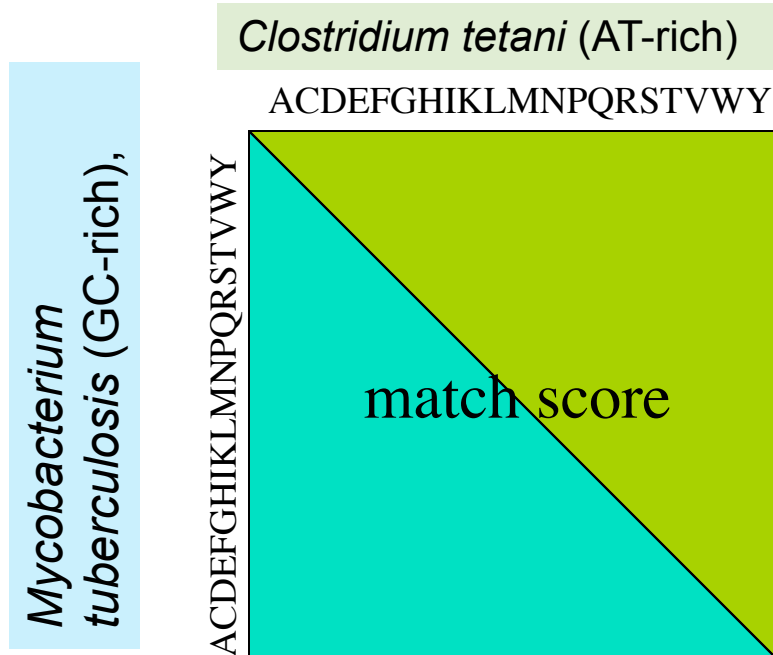
Linear gap penalty

Affine gap penalty

How many ways are there to do DP?

Asymmetric substitution matrices!

If two different species have different amino acid compositions, then the substitutions between those species are asymmetric, meaning $S[i \rightarrow j] \neq S[j \rightarrow i]$



For example, if tetanus has *more L* than tuberculosis. Then,

$$S[X_{\text{tetan}} \rightarrow L_{\text{tuber}}] > S[L_{\text{tetan}} \rightarrow X_{\text{tuber}}]$$

(where X is any amino acid)

Yu YK, Wootton JC, Altschul SF.

The compositional adjustment of amino acid substitution matrices.
Proc Natl Acad Sci U S A. 2003 Dec 23;100(26):15688-93.

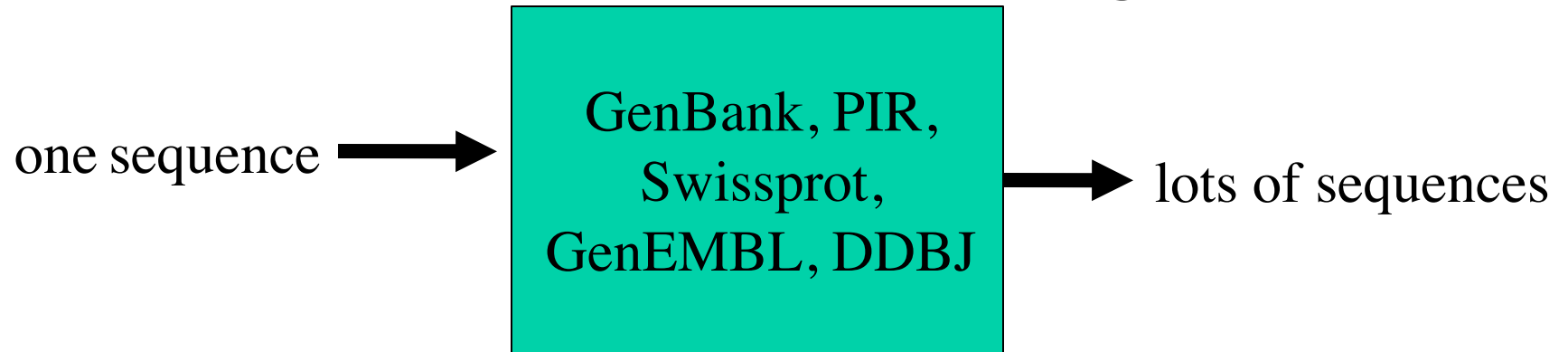
In class exercise.

Database search using NCBI blast

- Open “mystery sequence” on the course website.*
- Paste into a text file in UGENE
- right-click/Analyze/Query NCBI Blast database
- Select blastn, short, megablast. Expectation value=1. Max hits=20
- Wait....
- Open annotations.

*Provided by “crime scene investigators”

Database searching



Why do a database search?

Mol. Bio: Determination of gene function. Primer design.

Pathology, epidemiology, ecology: Determination of species, strain, lineage, phylogeny.

Biophysics: Prediction of RNA or protein structure, effect of mutation.

Searching millions of sequences

Given a protein or DNA sequence, we want to find all of the sequences in GenBank (over 17 million sequences!!) that have a good alignment score.

Each alignment score should be the *optimal* score (or a close approximation).

How do we do it?

DNA or Protein search?

- Advantages of searching **DNA** databases

Larger database. Does not assume a reading frame. Can find similarity in non-coding regions (introns, promotor regions). Can find *frameshift mutations*. Can find *pseudogenes*.

- Disadvantages

Slower. Not as sensitive. Ignores selective pressure at the protein level.

- Advantages of searching **protein** sequences

Faster. More sensitive. More biologically relevant.

- Disadvantages

Not applicable to non-coding DNA (promoters, introns, etc)

Searching using Dynamic Programming

SSEARCH Smith & Waterman

DP returns the **optimal alignment**, given the scoring function (usually *affine gap local* alignment)

Relatively slow, but more *sensitive*, and more *selective*, than FASTA and BLAST

Optimal.

sensitivity, selectivity

Searching using word matches

FASTA

W. Pearson, 1988

First searches for *k-tuples*, then links them. Results are similar to a **dot plot**. Finally, diagonals are scored using a substitution matrix, and the highest-scoring diagonals are joined.

High-scoring alignments are re-calculated using DP (local/affine).

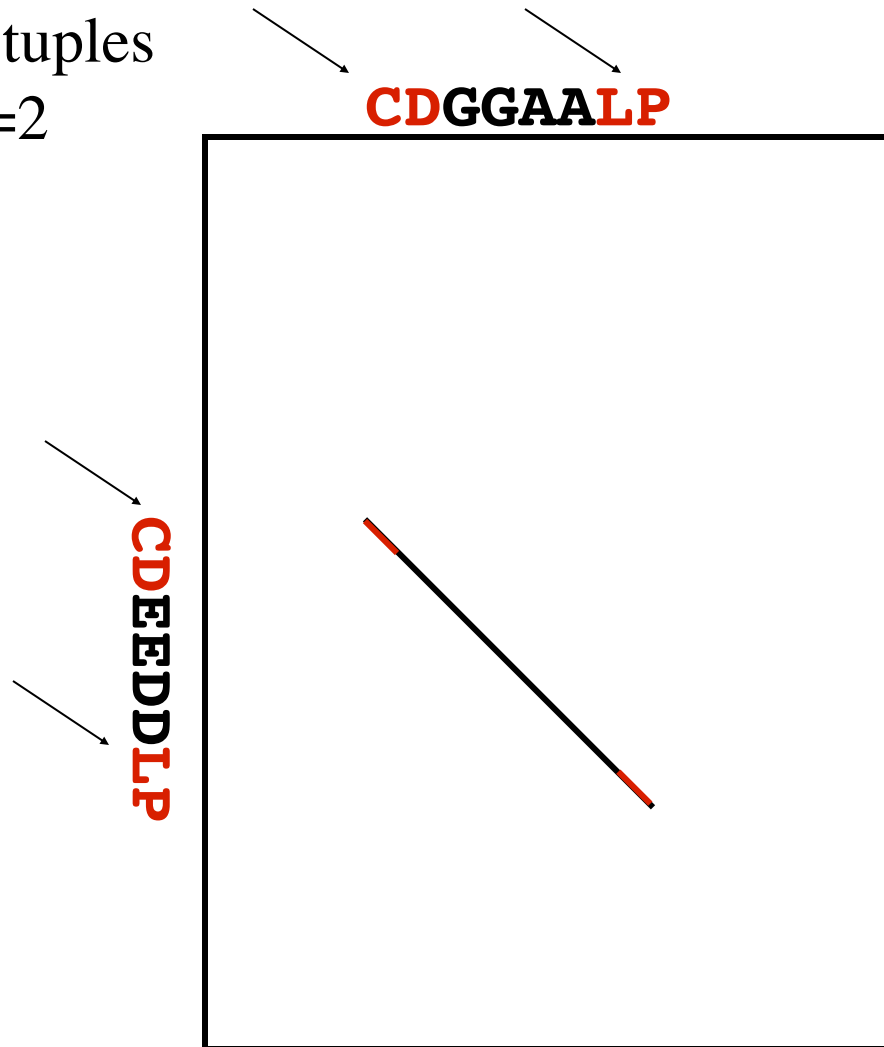
At least 50-times faster than SSEARCH. Not as sensitive. Final DP step makes it more *sensitive*, but less *selective*.

FASTA is a Heuristic alignment method, not Optimal.

heuristic

FASTA

k-tuples
k=2

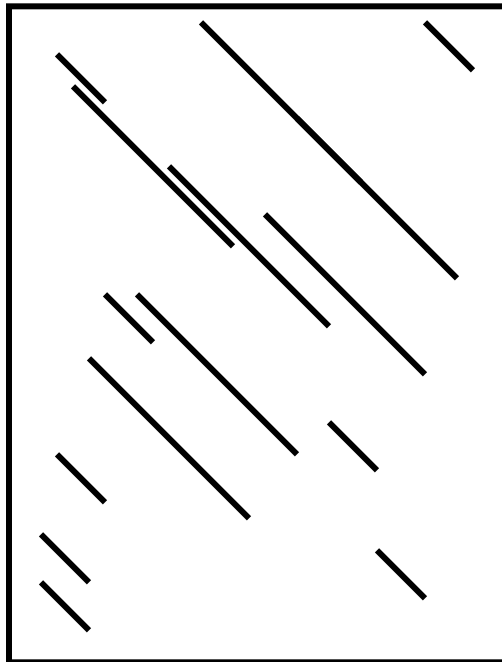


Finding identity matches is very fast.

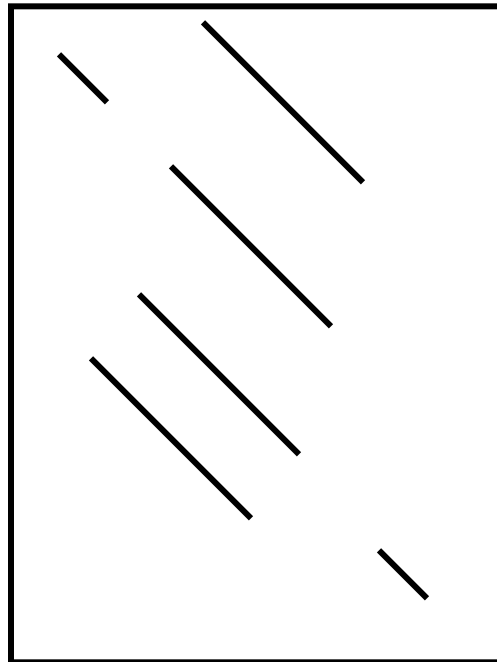
If two *k-tuples* are separated by exactly the same amount in both sequence, draw a diagonal. A gapless alignment.

FASTA

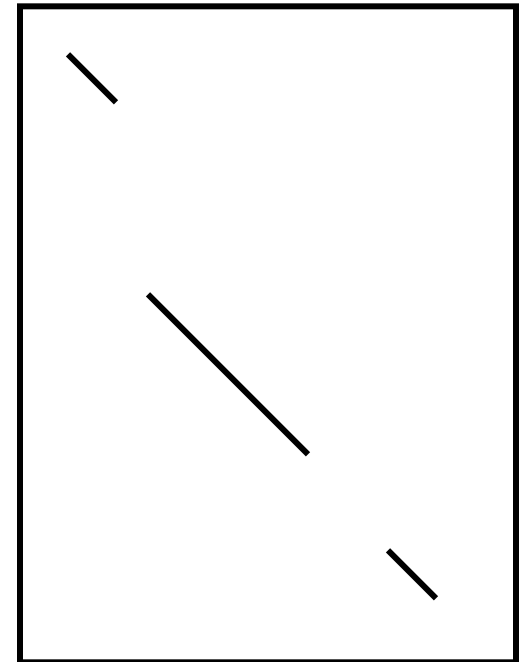
Find all gapless alignments



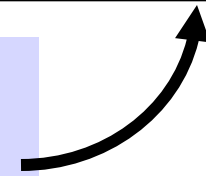
Score them using BLOSUM, keep the best



Connect them using simple affine gap. (gap ext.= 0)



If this alignment one of the best scores in the database search, go back and realign it to the query using DP.



Searching using lookup tables

BLAST

S. Altschul *et al.*

First make a set of lookup tables for all 3-letter (protein) or 11-letter (DNA) matches.

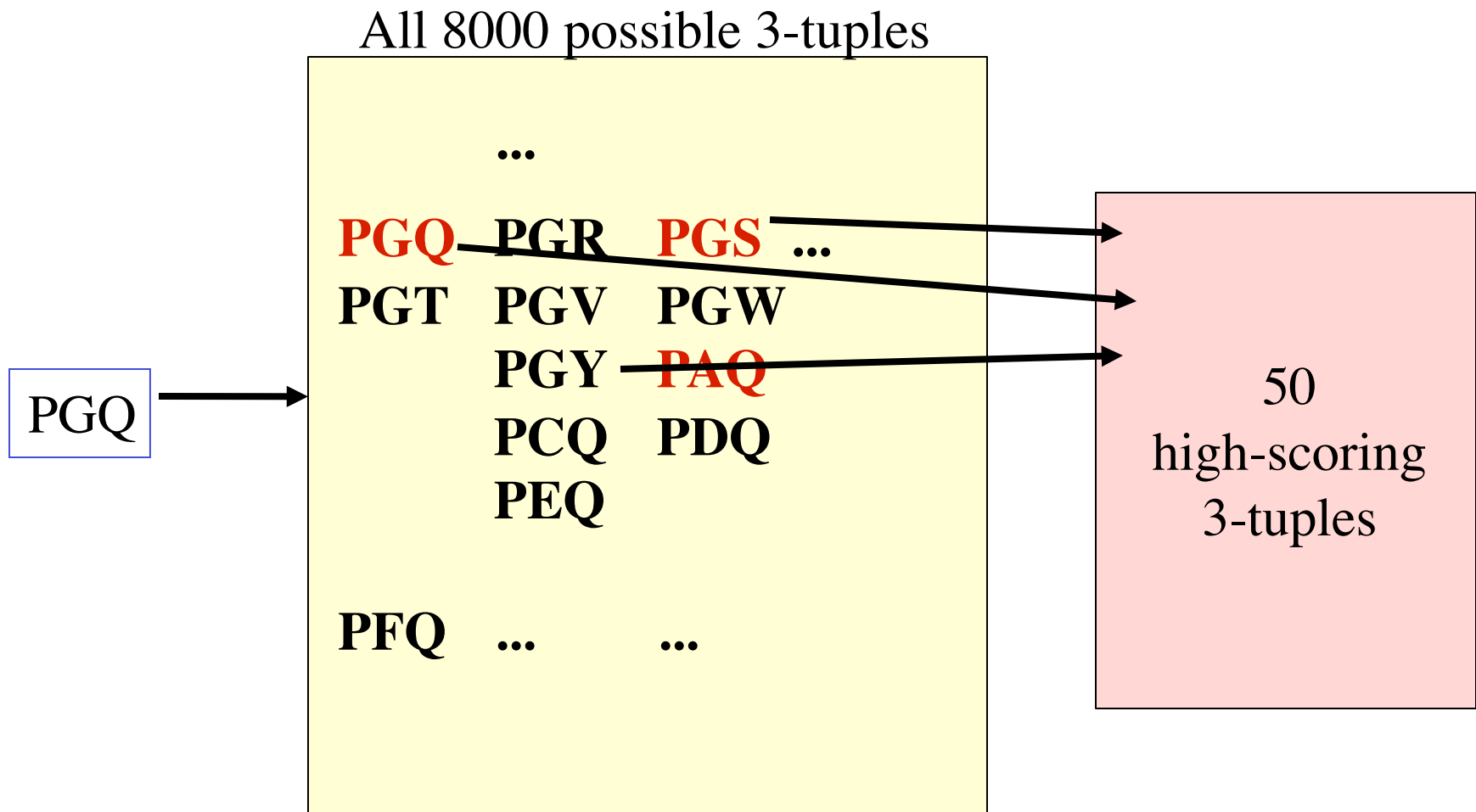
Make another lookup table: the locations of all 3-letter words in the database.

Start with a match, extend to the left and right until the score no longer increases.

Very fast. Selective, but not as sensitive as SSEARCH.
Good statistics.

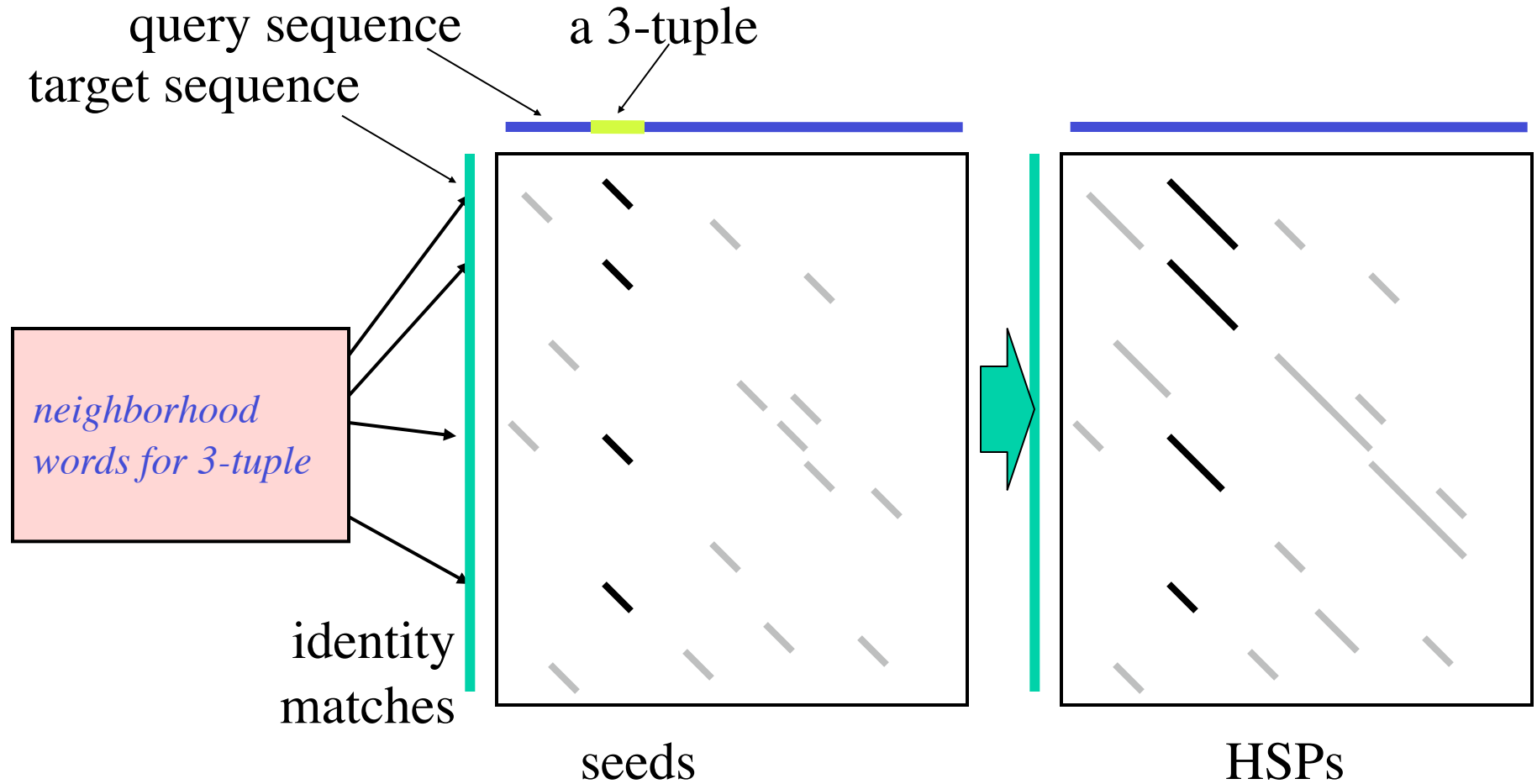
Heuristic.

BLAST, precalculations



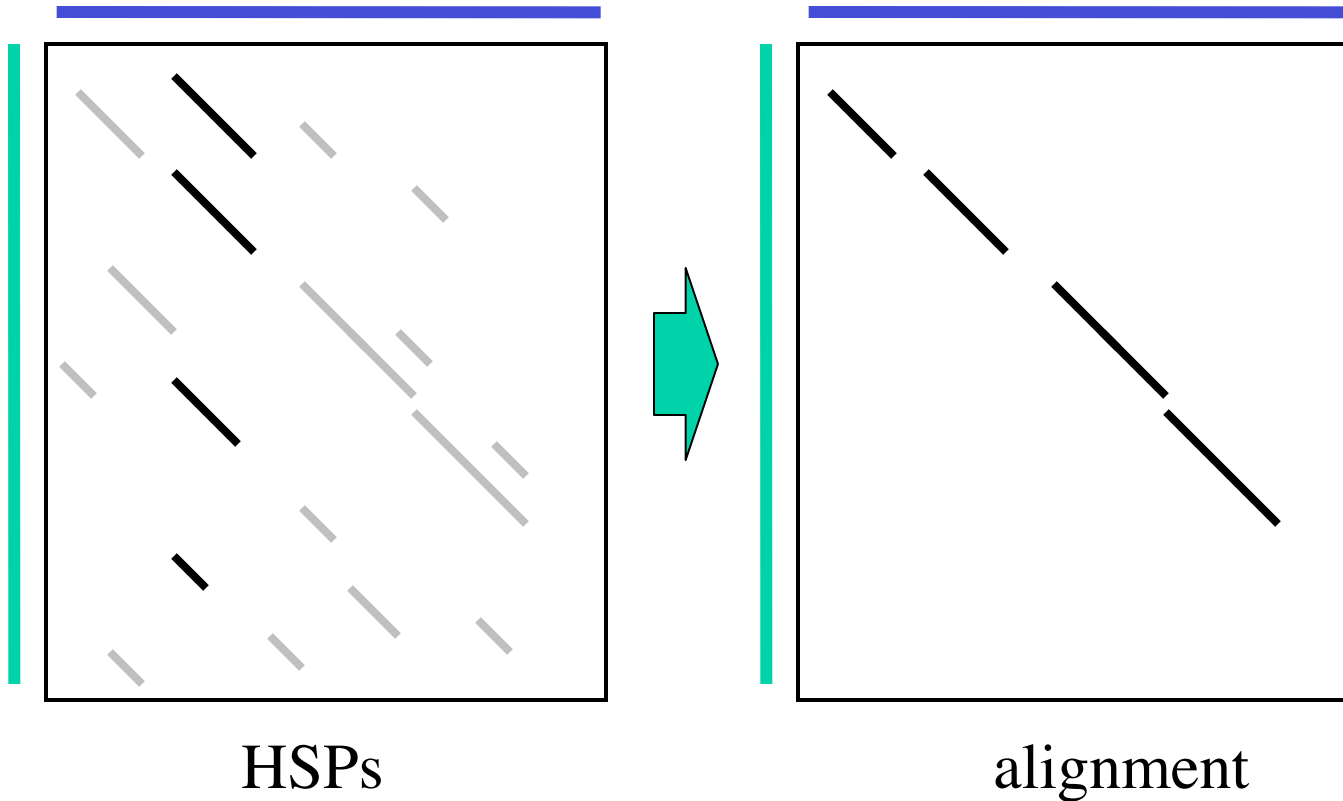
Each 3-tuple is scored against all 8000 possible 3-tuples using BLOSUM. The top scoring 50 are kept as that 3-tuple's "neighborhood words"

BLAST



For every 3-residue window, we get the set of 50 nearest neighbors. Use each word to get identity matches (seeds). Then extend the seed alignments as long as the score increases.

BLAST



The best extended seeds are called HSPs (high scoring pairs). The top scoring HSP is picked first, then the second (as long as it falls "northwest" or "southeast" of the first.), and so on.

Summary, database searches

- SSEARCH uses DP, slowest but most sensitive/selective. Optimal
- FASTA, much faster, Heuristic.
- BLAST, faster, uses lookup tables. Like FASTA + SSEARCH