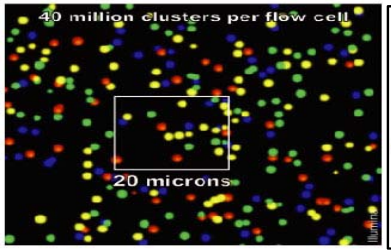
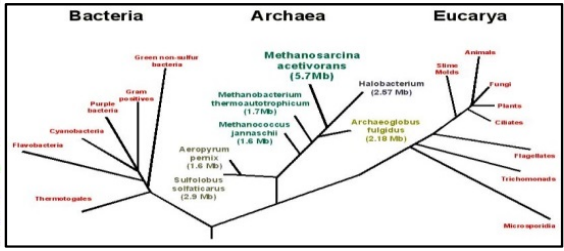
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



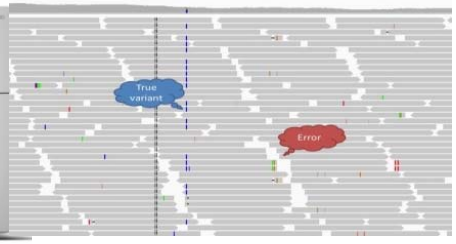


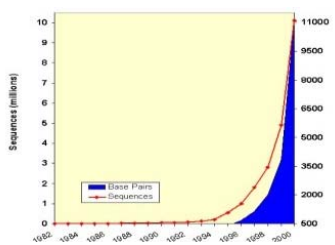
Sequence Database Search

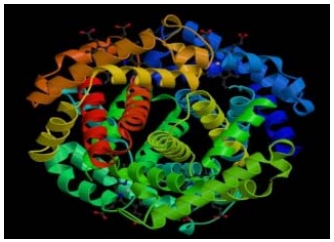
北京大学生物信息学中心 高歌

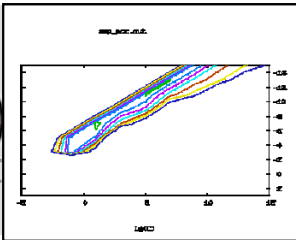
Ge Gao, Ph.D.

Center for Bioinformatics, Peking University

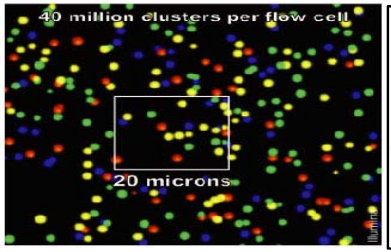


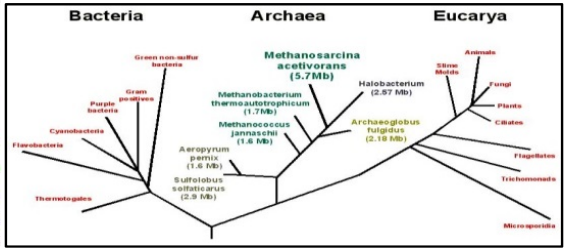






TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



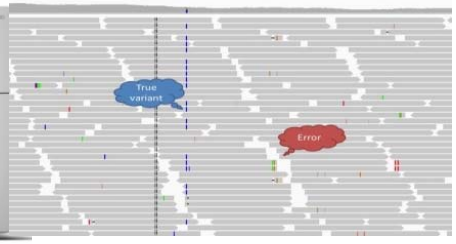


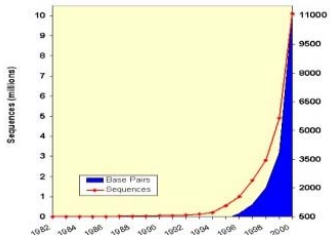
Unit 2:
BLAST Algorithm: a Primer

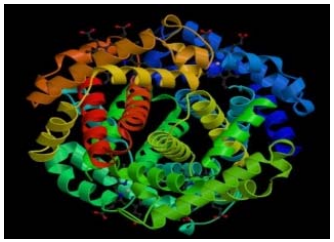
北京大学生物信息学中心 高歌

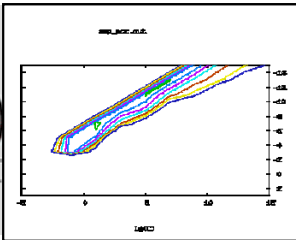
Ge Gao, Ph.D.

Center for Bioinformatics, Peking University









J. Mol. Biol. (1990) 215, 403–410

Basic Local Alignment Search Tool

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2]
Eugene W. Myers[3] and David J. Lipman[1]

[1]*National Center for Biotechnology Information*
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894, U.S.A.

[2]*Department of Computer Science*
The Pennsylvania State University, University Park, PA 16802, U.S.A.

[3]*Department of Computer Science*
University of Arizona, Tucson, AZ 85721, U.S.A.

(Received 26 February 1990; accepted 15 May 1990)

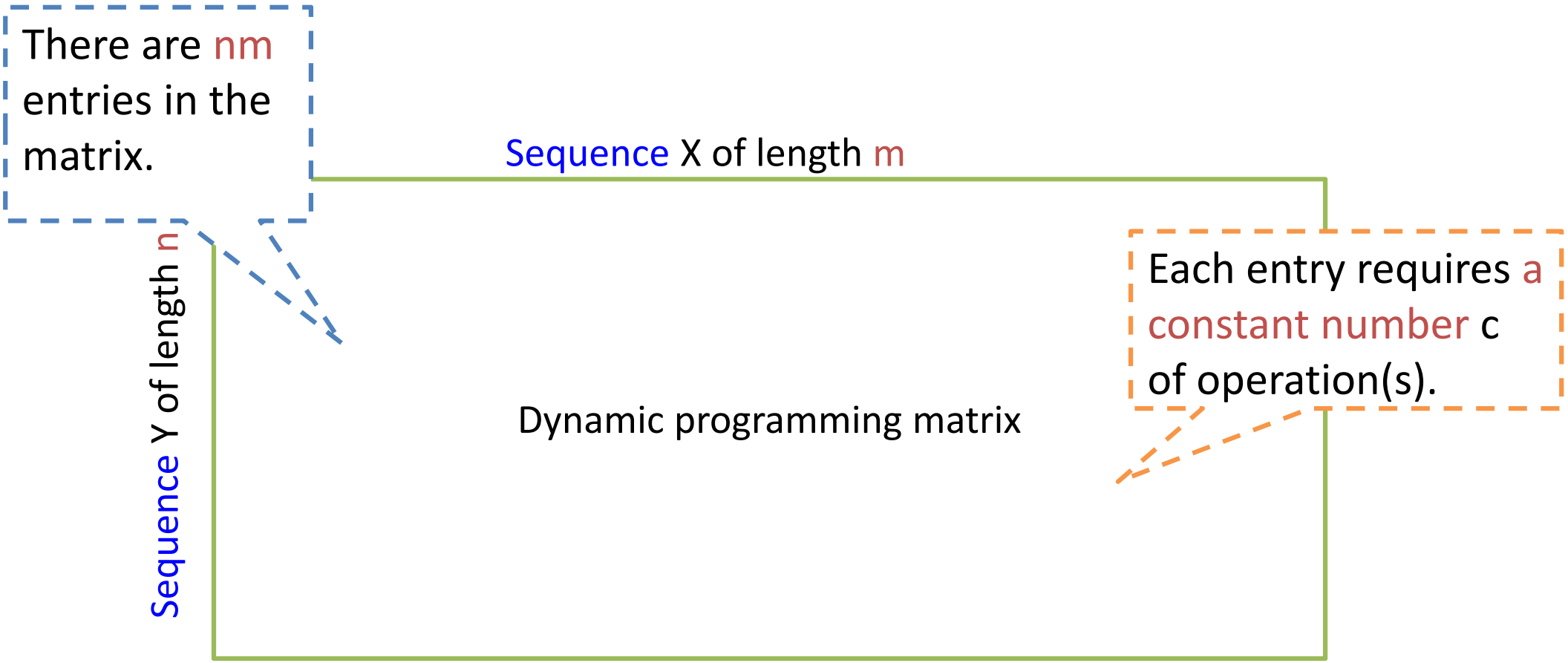
A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP scores allow an analysis of the performance of this method as well as the statistical significance of alignments it generates. The basic algorithm is simple and robust; it can be implemented in a number of ways and applied in a variety of contexts including straightforward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. In addition to its flexibility and tractability to mathematical analysis, BLAST is an order of magnitude faster than existing sequence comparison tools of comparable sensitivity.

Basic local alignment search tool

SF Altschul, W Gish, W Miller, EW Myers... - Journal of molecular ..., 1990 - Elsevier

A new approach to rapid sequence comparison, **basic local alignment search tool** (BLAST), directly approximates alignments that optimize a measure of **local** similarity, the maximal segment pair (MSP) score. Recent matlmmatical results on the stochastic properties of **...**

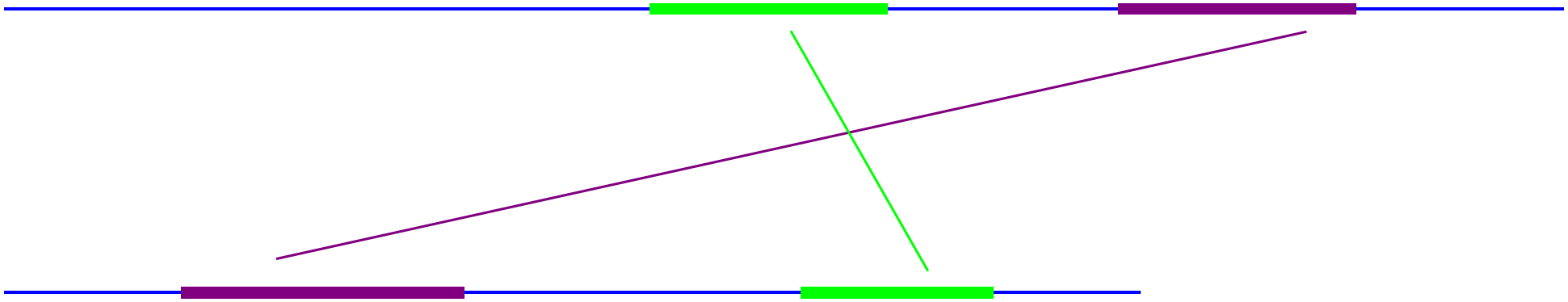
Cited by 47577 Related articles All 98 versions Cite



c*m*n operations needed in total, for one pair-wise alignment.

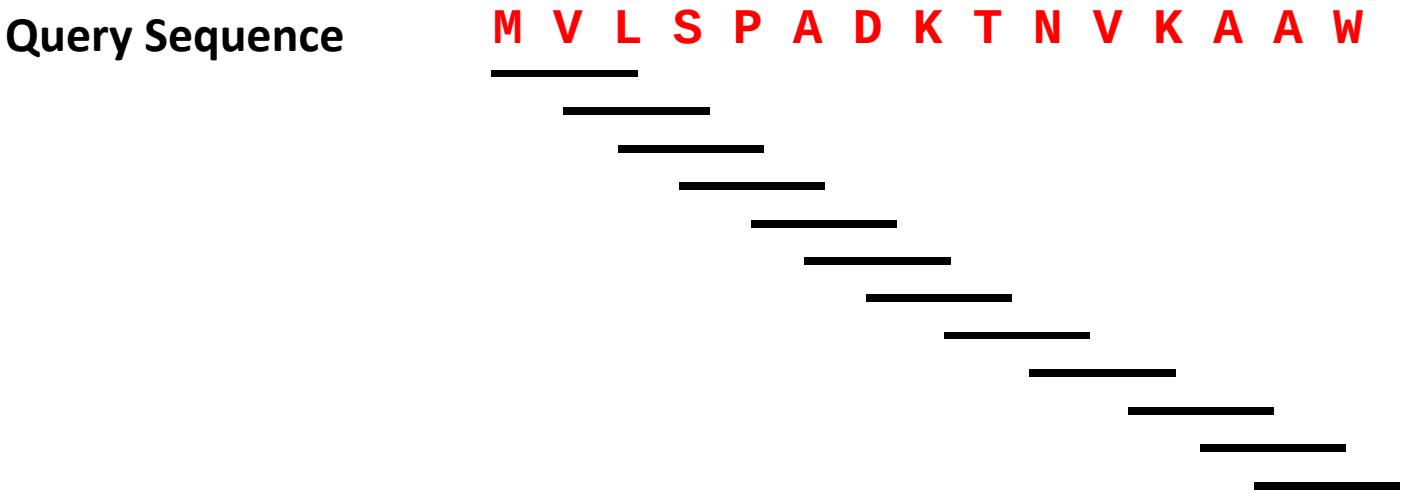
BLAST Ideas: Seeding-and-extending

1. Find matches (seed) between the query and subject
2. Extend seed into High Scoring Segment Pairs (HSPs)
   - Run Smith-Waterman algorithm on the specified region only.
3. Assess the reliability of the alignment.



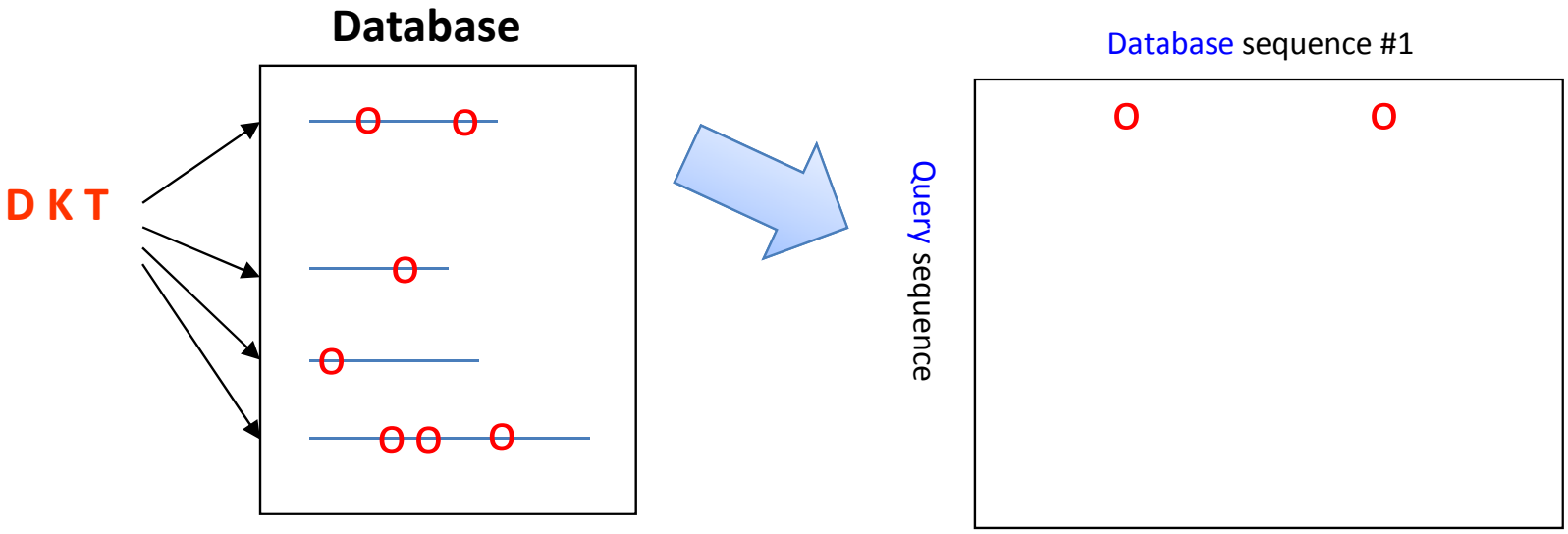
Seeding

For a given word length *w* (usually 3 for proteins and 11 for nucleotides), slicing the query sequence into multiple continuous "seed words"

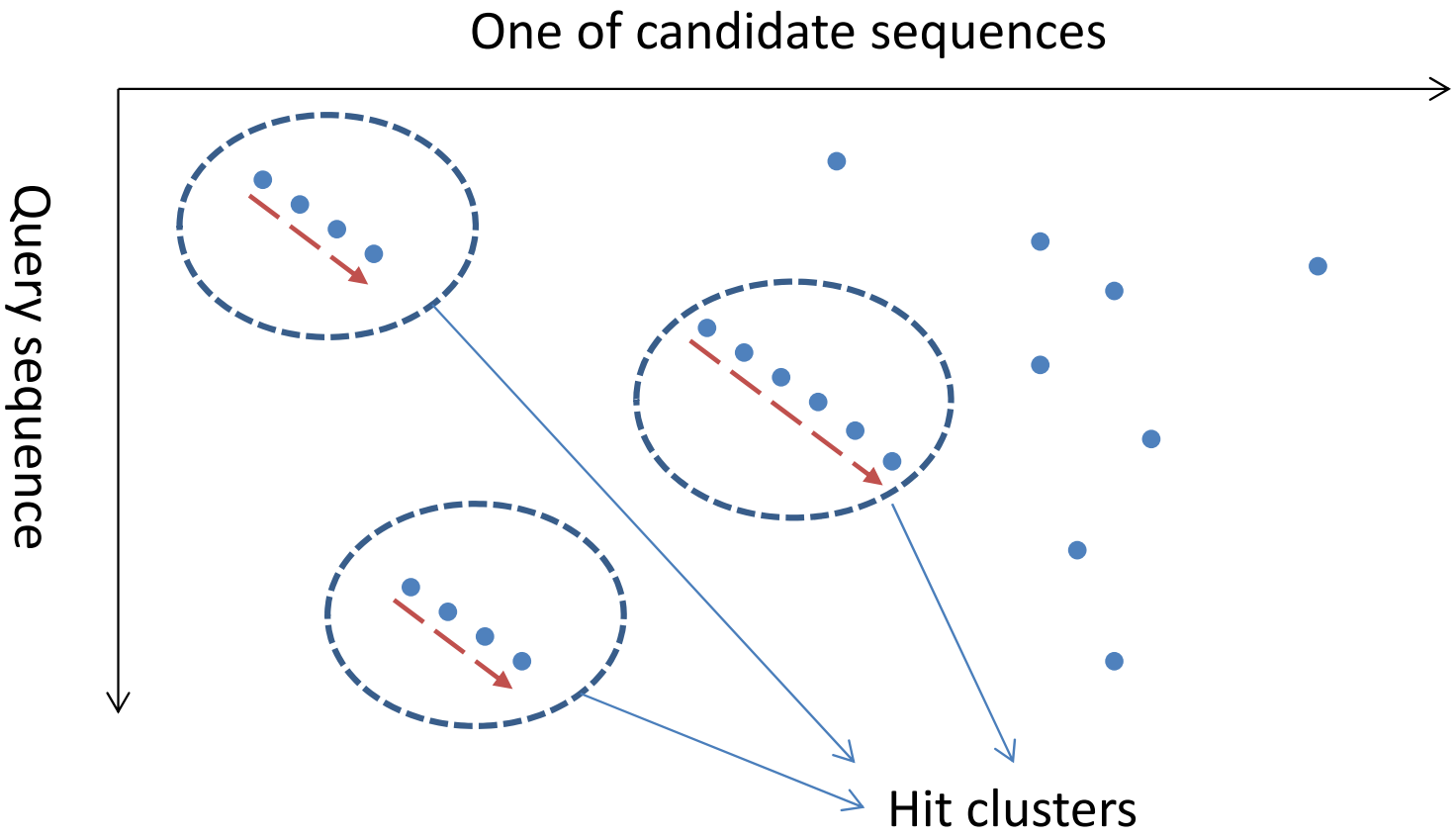


Speedup: Index database

The database was pre-indexed to quickly locate all positions in the database for a given seed.

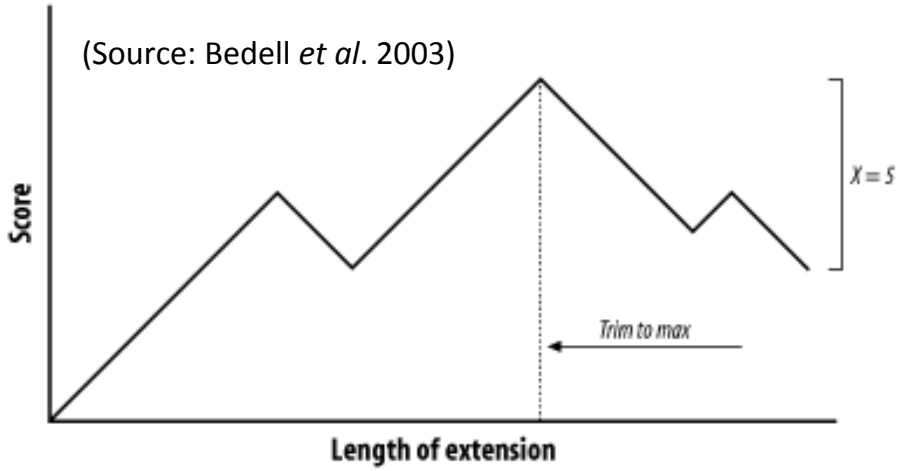


Diagonal *and* Two-hits



One of candidate sequences

Query sequence

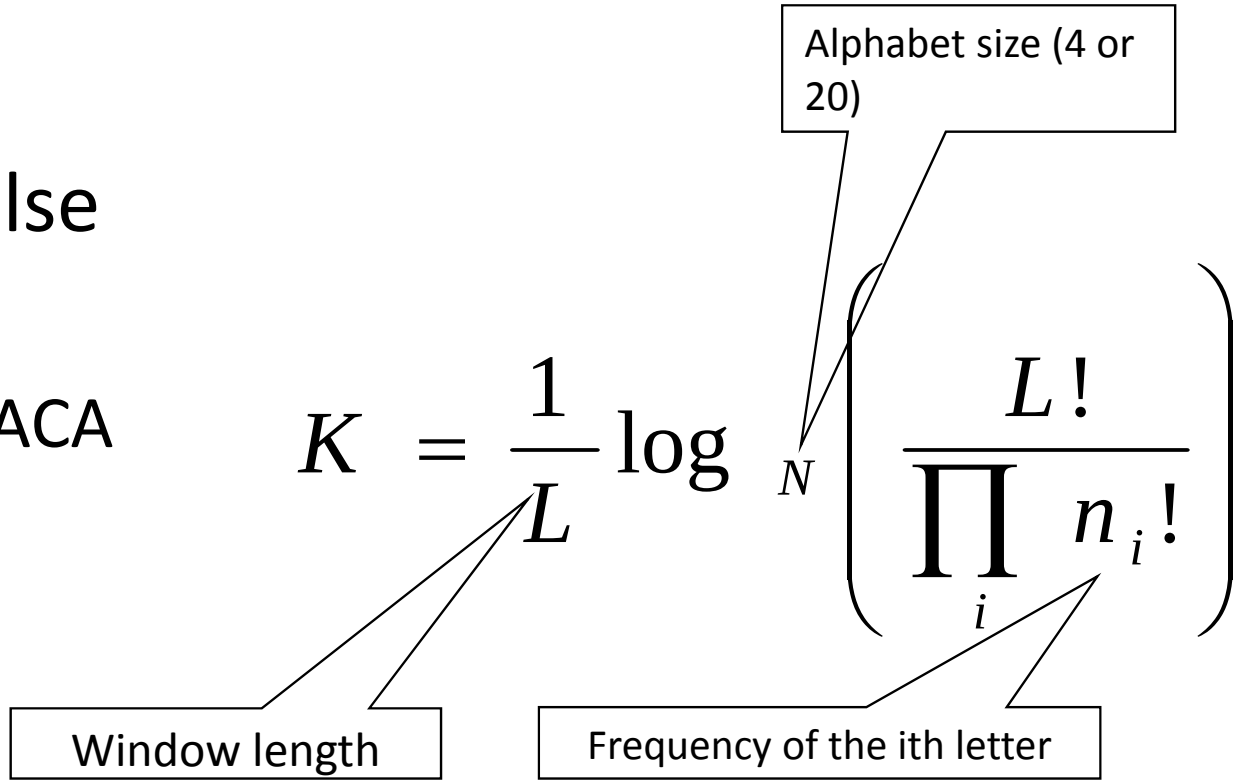


$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \\ 0 \end{cases}$$

Speedup: mask low-complexity

- Low complexity sequences yield false positives.
  - CACACACACACACACA
  - KLKLKLKLKLKL



For example, for typical microsatellite “CACACACACACACACA”, with window length 6:

$$K = \frac{1}{6}\log_4\left(\frac{6!}{n_A!*n_C!*n_G!*n_T!}\right)$$

$$= \frac{1}{6}\log_4\left(\frac{6!}{3!*3!*0!*0!}\right)$$

$$= \frac{1}{6}\log_4\left(\frac{6!}{3!*3!}\right)$$

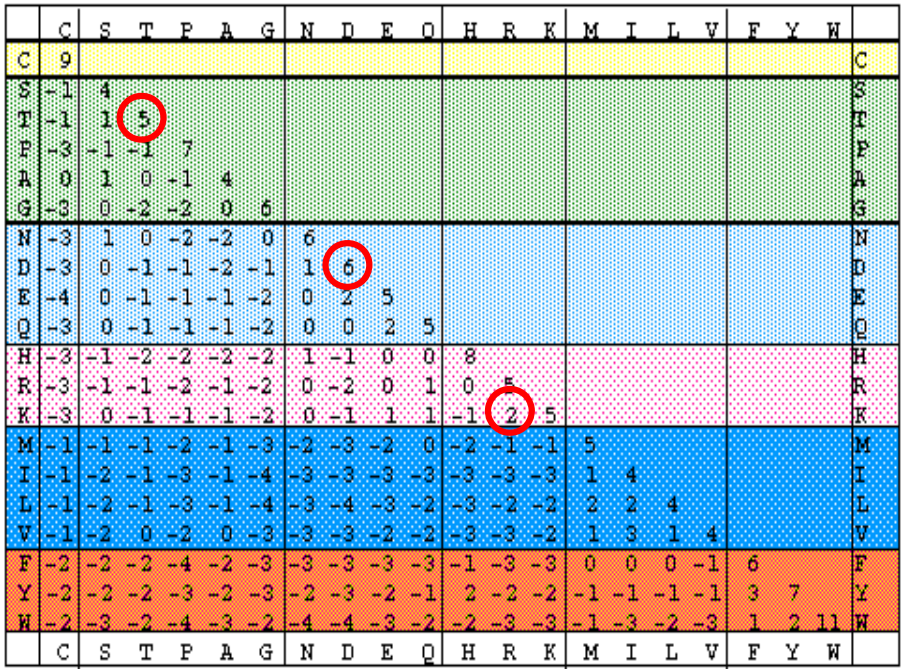
$$= \frac{1}{6}\log_4 20$$

$$= 0.36$$

To improve sensitivity, in addition to the seed word itself, the BLAST also use these highly similar “neighbourhood words” (based on the substitution matrix) for seeding.

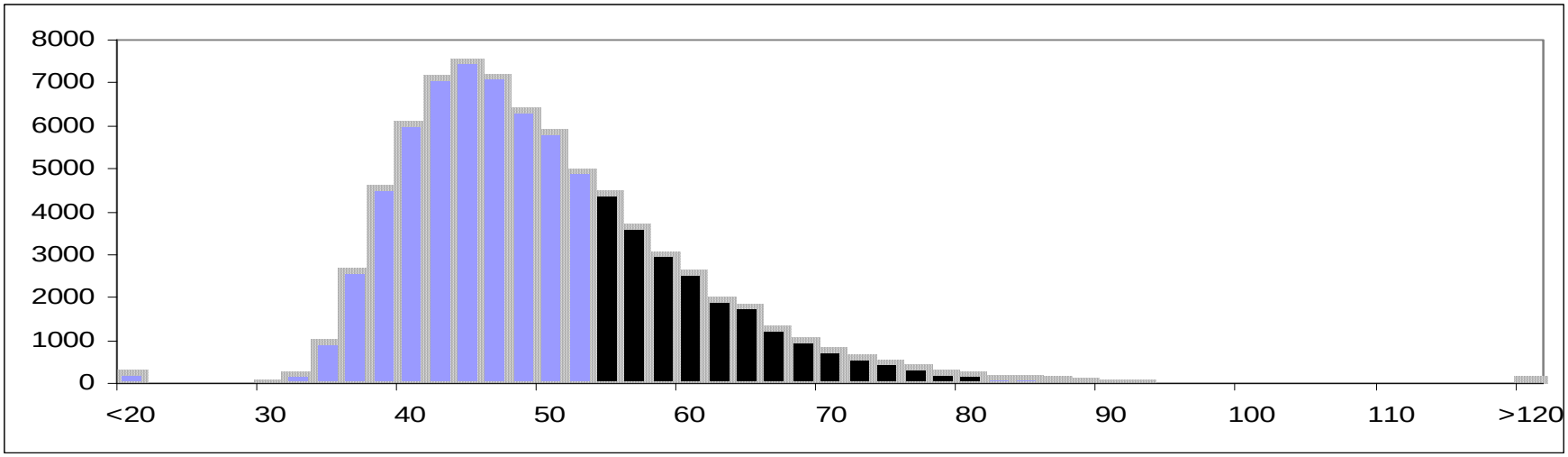
```
D K T
|   | ➔ 6+2+5 = 13
D R T
```

```
DKT 16
DRT 13
DET 12
DKS 12
DQT 12
EKT 12
DKA 11
DKN 11
DKV 11
DNT 11
DST 11
NKT 11
DAT 10
DDT 10
DHT 10
DKC 10
DKD 10
DKE 10
DKI 10
DKK 10
DKL 10
DKM 10
DKP 10
DKQ 10
DKR 10
DMT 10
DPT 10
DTT 10
QKT 10
SKT 10
```



Quality Assessment

Given the large data volume, it's critical to provide some measures for assessing the statistical significance of a given hit.



For a given amino acid, we have the chance of 1/20 to have a random match (as there are just 20 different amino acids in total).

Thus, for an amino acid sequence with length L, the probability of having a random match across the full length is $(1/20)^L$

Say you have a 6 AA peptide (not so unusual, e.g. the Tryptophyllin-T2-6 in *Phyllomedusa azurea* or "Orange-legged monkey frog,橙腿猴树蛙" is a 6 AA peptide), then the odd would be $(1/20)^6 = 1.56 * 10^{-8}$
Looks not so big, huh?

But you're searching Swiss-Prot databases which contains 192,206,270 amino acids in 540,958 sequences (Sept 18th, 2013). Then you would expect to have $(1/20)^6 * 192,206,270 = 3.00$ matches by chance, for any 6 AA peptide!

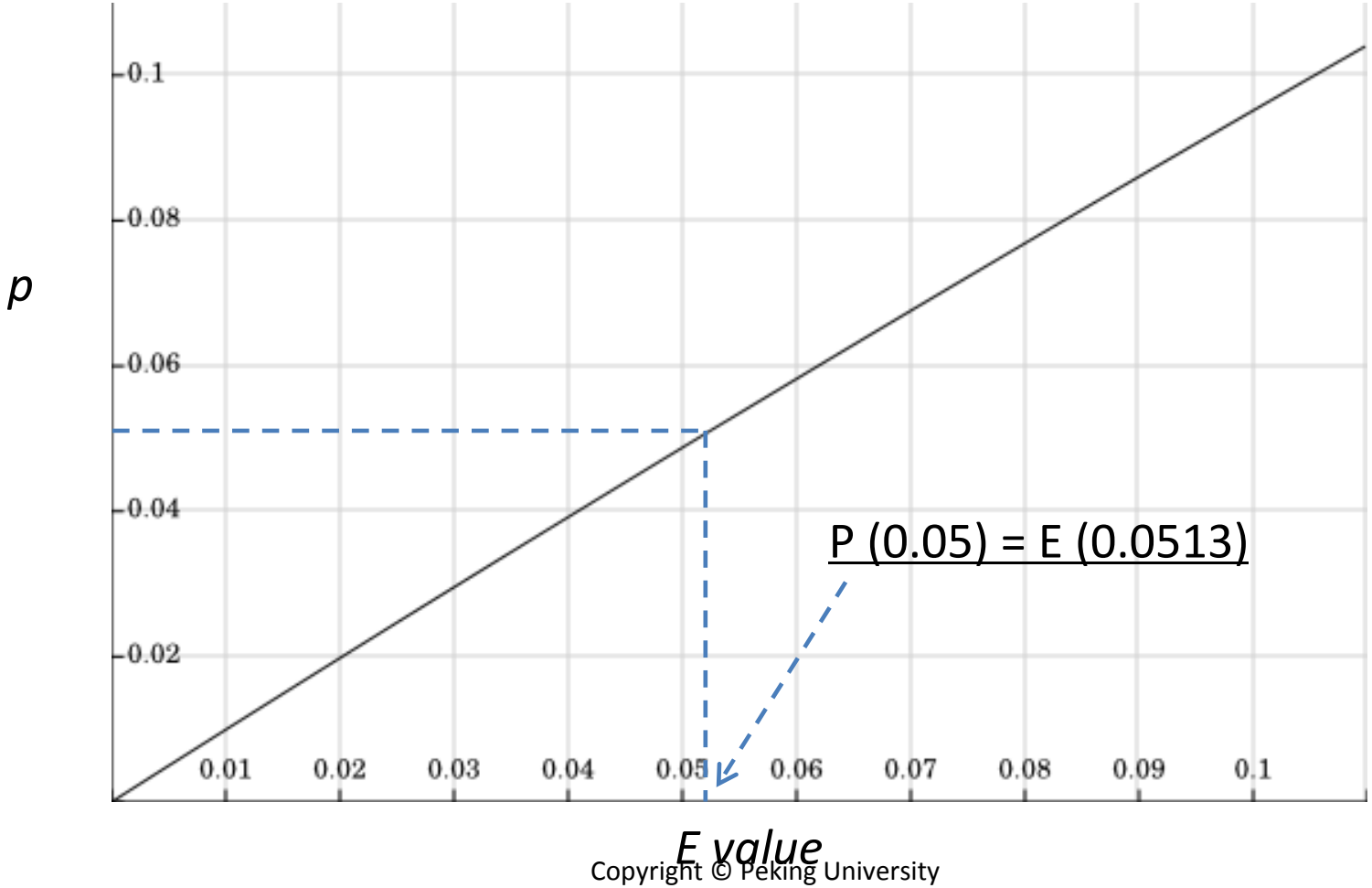
E-Value: How a match is likely to arise by chance

- The expected number of alignments with a given score that would be expected to occur at random in the database that has been searched
  - e.g. if E=10, 10 matches with scores this high are expected to be found by chance

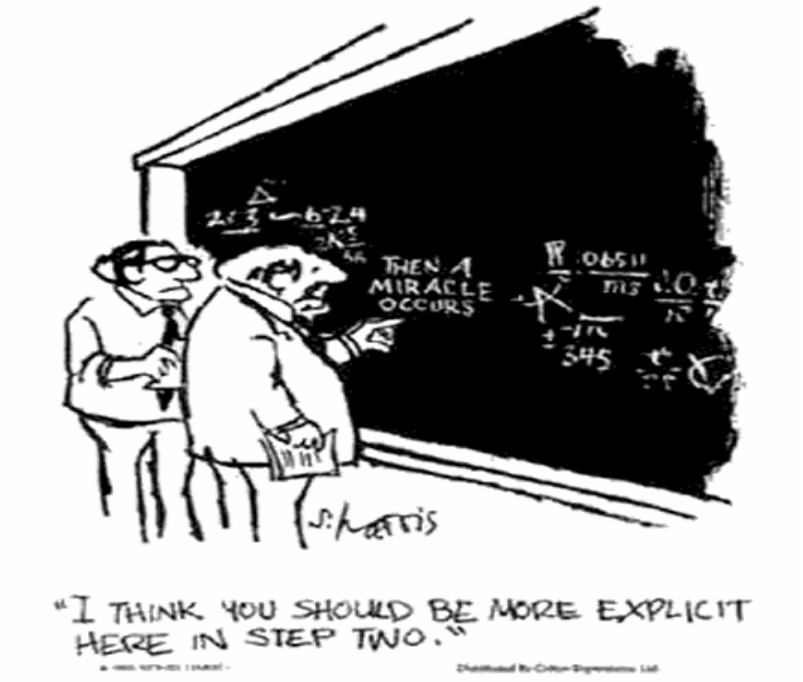
$$E = kmne^{-\lambda S}$$

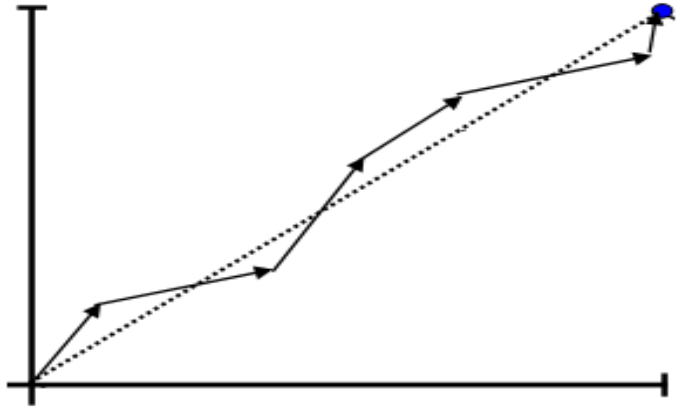
$$E = kmne^{-\lambda s}$$

$$p = 1 - e^{-E}$$

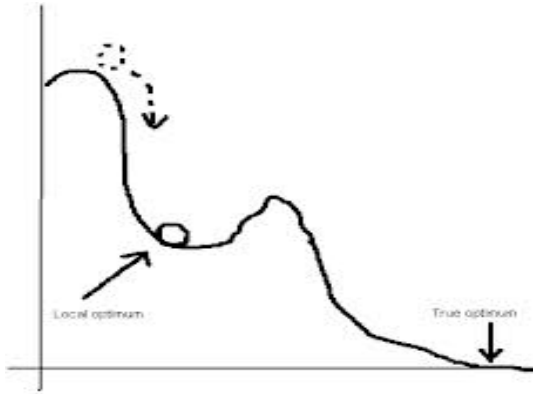


Heuristic (pronounced hyu-RIS-tik, Greek: "Εὑρίσκω", "**find**" or "**discover**") refers to experience-based techniques for problem solving, learning, and discovery. (Source: Wikipedia)





Not best,
but good *enough*



- Key heuristics in BLAST
  - Seeding-and-extending: looking for seeds of high scoring alignments ONLY

  ➔Use dynamic programming selectively

- Tradeoff: speed vs. sensitivity
  - Empirically, 1000 ~ 10000 times faster than plain Dynamic-Programming-based local alignment
  - But suffer from low sensitivity, especially for distant sequences (e.g. *E.coli* → human)

Summary Questions

- Why do you need E-value?

- Could you give an alignment that the Smith-Waterman algorithm, but not the BLAST algorithm, can identify? Explain your case.

生物信息学：导论与方法
Bioinformatics: Introduction and Methods