Chapter 6

Database Searching For Similar Sequences

- To explain the differences between an alignment of two defined sequences and the search for homologs in databases
- To be able to use tools like Fasta and BLAST
- To understand the statistical significance of a score

- Dynamic programming algorithms
  - optimal local alignment (Smith and Waterman)
  - database searches take an hour to minutes with dedicated hardware (parallel architecture)
  - MPSearch (fast implementation of the S-W method)

- Hash coding algorithms
  - Introduction of heuristics
  - Gain of speed at the expense of some loss in sensitivity

    (50-100 times faster than the S-W method)
  - **BLAST** (Altschul *et al.*, NCBI) and **FASTA** (Pearson and Lipman, the University of Virginia)

Sequence alignment using lookup tables for common words

Seq 1 APNGTSCHQE Seq 2 GCHPLSAGQD

Position 1 2 3 4 5 6 7 8 9 10 1 234 5 6 7 8 9 10

Amino acid	Position in seq. 1	Position in seq. 2	Offset (1-2)
A	1	7	-6
P	2	4	-2
N	3	-	
G	4	1, 8	3, -4
T	5	-	
S	6	6	0
C	7	2	5
H	8	3	5
Q	9	9	0
E	10	-	
L	-	5	
D	-	10	

```
Position        1 2 3 4 5 6 7 8 9 10
Seq 1       ACNGTSCHQE
Seq 2         GCHCLSAGQD
Offset      +5
```

Common word or ktuple = 2
No gap allowed

Lookup (hash) table for sequence alignment

sequence	position							
	1	2	3	4	5	6	7	8
(*s*)	A	A	T	A	A	T	G	C
(*t*)	C	T	A	A	T	C	C	

Offset +1
AATAATGC
-CTAATCC

Offset +2
AATAATGC
--CTAATCC

Offset -2
--AATAATGC
CTAATCC

2-ktuples	Position in sequence *s*	
AA	1	4
AT	2	5
GC	7	
TA	3	
TG	6	

	1	2	3	4	5	6
2-ktup	CT	TA	AA	AT	TC	CC
offset+1		1	2	3	3	3
offset–2			1	2	2	2
etc						

Lookup (hash) table for sequence alignment (continued)

```
S    AATAATGC
     ||||  |
t    CTAATCC
```

	A	A	T	A	A	T	G	C	Matches on diag
C								O	
T			●			O			1
A	●	O		●	O				0
A	O	●		O	●				0
T			●			●			1
C								O	0
C								●	2
Matches on diag		0	0	0	1	3	1	1	**5**

Hashing and chaining

Imagine the following sequence *q*
(l = 22 and ktuple =2)

```
          10        20
TCGGATTCGT ACGGTACGGA TC
```

Here are the k-tuples

```
                                  10                                  20
TC,CG,GG,GA,AT,TT,TC,CG,GT,TA,AC,CG,GG,GT,TA,AC,CG,GG,GA,AT,TC
```

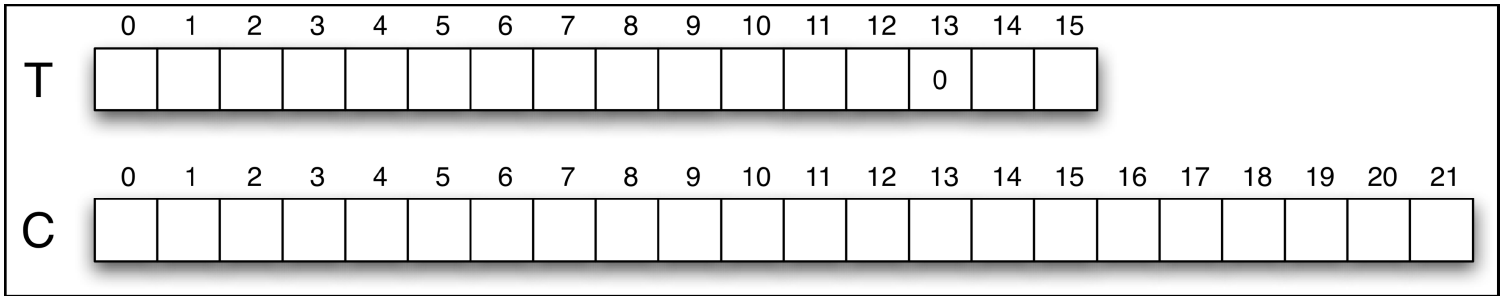
We chose a system allowing vectorial registration of the pairs

$Idx(ll) = v(ll_1)4^1 + v(ll_2)4^0$ v(*A*) = 0, v(*C*) = 1, v(*G*) = 2 and v(T) = 3

```
0                                   10                                  20
13,6,10,8,3,15,13,6,11,12,1,6,10,11,12,1,6,10,8,3,13
```

```
                    10              20
Seq q   TCGGATTCGT ACGGTACGGA TC
```

```
0                                   10                                  20
TC,CG,GG,GA,AT,TT,TC, CG,GT,TA, AC,CG,GG,GT, TA,AC,CG,GG,GA,AT,TC
13,6,10,8,3,15,13,6,11,12,1,6,10,11,12,1,6,10,8,3,13
```



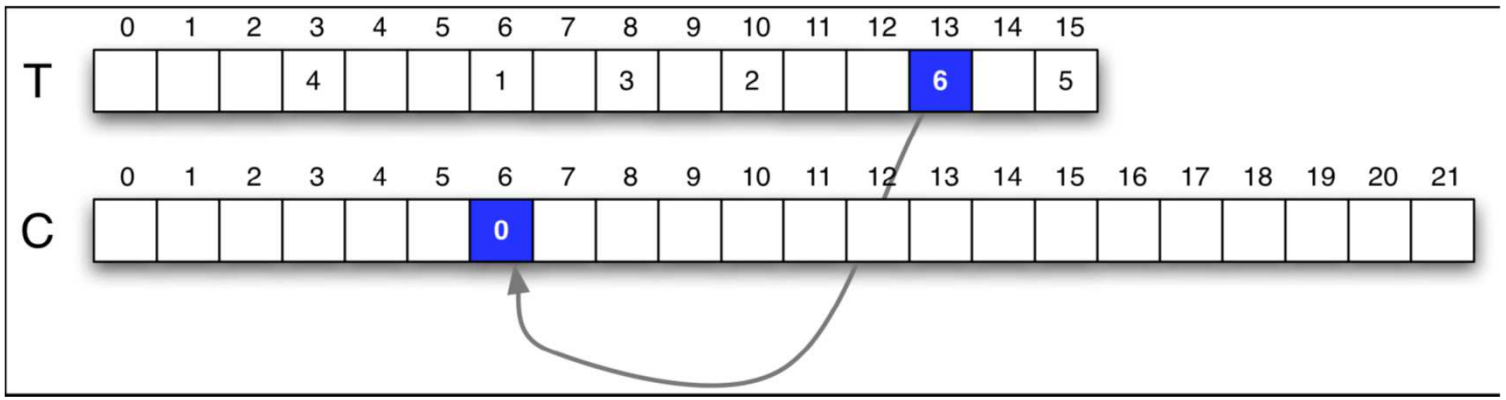
The first tuple TC start at the 0 position and has an index value of 13

0 10 20

TCGGATTCGT ACGGTACGGA TC

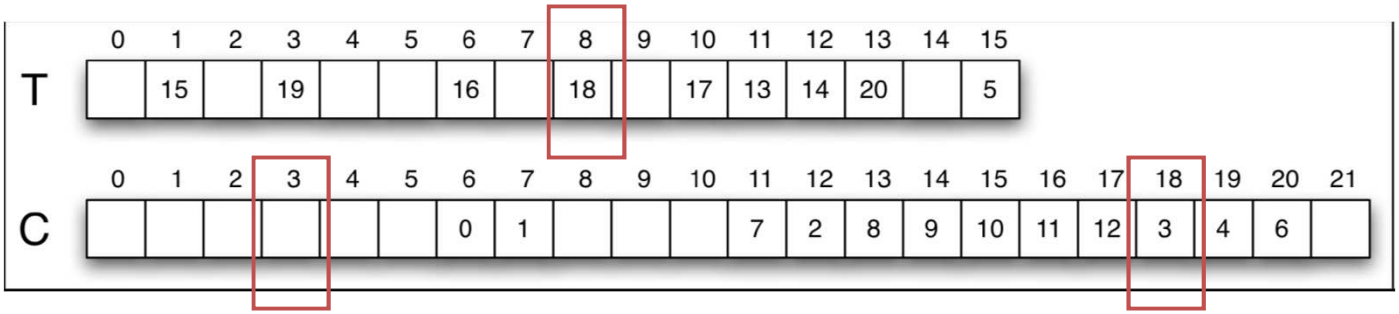
TC,CG,GG,GA,AT,TT,TC, CG,GT,TA, AC,CG,GG,GT, TA,AC,CG,GG,GA,AT,TC

13,6,10,8,3,15,13,6,11,12,1,6,10,11,12,1,6,10,8,3,13



Go on filling each cell;

If an index value already registered is met, move it to C and insert the new location in T



The table allows easy detection of repeated tuples and similar region

Seq *t* **GCTGATTCCA TGGGTAACCT GA**

0 10 20

GC,CT,TG,**GA**,**AT**,**TT**,**TC**,CC,CA,**AT**,TG,**GG**,**GG**,**GT**,**TA**,AA,**AC**,CC,CT,TG,**GA**

9, 7, 14 **8**, **3**, **15**,**13**, 5, 4,**3**, 14,**10**,**10**,**11**,**12**, 0, **1**, 5, 7,14, **8**

Position in seq *q*

3 4 5 0
18 19 6
20

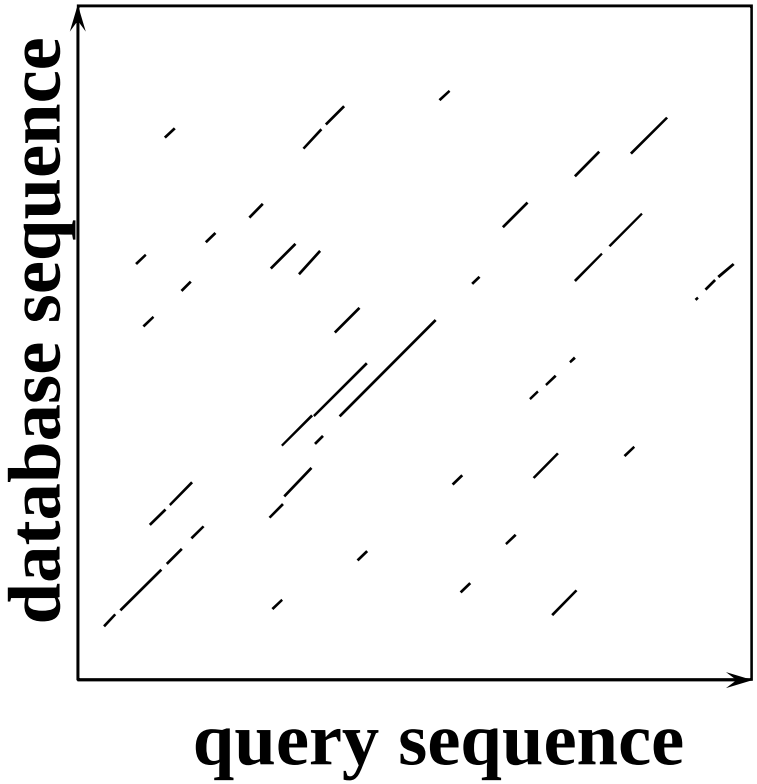
12 8 9
17 13 14

Offset = 0
Offset = 15

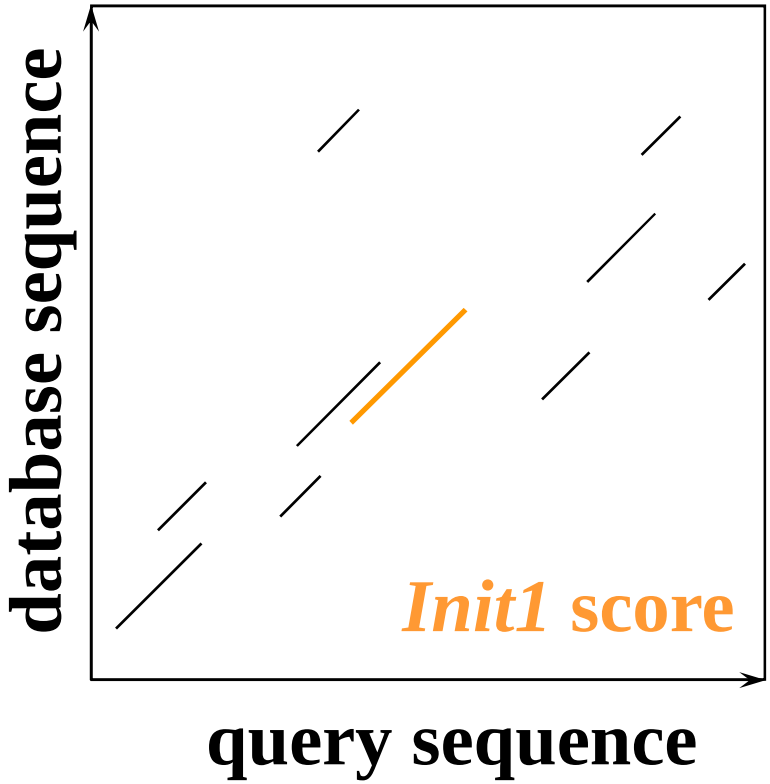
1. FASTA

Method

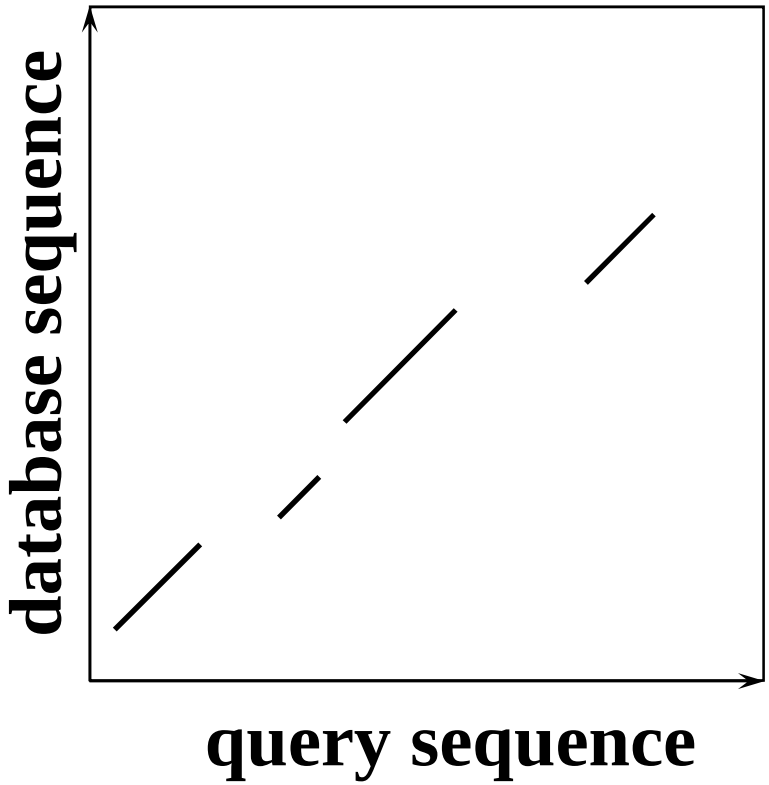
A. Search "k-tuples" common between query sequence and database sequence



B. The 10 best-matching regions are evaluated using a scoring matrix and gap penalty

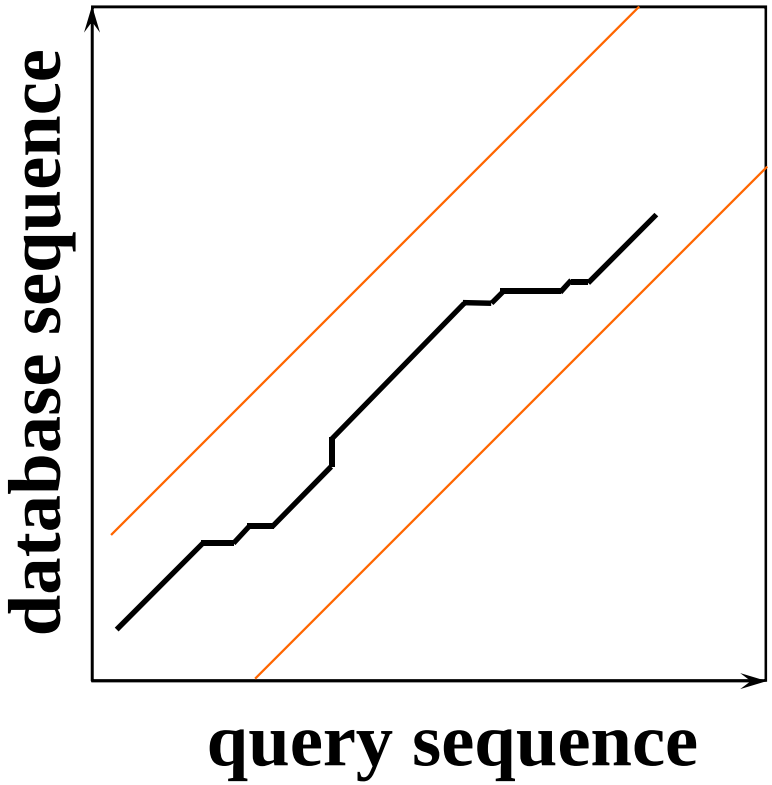


C. Several segments on different "diagonals" are joined into longer regions of score *initn*



***Initn* score**

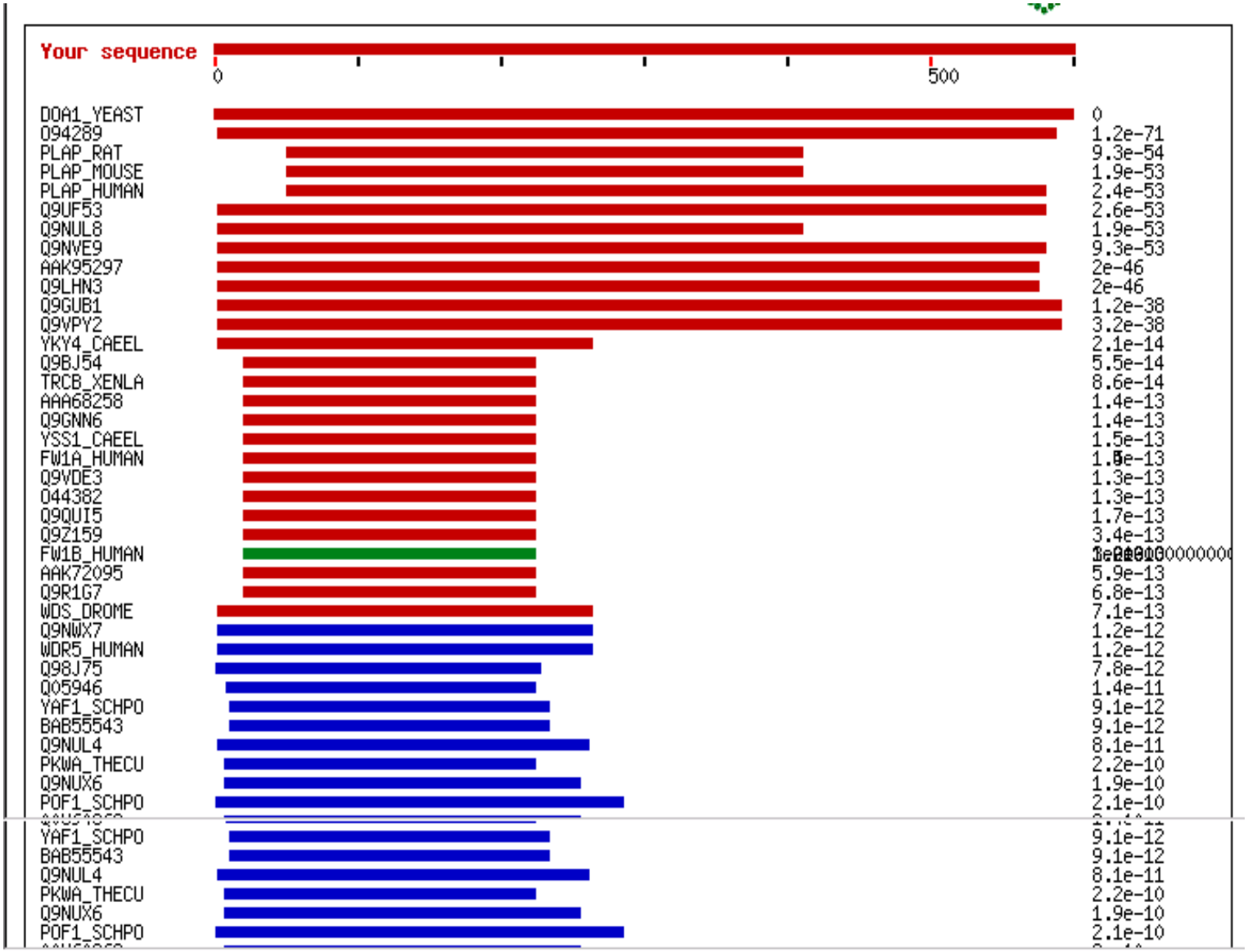
D. Best local alignment using the Smith-Waterman method within a restricted area (no more true in FASTA3)



***Opt* score**

Example of an FASTA output

a) Graphical overview of similarity



b) statistics

```
Searching /cbi/services/fasta/fastadb/swall library

        opt      E()
< 20    980      0:=
  22      5      0:=           one = represents 1168 library sequences
  24      9      1:*
  26     24     15:*
  28    133    157:*
  30    858    955:*
  32   4148   3695:===*
  34   9615  10019:========*
  36  19871  20577:==================*
  38  37382  34006:==============================*===
  40  54536  47436:=========================================*======
  42  64336  57984:==================================================*======
  44  70062  63962:========================================================*=====
  46  65798  65147:=========================================================*=
  48  61450  62371:==========================================================*
  50  53720  56913:=============================================  *
  52  46926  50036:======================================== *
  54  39252  42740:==================================   *
  56  33537  35701:============================= *
  58  26583  29310:========================  *
  60  22436  23743:====================*
  62  18193  19035:================*
  64  13693  15138:============*
  66  11084  11965:==========*
  68   8607   9411:========*
  70   6849   7375:=======*
  72   5554   5763:====*
  74   4197   4493:===*
  82   1429   1615:=*
  84   1065   1279:=*
  86    808    990:*
  88    585    766:*           inset = represents 19 library sequences
  90    541    593:*
  92    346    458:*          :====================      *
  94    338    355:*          :===================*
  96    254    274:*          :================*
  98    223    212:*          :============*
 100    179    164:*          :=========*=
 102    149    127:*          :=======*=
 104    143     98:*          :======*==
 106     73     76:*          :===*
 108     83     59:*          :===*=
 110     65     46:*          :==*=
 112     45     35:*          :=*=
 114     50     27:*          :=*=
 116     41     21:*          :=*=
 118     30     16:*          :*=
>120    929     13:*          :*=======================================
221373654 residues in 694597 sequences
 statistics extrapolated from 60000 to 693540 sequences
  Expectation_n fit: rho(ln(x))= 6.0577+/-0.000182; mu= 5.4114+/- 0.010
 mean_var=81.7127+/-16.658, 0's: 147 Z-trim: 55  B-trim: 2594 in 1/64
 Lambda= 0.1419
 Kolmogorov-Smirnov  statistic: 0.0329 (N=29) at  46
```

c) Comparison scores and sequence alignments

```
FASTA (3.39 May 2001) function [optimized, BL50 matrix (15:-5)] ktup: 2
 join: 38, opt: 26, gap-pen: -12/ -2, width:  16
 Scan time: 11.190
The best scores are:                                      opt bits E(693540)
SWALL:DOA1_YEAST P36037 DOA1 PROTEIN.              ( 715) 4314  893       0
SWALL:O94289 O94289 WD REPEAT-CONTAINING PROTEIN.  ( 713) 1287  274 1.2e-71
SWALL:PLAP_RAT P54319 PHOSPHOLIPASE A-2-ACTIVATIN  ( 647)  996  214 9.3e-54
SWALL:Q9NUL8 Q9NUL8 CDNA FLJ11281 FIS, CLONE PLAC  ( 544)  990  213 1.9e-53
SWALL:PLAP_MOUSE P27612 PHOSPHOLIPASE A-2-ACTIVAT  ( 646)  991  213 1.9e-53
SWALL:PLAP_HUMAN Q9Y263 PHOSPHOLIPASE A-2-ACTIVAT  ( 738)  990  213 2.4e-53
SWALL:Q9UF53 Q9UF53 HYPOTHETICAL 87.2 KDA PROTEIN  ( 795)  990  213 2.6e-53
SWALL:Q9NVE9 Q9NVE9 CDNA FLJ10780 FIS, CLONE NT2R  ( 795)  981  211 9.3e-53

>>SWALL:O94289 O94289 WD REPEAT-CONTAINING PROTEIN.        (713 aa)
 initn: 792 initl: 475 opt: 1287  Z-score: 1423.7  bits: 273.9 E(): 1
Smith-Waterman score: 1291;  37.202% identity (40.717% ungapped) in 6

               10        20        30        40        50
sp       MGYQLSATLKGHDQDVRDVVAVDDSKVASVSRDGTVRLWSK-DDQWLGTVVYTGQGFLN
           :.::  : :: :::: : ....  ..:::::::  .: . . .:       . .:::.:
SWALL: MTSYELSRELGGHKQDVRGVCSISNELIGSASRDGTYSVWEQINGEWTPHFYENHEGFVN
              10        20        30        40        50        60

       60        70        80        90       100       110
sp     SVCYDSEKELLLFGGKDTMINGVPLFATSGEDPLYTLIGHQGNVCSLS-FQDGVVISGSW
        :::    .    ::.:    :. :  .. ..: : :.::..:.:: : ... ..:.:::
SWALL: CVCYVPAIDKNSRGGQDKC--GI-LQEVGTNSPSYYLFGHESNICSASALNSETIITGSW
              70        80        90       100       110
```

2. Basic Local Alignment Search Tool (BLAST)

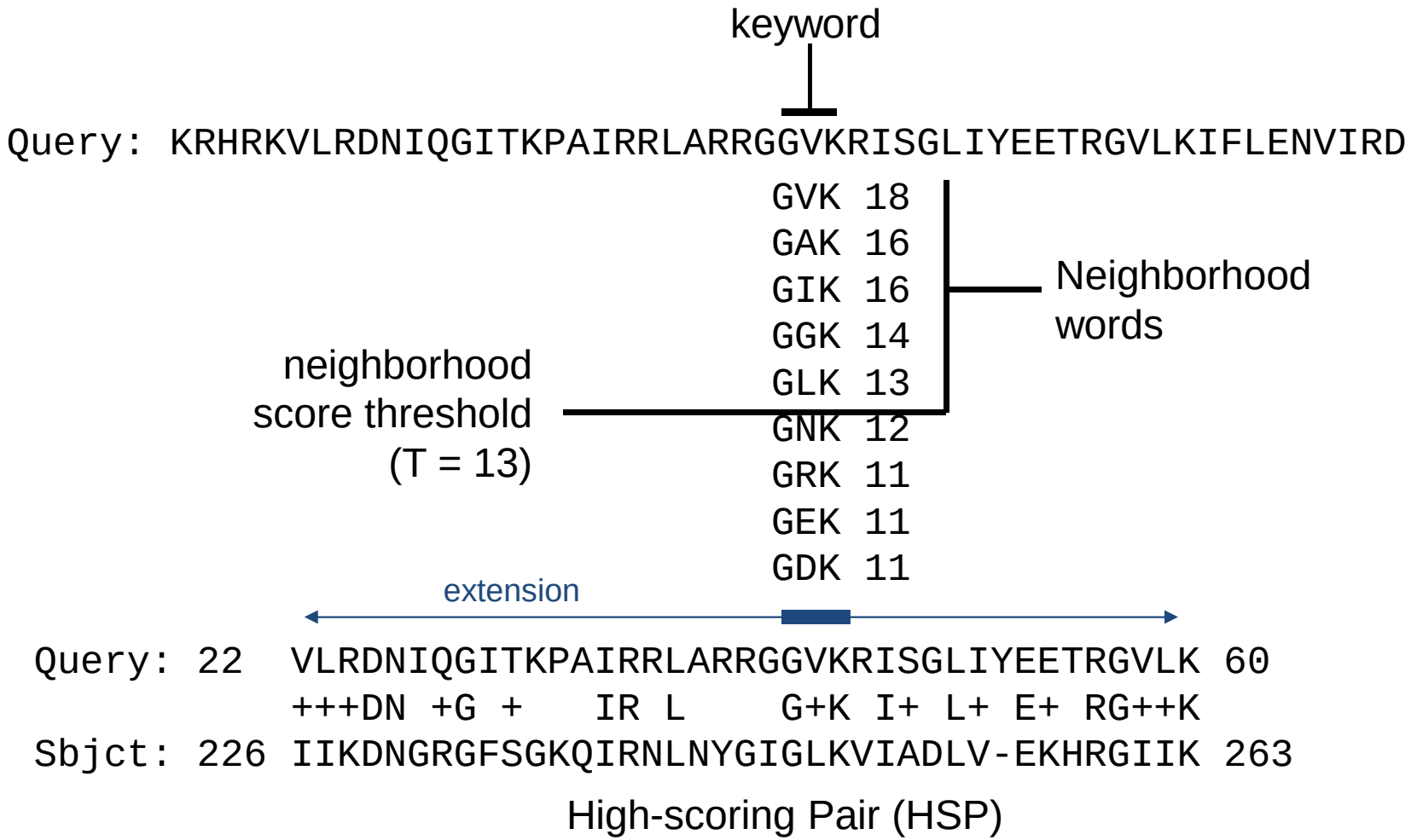
- Great improvement in speed, with a modest decrease in sensitivity
- Minimizes search space instead of exploring entire search space between two sequences
- Finds short exact matches ("seeds"), only explores locally around these "hits"

BLAST algorithm

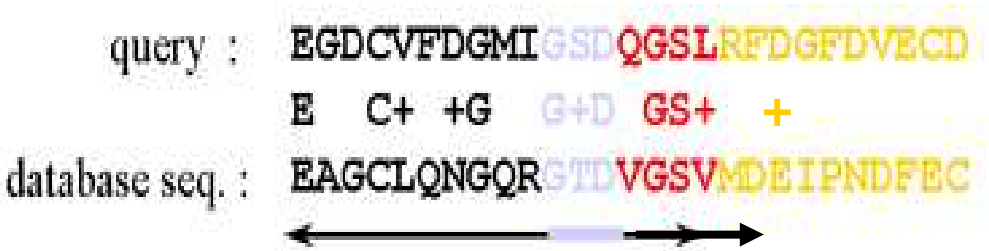
- Keyword search of all words of length *w* from the query of length *n* in database of length *m* with score above threshold
  - *w* = 11 for DNA queries, *w* =3 for proteins
  - Matches with any other combination of 3 amino acids are also evaluated using a scoring matrix to generate a list of neighbourhood words (cutoff score T)

- Local ungapped alignment extension for each found keyword
  - Extend result until longest match above threshold is achieved
- Running time *O*(*nm*)

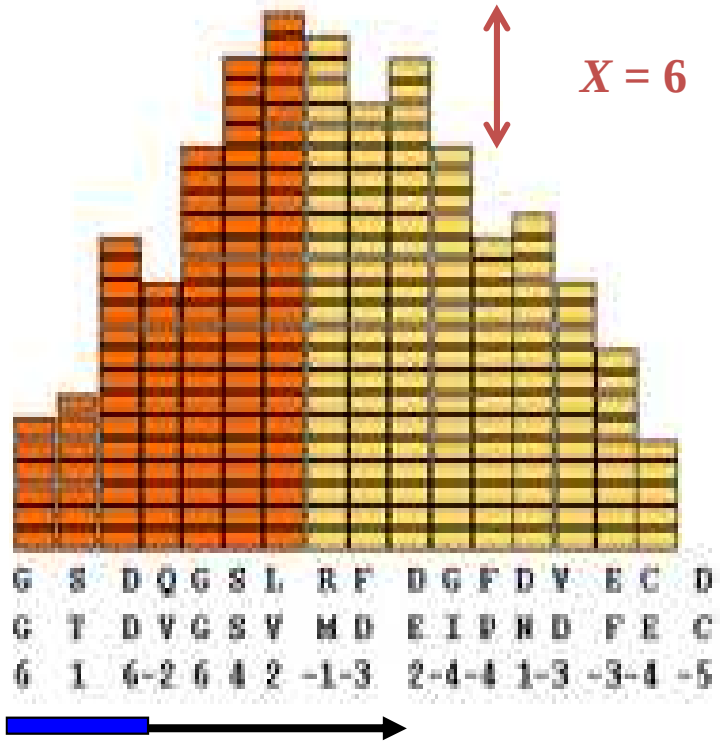
BLAST algorithm



Extension of a word into an High Scoring Segment Pair



Score = 13



The extension is terminated after the cumulative score drops off by 6 (*G/F*).

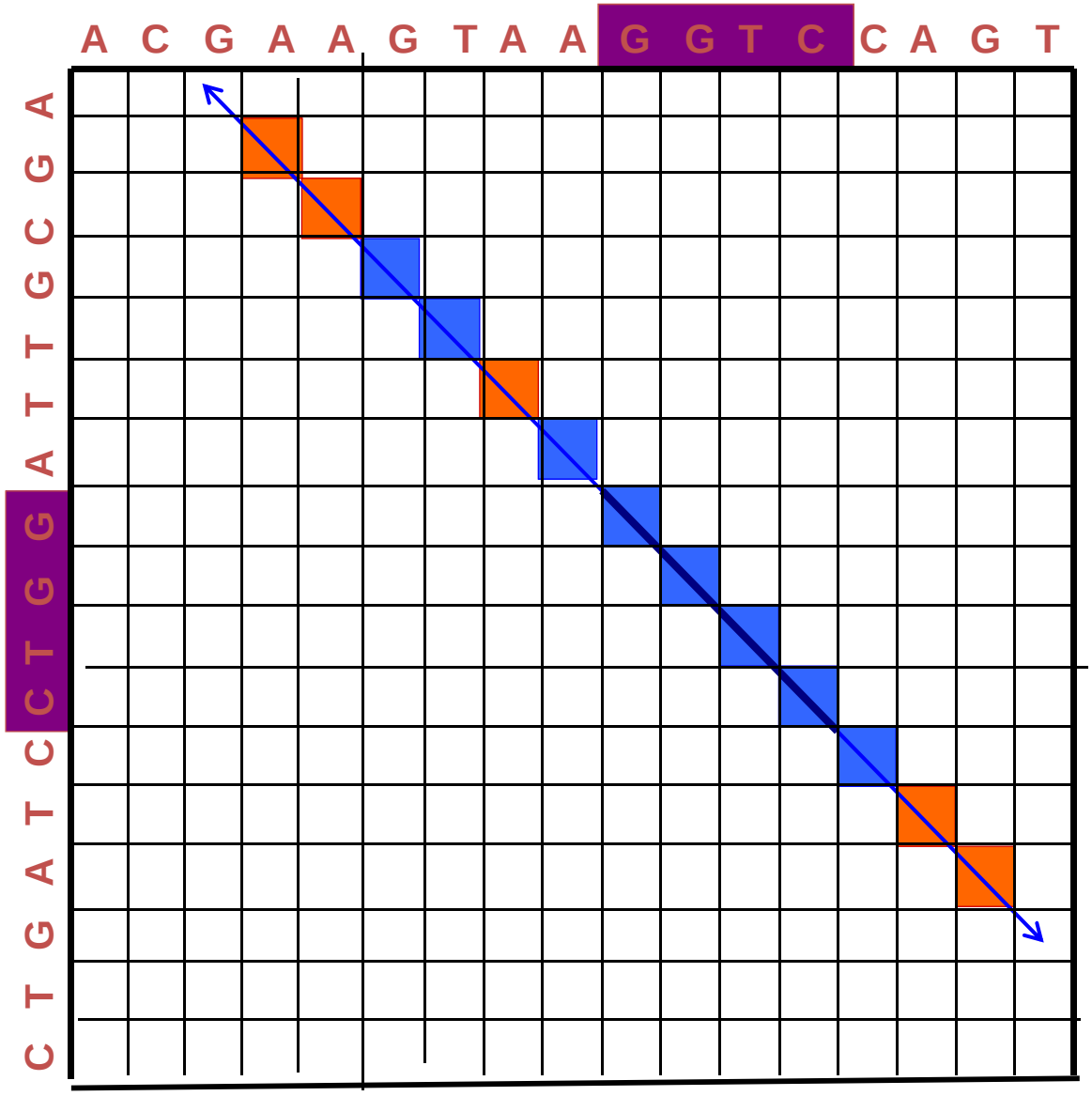
HSP score = 19 +13 + 8 = 40

Original BLAST: Example

- $w = 4$, $T = 4$
- Exact keyword match of **GGTC**
- Extend diagonals with mismatches until score is under 50%
- Output result

  GTAAGGTCC

  GTTAGGTCC



From lectures by Serafim Batzoglou (Stanford)

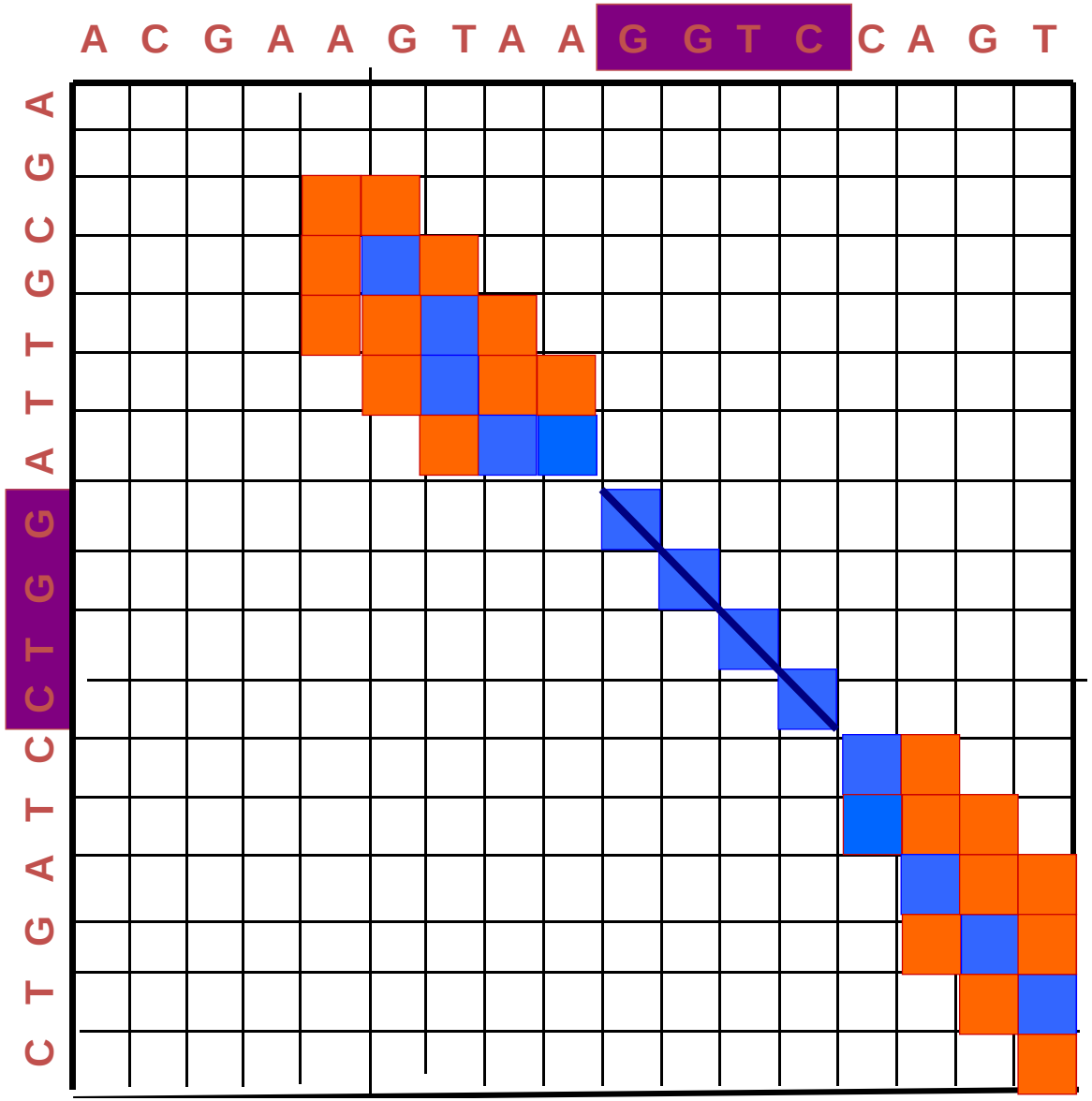
BLAST 2

- in the 1st version, the extension process in each direction was stopped when the accumulated score stopped increasing and had just begun to fall a small amount (X) below the best score.
- in BLAST2, only words lying on the same diagonal and within distance A of each other are joined and extended as described above.

BLAST2 (contd)

- Original BLAST exact keyword search, THEN:
- Extend with gaps around ends of exact match until score <*T*, then merge nearby alignments
- Output result

GTAAGGTCC–AGT
GTTAGGTCCTAGT

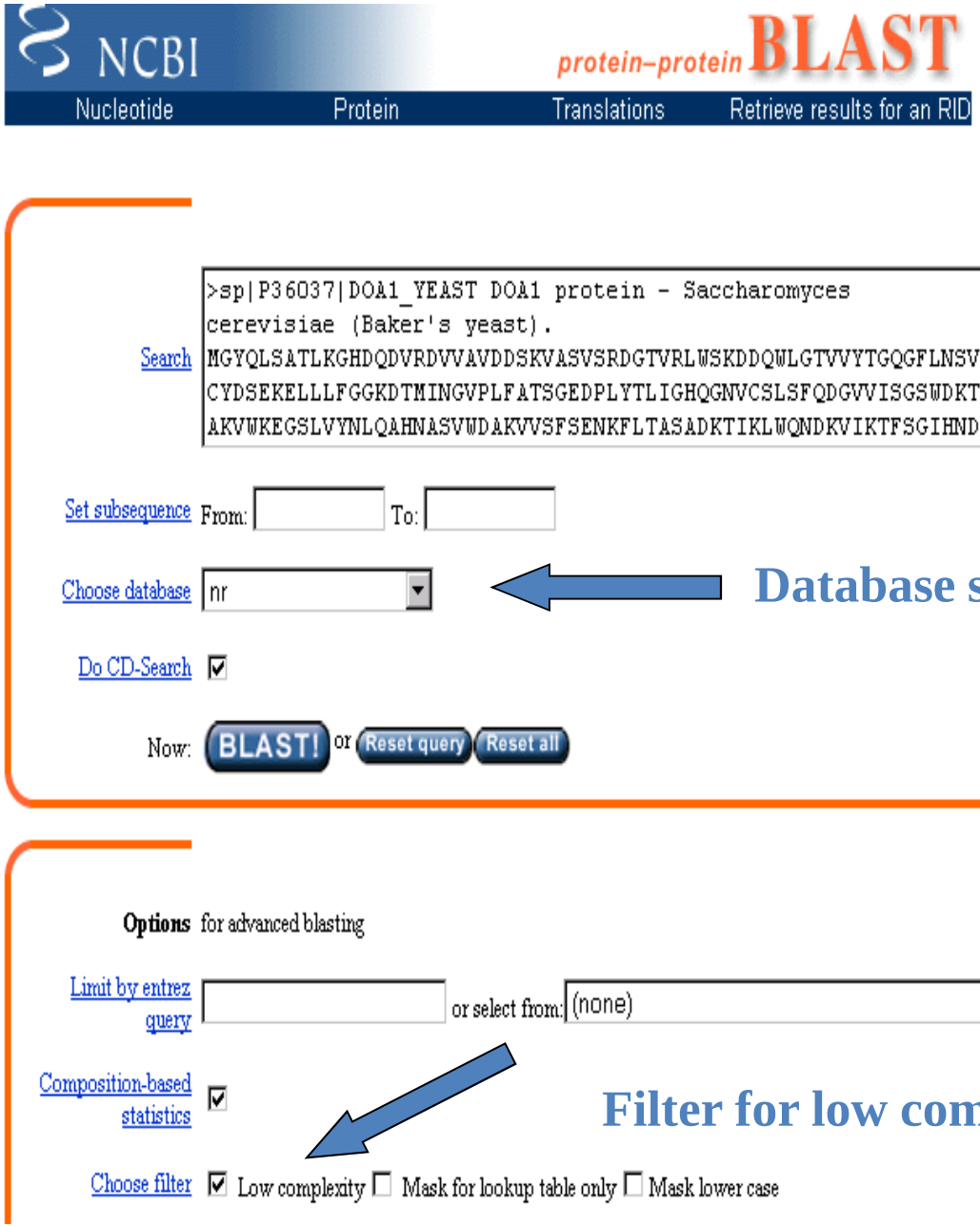


From lectures by Serafim Batzoglou (Stanford)

- For each HSP score greater than a cut-off score S, an optimal alignment with gaps is produced with the Smith-Waterman method. The score is obtained and the expect value (E) for that score is calculated

- The match is reported when the expect score satisfies the threshold parameter E

Example of a BLASTP database search



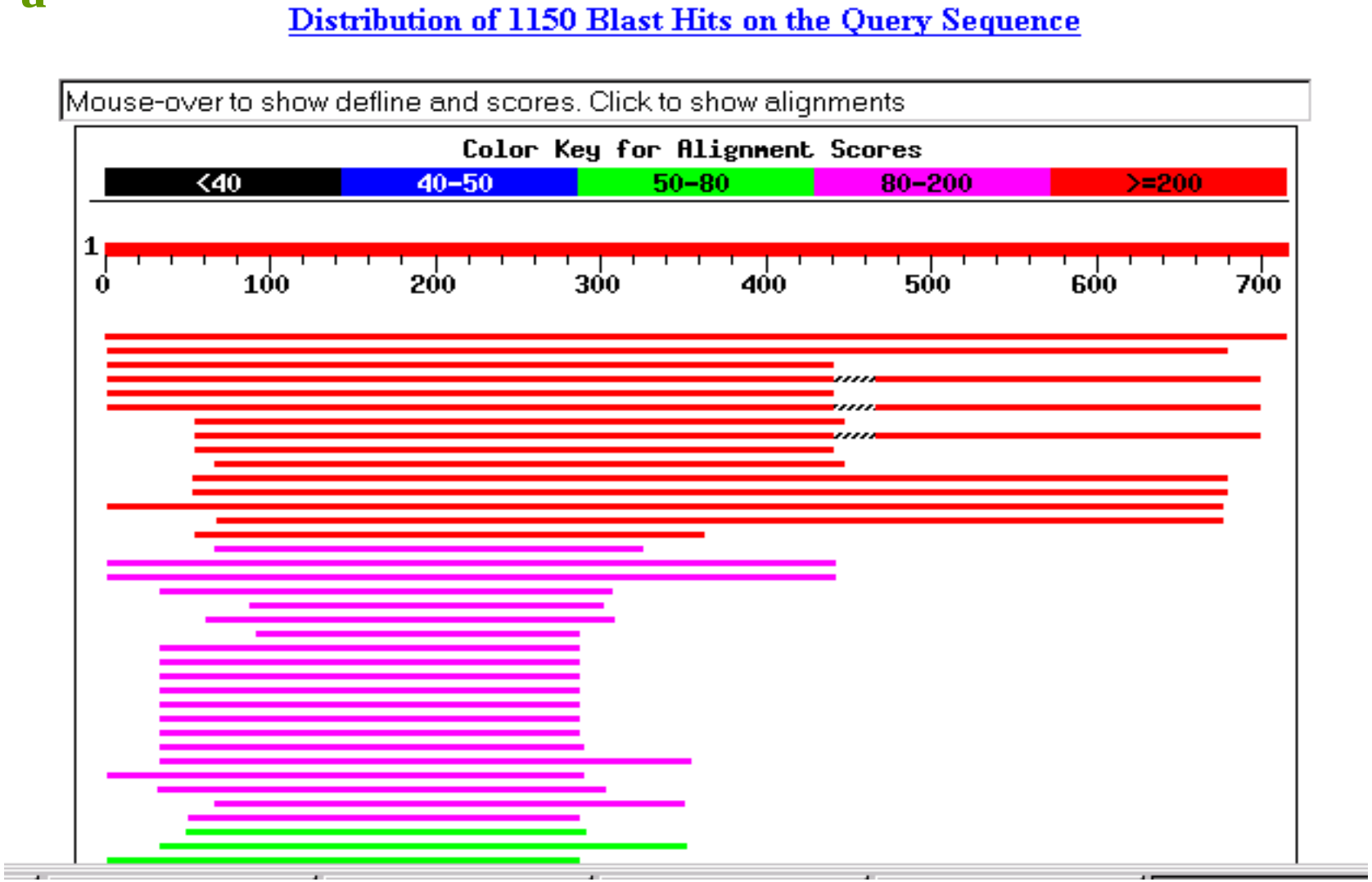
BLAST - Sequence Filtering

- Some regions of DNA and protein sequences consist of long repeated runs of a single residue or a pattern of residues.
- These repetitive sequences are not usually of interest to biologists, and close matches to these sequences may « mask out » lower-scoring matches to homologous sequences.
- By default, BLAST filters these sequences of low complexity, using the SEG program (protein).

HILCDEVNEGDEENEDFLPS
HILCXXXXXXXXXXXXFLPS

Example of BLASTP output (a and b)

a



b

Sequences producing significant alignments:		Score (bits)	E Value
gi|6322636|ref|NP_012709.1|	Required for normal intracellul...	1363	0.0
gi|7493716|pir||T40729	WD repeat-containing protein - fissi...	342	8e-93
gi|7023843|dbj|BAA92105.1|	(AK002143) unnamed protein produ...	298	2e-79
gi|14738164|ref|XP_035969.1|	phospholipase A2-activating pr...	297	2e-79
gi|14738161|ref|XP_035968.1|	phospholipase A2-activating pr...	297	2e-79
gi|7023020|dbj|BAA91803.1|	(AK001642) unnamed protein produ...	295	1e-78
gi|2507098|sp|P54319|PLAP_RAT	PHOSPHOLIPASE A-2-ACTIVATING ...	293	3e-78
gi|4758934|ref|NP_004244.1|	phospholipase A2-activating pro...	293	5e-78
gi|5326866|gb|AAD42075.1|AF145020_1	(AF145020) phospholipas...	290	3e-77
gi|2507097|sp|P27612|PLAP_MOUSE	PHOSPHOLIPASE A-2-ACTIVATIN...	290	3e-77

Alignments

>gi|6322636|ref|NP_012709.1| Required for normal intracellular ubiquitin metabolism and for normal rates of proteolysis of ubiquitin-dependent proteolytic substrates in vivo; Doa1p [Saccharomyces cerevisiae]
gi|549752|sp|P36037|DOA1_YEAST DOA1 PROTEIN
gi|539311|pir||S38051 DOA1 protein - yeast (Saccharomyces cerevisiae)
gi|473137|emb|CAA53560.1| (X75951) ORF6, F715 [Saccharomyces cerevisiae]
gi|486381|emb|CAA82058.1| (Z28213) ORF YKL213c [Saccharomyces cerevisiae]
gi|1086570|gb|AAA82258.1| (U39947) Doa1p [Saccharomyces cerevisiae]
Length = 715

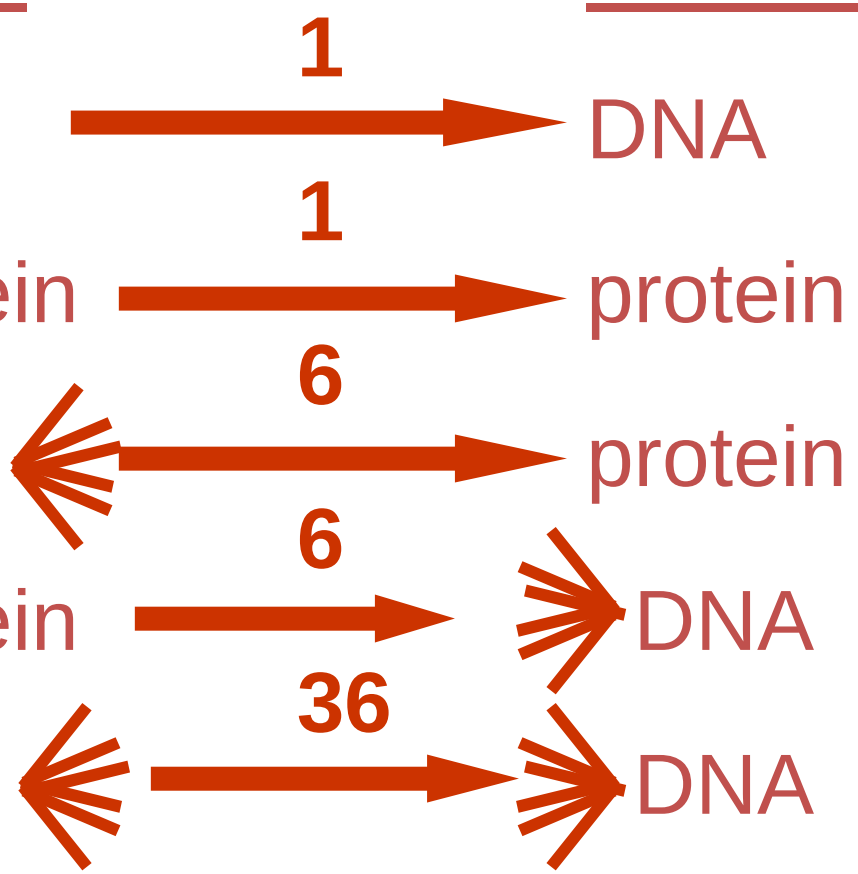
Score = 1363 bits (3529), Expect = 0.0
Identities = 681/715 (95%), Positives = 681/715 (95%)

```
Query: 1   MGYQLSATLKGHXXXXXXXXXXXXXXXXXXXXXXGTVRLWSKDDQWLGTVVYTGQGFLNSV 60
           MGYQLSATLKGH                       GTVRLWSKDDQWLGTVVYTGQGFLNSV
Sbjct: 1   MGYQLSATLKGHDQDVRDVVAVDDSKVASVSRDGTVRLWSKDDQWLGTVVYTGQGFLNSV 60

Query: 61  CYDSEKELLLFGGKDTMINGVPLFATSGEDPLYTLIGHQGNVCSLSFQDGVVISGSWDKT 120
           CYDSEKELLLFGGKDTMINGVPLFATSGEDPLYTLIGHQGNVCSLSFQDGVVISGSWDKT
Sbjct: 61  CYDSEKELLLFGGKDTMINGVPLFATSGEDPLYTLIGHQGNVCSLSFQDGVVISGSWDKT 120
```

Choose a BLAST program

Program	Input		Database
blastn	DNA	1	DNA
blastp	protein	1	protein
blastx	DNA	6	protein
tblastn	protein	6	DNA
tblastx	DNA	36	DNA



3. Statistics of sequence similarity scores

- Given an alignment score, how strong is the similarity it represents?
- What is the probability of having such a high score with unrelated sequences?
- What is the expecter number of alignments having such a high score with unrelated sequences from a DB search?

Modelling a random DNA sequence alignment

- all sequences have the same length
- no deletion or insertion

```
A C C G T T A G G G
A C T T T T A G G A
* *     * * * * *
```

Number of matching sites (m) = 7

Size of the sequences (N) = 10

The probability of an identical match is : 4 x 1/4 x 1/4 = 0.25 (equal frequency)

The probability that there are *m* matching sites is given by a binomial distribution

$$P(m) = C_m^N a^m (1-a)^{N-m}$$

$$\mu = N\,a$$

$$\sigma^2 = N\,a\,(1-a)$$

- **Approximation of binomial distribution by normal distribution**

$$Y(m) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\left[-\frac{(m-\mu)^2}{2\sigma^2}\right]}$$

if $z = (m_{obs} - \mu) / \sigma$

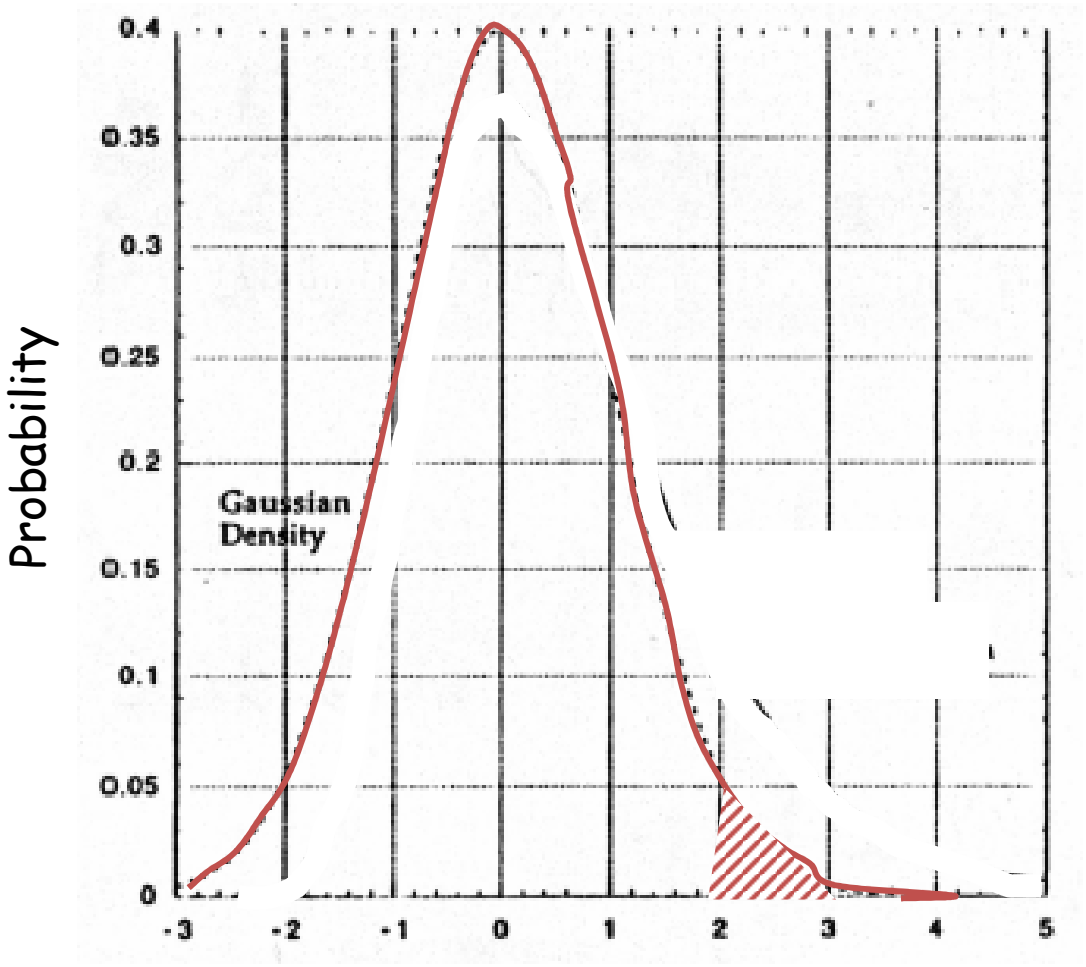
then

$$Y(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

$\mu = 0$ and $\sigma^2 = 1$

For a z-value of 2, probability is 0.054

Standard normal distribution



p(z > 2) = 0.023

P (0≤Z≤2) = 0,4772 as given in the Table

Suppose we observe 120 base matches in a sequence of length 400 (N)

- $\mu = 0.25\ N$ *and* $\sigma^2 = 0.1875N$
- $z = (120 - 100)/\sqrt{75} = 2.309$

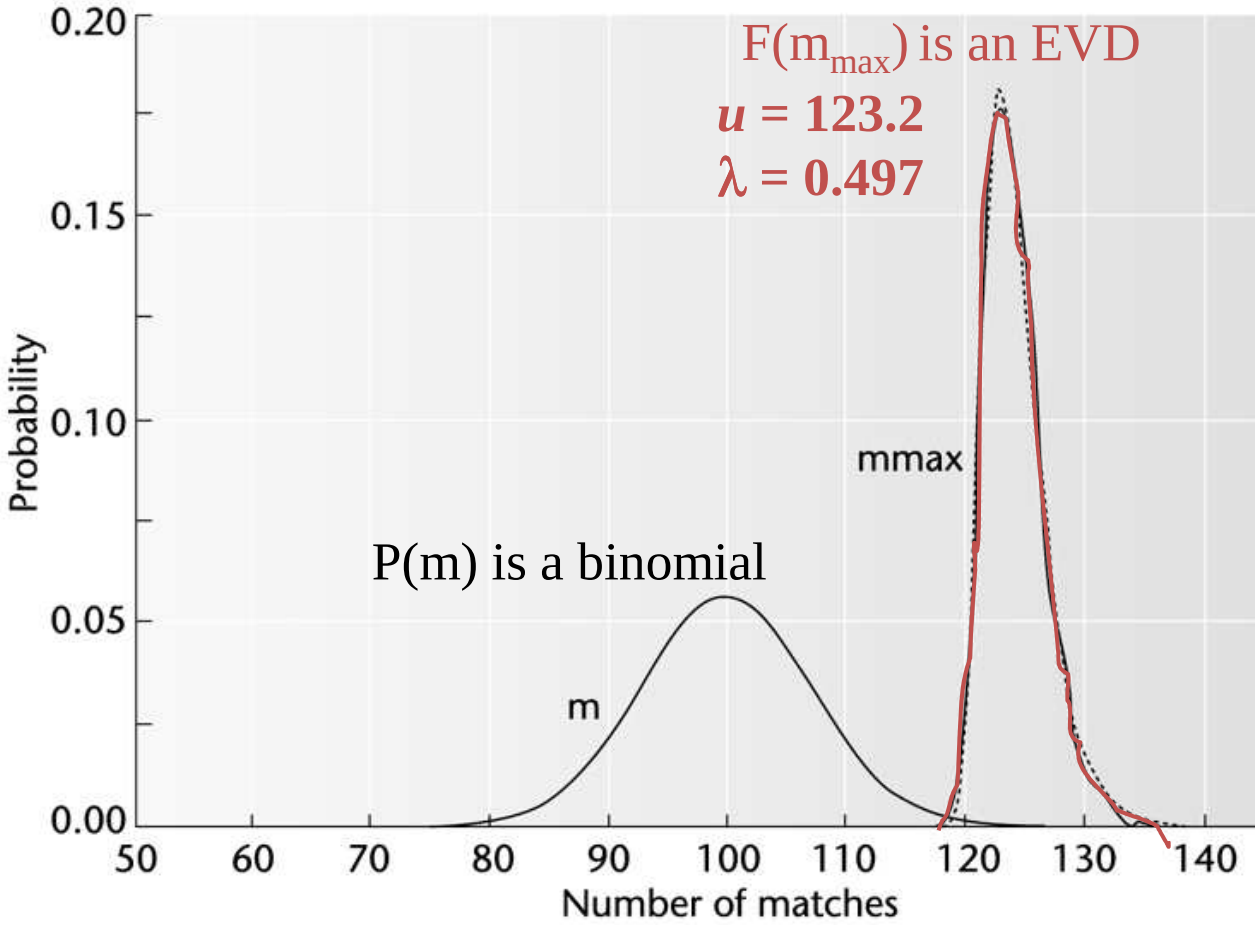
- *For z of 2.309, p (z>2.31) = 0.0105*

- *Significant match even though the percentage identity is quite low (30%)*

Note that Pearson and Lipman recommend a *z* value of > 10 to be significant

Simulation shows that probability distribution for DB search scores follows an Extreme Value Distribution (EVD)

- Generate random 2000 sequences (G+C) with a length 200 nucleotides (N)
- Calculate m_{max} (score of the most closely matching sequence from the DB) and m (match between pairs of sequences starting at the same ordinate)



We are taking the maximum score of a large number of alignments !

A score of130 matches is just big enough to be significant according to EVD as p(130)= 0.034 Compare with p(Z=4.2) <0,0001 if a (wrong) Normal distribution was considered

- **Extreme value distribution**

$$F(S_{max}) = \lambda\, e^{-\lambda(S_{max}-u)} \exp\left(-e^{-\lambda(S_{max}-u)}\right)$$

$$p(S \geq s_{obs}) = 1 - \exp\left(-e^{-\lambda(s_{obs}-u)}\right)$$

If $S' = \lambda (S - u)$

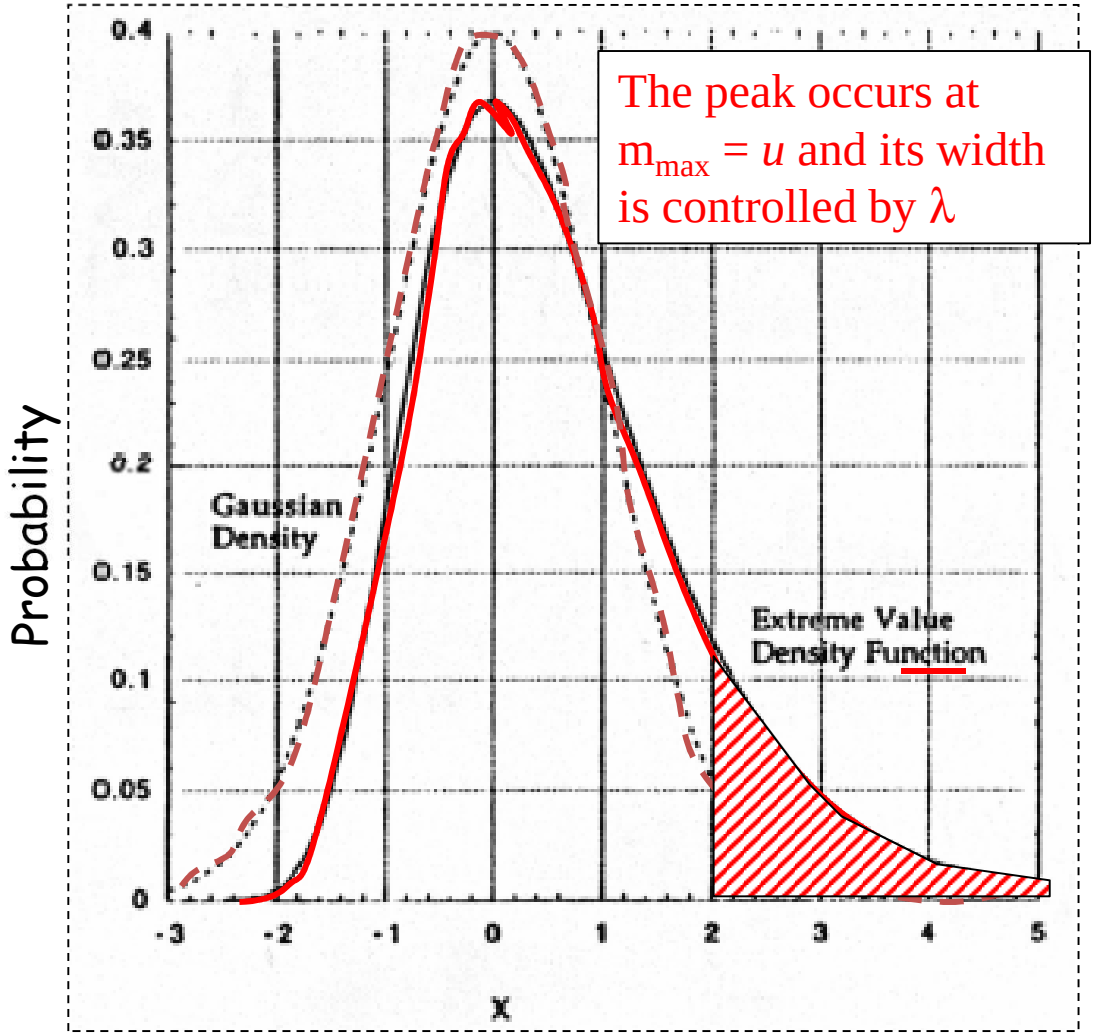
$$F_{S'} = \exp\left[-s' - e^{-s'}\right]$$

$u = 0$ and $\lambda = 1$

$$p(S' \geq s'_{obs}) = 1 - \exp[-e^{-s'_{obs}}]$$

For a S'= 2
p(S'≥ 2) = 0.13

To be compared with 0.023
if Normal distribution



Note that $P(S'>x) \cong e^{-x}$ for high values of x

x	$1 - exp(-e^{-x})$	e^{-x}
0	0.63	1
1	0.308	0.368
2	0.127	0.135
3	0.0486	0.0498
4	0.0181	0.0183

FASTA Statistics and scores

Fasta scores are expressed as normalized Z scores

$$Z = 50 + 10z \qquad \text{with } z = (S - \mu)/\sigma$$

The λ and u parameters of the EVD are expressed in terms of μ and σ

$$\lambda = 1.2825/\sigma \text{ and } u = \mu - 0.4500\sigma$$

The following equation

$$p(S \geq x) = 1 - \exp[-e^{-\lambda(S-u)}]$$ is modified in :

$$p(z) \simeq 1 - exp\,[-e^{(-1.2825\,z - 0.5772)}]$$

$$E(z) \simeq p(Z \geq z) \cdot D$$ with D: number of library sequences

BLAST statistics

- Significance of a score is expressed as the E(xpectation) value, the number of alignments between your sequence and randomly chosen sequences giving a score as good as the one observed

$$E(S>x) = KMN\, e^{-\lambda x}$$

Where M and *N* are effective lengths of the query and database sequences; their product is the effective search space

K and λ are parameters determined from the EVD

$$E(S_{bits}) = MN/2^{S_{bits}} \quad \text{with} \quad S_{bits} = \frac{\lambda S - \ln K}{\ln 2}$$

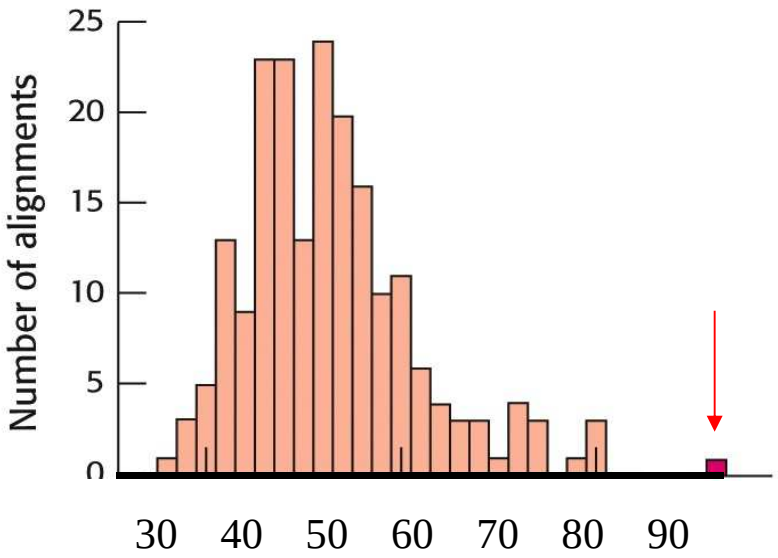
Bit scores allow you to compare results between different database searches, even using different scoring matrices.

PRSS Statistics for pairwise comparison

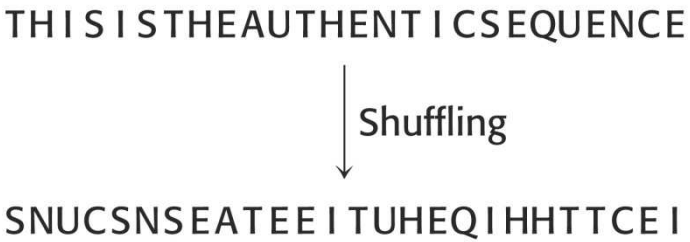
S-W score = 97

Alignment between human myoglobin and plant leghemoglobin gave a S-W score of 97 (BLOSUM50, GOP=12, GEP= 2)

Statistics of pairwise global alignments



Distribution of scores for shuffled sequences (PRSS)



From the random score distribution calculate the probability that a score reflects a true alignment (BLOSUM 50, -10,-2; K=0.03052, λ = 0.1750, MN = 154 x 154000; 1000 permutations)

$S_b = (\lambda S - \ln K)/\ln 2$

$S_b = (97 \times 0{,}1750 - \ln 0.03052)/\ln 2 = 29.5$

$E(S_b > x_b) = MN2^{-xb} = MN/2^{xb}$

$= 154 \times 154000/2^{29.5} = 0.031$

Statistics of coin tosses - expectation

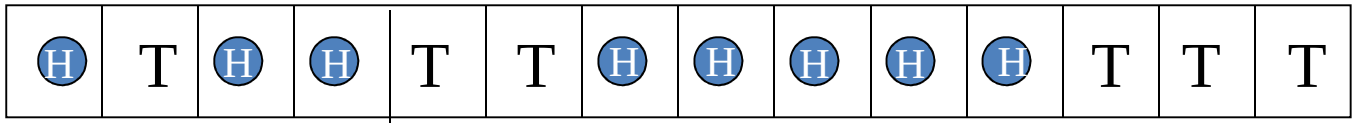
- $p(\mathrm{H}) = p(\mathrm{T}) = 0.5$
- $p(\mathrm{HHHTH}) = p(\mathrm{HTTTH}) = p(\mathrm{HHHHH}) = (1/2)^5$
- What is the expectation of the number of heads, E(H), in 5 flips?
  - p (HHHHH) = p(TTTTT) = 1/32
  - p (HHHHT) = p (TTTTH) = 5/32
  - p (HHHTT) = p (TTTHH) = 10/32 ($C_3^5 = \frac{5!}{3! \times 2!}$ possibilities out of 32)

- E (H) = (5 + 0) x 1/32 + (4 + 1) x 5/32 + (3 + 2) x 10/32 = 2.5

Expected value (or **mathematical expectation**, or **mean**) = the sum of the probability of each possible outcome of the experiment multiplied by the outcome value

Note that E(H) = p(H) x N
= 0.5 x 5

What is the expected number of times that 5 heads in a row occur by chance in 14 flips?



- The probability of 5 heads in a row is :

$$p(5) = (1/2)^5 = 1/32$$

- But since there were 10 places that one could have obtained 5 heads in a row, the expected number of times that 5 heads in a row occur by chance is :

$$E(5) \cong 10 \times 1/32 = 0.31$$

$$E(x) \cong p(x) \cdot N \quad \text{for } N \gg x$$

The expected length of the longuest run Rn increases as log of number of flips

- E(# of H of length m) ~ $n\ p^m$
- if the longest run has to be seen once,

$$1 \cong n\ p^{Rn}$$

$$1/n \cong p^{Rn}$$

$$-\log_e(n) \cong Rn \log_e(p)$$

$$-\log_e(n)/\log_e(p) \cong Rn$$

$$Rn \cong \log_{(1/p)}(n)$$

If $n = 100$ tosses, then $R \approx \log_2 100 = 6.65$

The expected length of the longest match for aligned DNA sequences of 100 nucleotides (p = 0.25) is :

$$Rn \approx 2 \times \log_4 100 = 6.65$$

From Head runs to alignment scores

The expected number of matching words E(l) between 2 sequences (lengths n and m) is equivalent to the expected number of heads in the coin-tossing sequence

- $P(l \geq m) = p^m$ with p = probability that 2 bases match

If $\lambda = -\ln(p) = \ln(1/p)$

- $P(l \geq m) = e^{-\lambda m}$
- $E(l \geq m) = KMN\, e^{-\lambda m}$

n ways of choosing the first starting site and *m* ways of choosing the second starting site

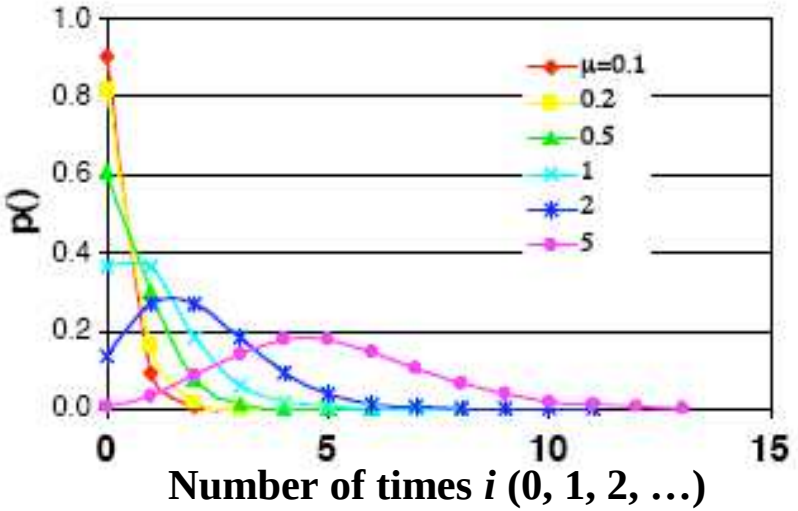
K accounts for overlap between words starting at neigboring points (<1)

It is more useful to use alignment scores instead of lengths for comparing sequences

- $E(S \geq x) = KMN\, e^{-\lambda x}$

- The probability of having an alignment with a score higher than the expected one is predicted by the Poisson distribution where the mean μ is given by E(S)

Poisson distribution for different mean (μ) values



$$P(\mu,i)=(\mu^i e^{-\mu})/i!$$

$$p(i > 0) = 1 - p(0) = 1 - \mu^0 e^{-\mu}/0! = 1 - e^{-\mu}$$

$$P(S) = 1-\exp(-KMNe^{-\lambda S})$$

- The number of alignments in a database that exceeds the mean score E(S >x) = P.D (see FASTA)