BLAST

Basic Local Alignment Search Tool

Less accurate than Smith-Waterman, but over 50 time faster.

1. Find ungapped matches of a small fixed length, w, that score at least T.

2. Extend matches in both directions in an attempt to find an alignment with a score exceeding S.

Segment pairs whose scores cannot be improved by extending or trimming are called high scoring pairs (HSPs).

Typical values for w are 3 when aligning proteins and 11 when aligning nucleic acids.

e-values and p-values

The expected number of HSPs with a score of at least S is given by the formula:

$$E = Kmne^{-\lambda S}$$

Doubling the length of the query sequence (m) or the size of the database (n) should double the number of HSPs.

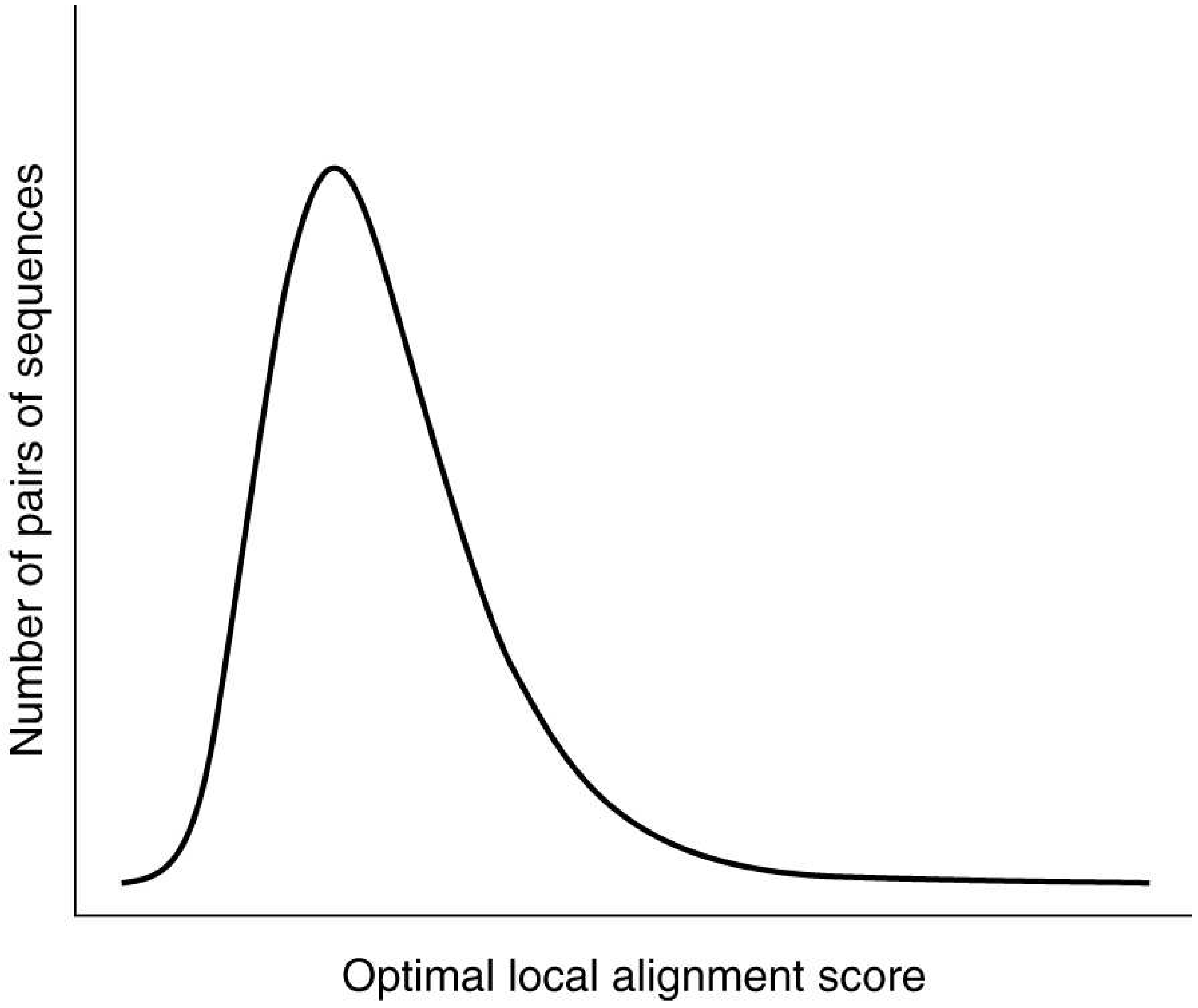
To obtain score $2x$, score x must be obtained twice in a row.
So one expects E to decrease exponentially with score.

The probability of observing a score $\geq S$ is:

$$1 - \exp(-Kmne^{-\lambda S})$$

This is the p-value.

Extreme value distribution



FASTA

k-tuples, strings of length k.

k = 1 - 2 for proteins and 4-6 for nucleic acids.

Construct a look-up table with all k-tuples in the database.

Look up all k-tuples from the query string and mark matching database k-tuples. Sort matches by the difference in their indices (i-j).

Nearby matches on the same diagonal are joined to form an ungapped local alignment region.

Join nearby high scoring regions on different diagonals.

For the best regions, perform dynamic programming in a window around the region.

Applications of pattern matching: DNA

Identifying whether a DNA molecule has a subsequence that will be recognised by a protein.

- Restriction enzymes that cut DNA.
  HindII (the first identified restriction enzyme) cuts “GT[TC][AG]AC”
  EcoRI cuts “GAATTC” between G and A
  http://rebase.neb.com/rebase/rebase.html

- DNA methylation
  e.g. E. coli DNA adenine methyltransferase (DAM) recognises GATC

- Transcription factor binding sites.
  their presence can promote or block transcription

Applications of pattern matching: PROSITE

Members of some protein families can be recognised by the presence of a specific pattern in a protein’s sequence, e.g.

ATP/GTP-binding site motif A (PS00017)

```
[AG]-x(4)-G-K-[ST]
```

Zinc finger C2H2 type domain signature (PS00028)

```
C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H
```

EGF-like domain (PS00022)

```
C-x-C-x(2)-{V}-x(2)-G-{C}-x-C
```

http://prosite.expasy.org/

Patterns vs Profiles

Patterns are qualitative

— they either match or they don’t!

Profiles are quantitative

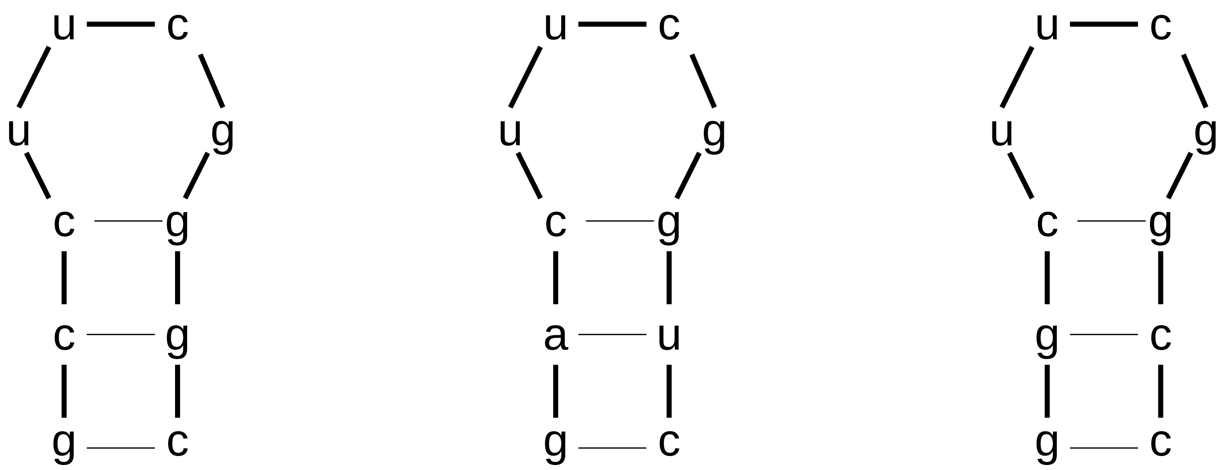
— numerical weights are associated with matches and mismatches at various positions

— can give greater sensitivity, allowing family membership to be detected even if the family has only a few highly conserved sequence positions

— Hidden Markov Model are commonly used to derive profiles

Covariance

Can give clues about base pairing and RNA secondary structure.



```
 *      *
gccuucgggc
gacuucgguc
ggcuucggcc

(((....)))
```