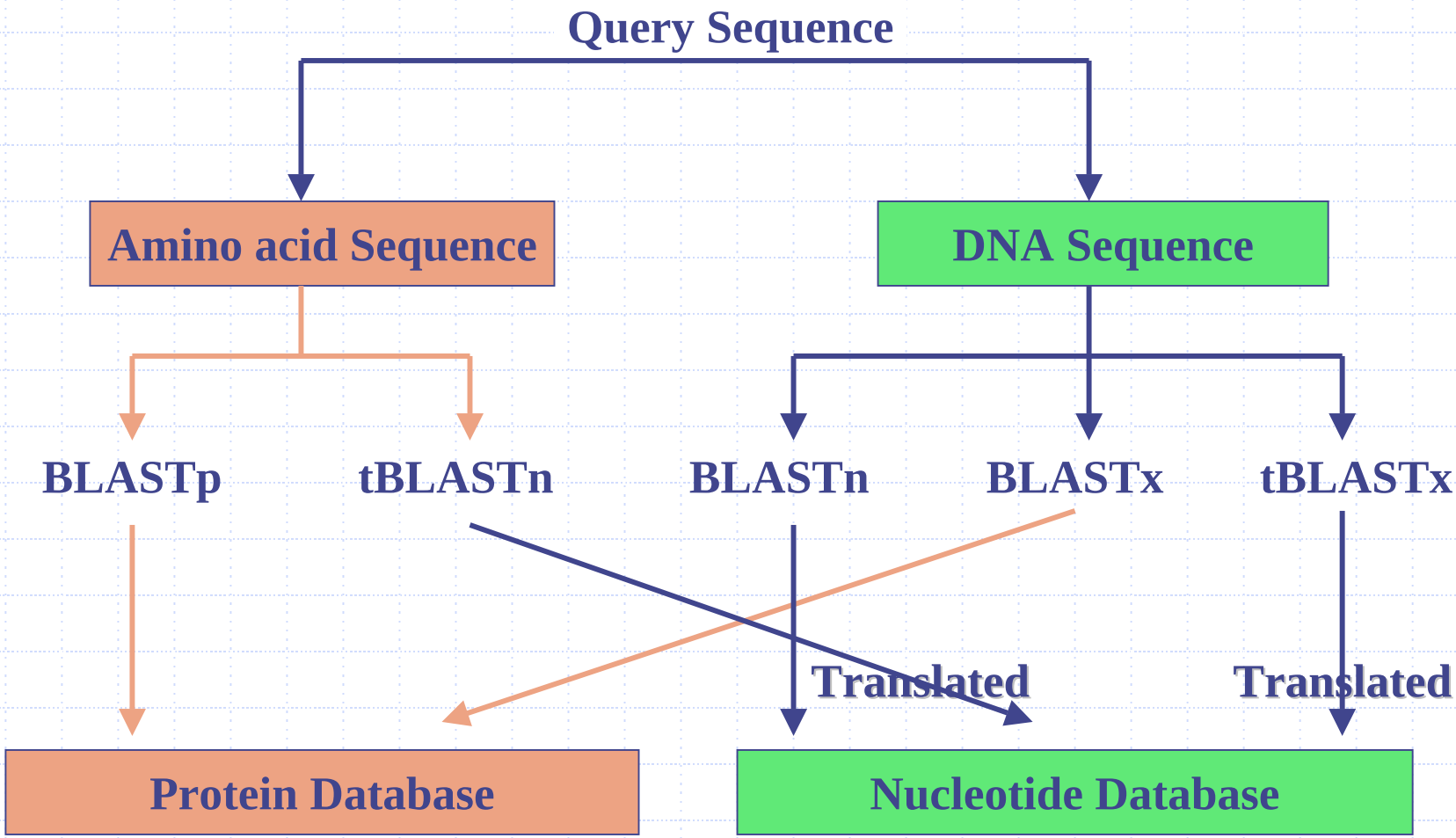
Blast BLAST

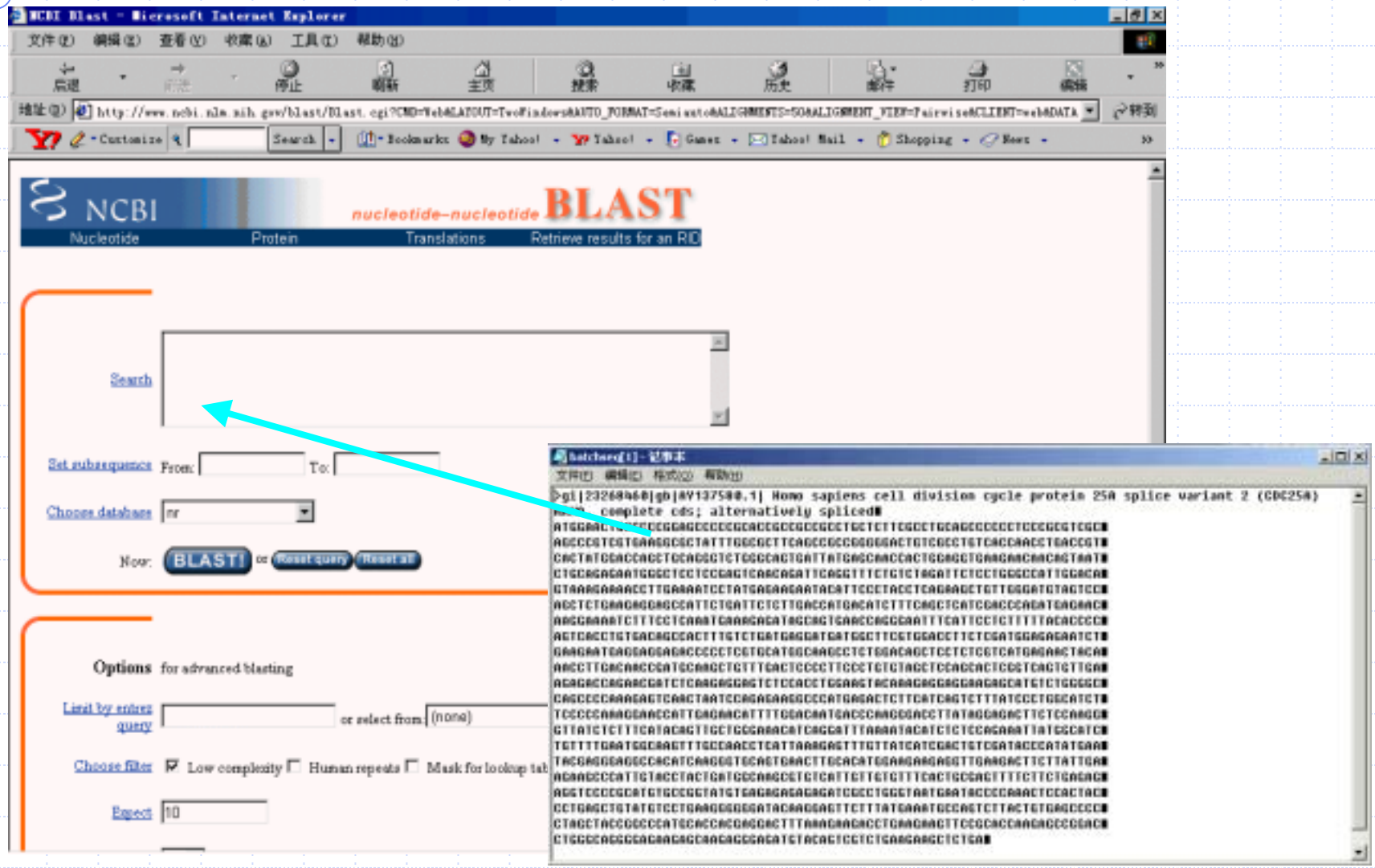
1. Read BLAST results
2. Dynamic programming
3. Matrix
4. Entropy (H)
5. Score and E-value
6. λ and κ

Topic 2

1. BLAST*all*


Query Sequence
Amino acid Sequence
DNA Sequence
BLASTp
tBLASTn
BLASTn
BLASTx
tBLASTx
Translated
Translated
Protein Database
Nucleotide Database

Web-based BLAST Input



Web-based BLASTN Output

BLASTN 2.2.4 [Aug-26-2002]

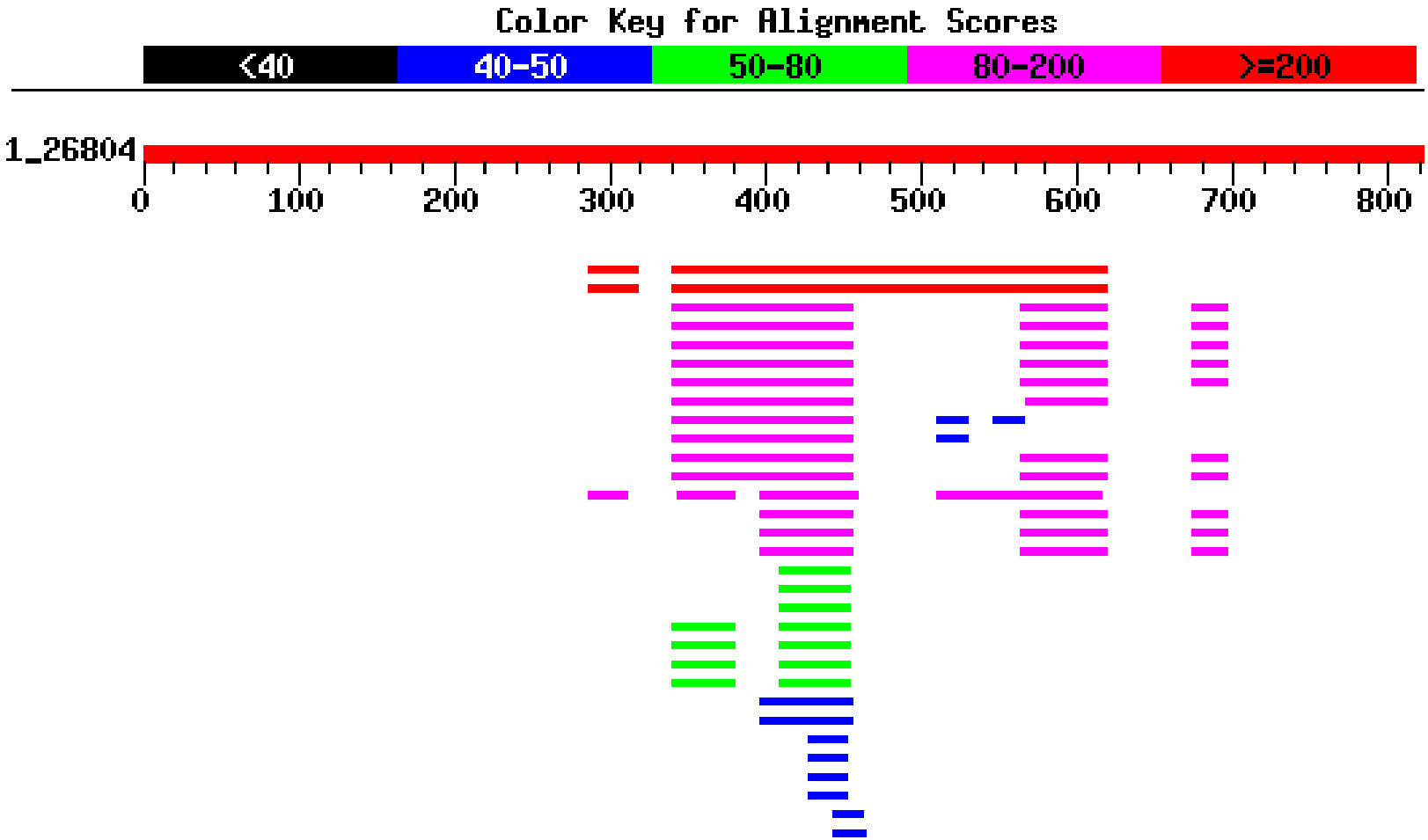
Reference:
Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

RID: 1031981000-015047-25897

Query= cluster8190_1
(1447 letters)

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences)
1,367,736 sequences; 6,442,224,816 total letters

Web-based BLAST Output



Web-based BLAST Output

Sequences producing significant alignments:	Score (bits)	E Value
gi|21314556|gb|AF513016.1| Sus scrofa myosin regulatory lig...	654	0.0
gi|21261723|emb|AJ487671.1|SSC487671 Sus scrofa mRNA for my...	654	0.0
gi|17226389|gb|AF440218.1|AF440218 Canis familiaris ventric...	521	e-145
gi|22061915|ref|XM_027060.4| Homo sapiens myosin, light pol...	420	e-114
gi|21411328|gb|BC031006.1| Homo sapiens, myosin, light poly...	420	e-114
gi|21410232|gb|BC031008.1| Homo sapiens, myosin, light poly...	420	e-114
gi|16198354|gb|BC015821.1|BC015821 Homo sapiens, myosin, li...	420	e-114
gi|516580|gb|S69022.1|S69022 Homo sapiens myosin light chai...	420	e-114
gi|4557774|ref|NM_000432.1| Homo sapiens myosin, light poly...	404	e-109
gi|34845|emb|X66141.1|HSMYLC2 H.sapiens mRNA for cardiac ve...	404	e-109
gi|2460246|gb|AF020768.1|AF020768 Homo sapiens cardiac vent...	400	e-108
gi|1220300|gb|M22815.1|HUMCMLC Human (clone PWHCLC2-8) card...	385	e-103
gi|34686|emb|X14332.1|HSMLC2 Human mRNA for ventricular myo...	371	2e-99
gi|56682|emb|X07314.1|RNMLC2R Rat heart myosin light chain ...	264	3e-67
gi|20841890|ref|XM_132324.1| Mus musculus myosin light chai...	262	1e-66
gi|6754777|ref|NM_010861.1| Mus musculus myosin light chain...	262	1e-66
gi|12832295|dbj|AK002367.1| Mus musculus adult male kidney ...	262	1e-66
gi|199984|gb|M91602.1|MUSMYLTCH M.musculus myosin light cha...	262	1e-66

Web-based BLAST Output

>gi|21314556|gb|AF513016.1| Sus scrofa myosin regulatory light chain ventricular isoform (MLC-2V) mRNA, complete cds
Length = 672

Score = 654 bits (330), Expect = 0.0
Identities = 359/370 (97%)
Strand = Plus / Plus

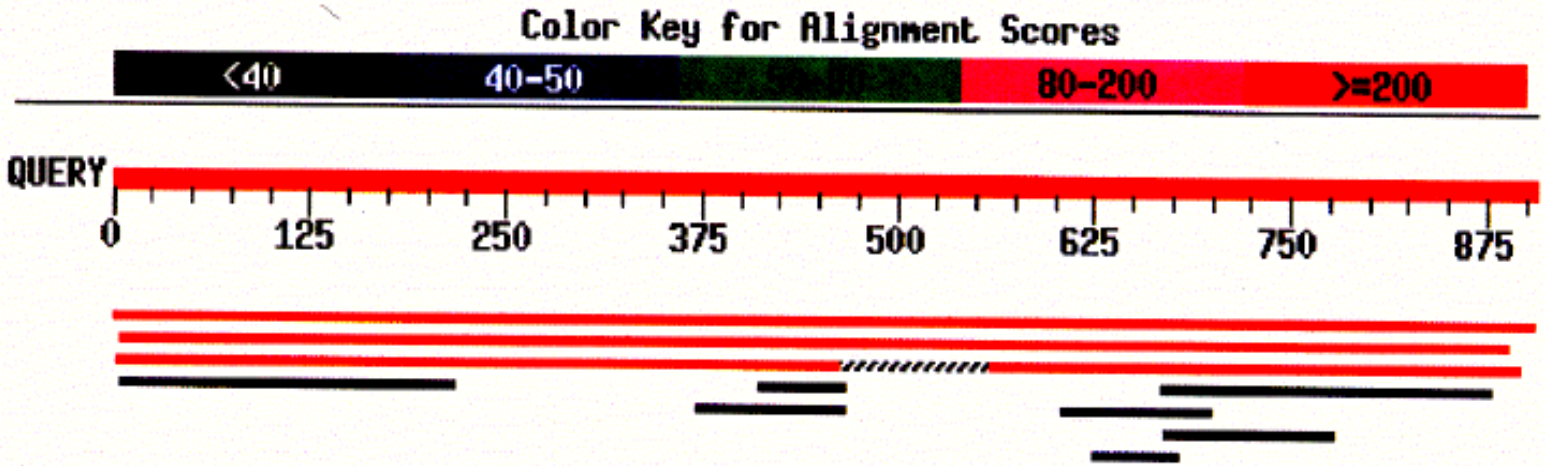
```
Query: 95  tccaccatgtcacctaagaaagccaagaagagagcagatggagccaaagnnnncgtgttc 154
           ||||||||||||||||||||||||||||||||||||||||||||||||||||       ||||||||
Sbjct: 1   tccaccatgtcacctaagaaagccaagaagagagcagatggagccaattccaacgtgttc 60


Query: 155 tccatgtttgaacagacccagattcaggaatttaaggaggccttcaccatcatggaccag 214
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 61  tccatgtttgaacagacccagattcaggaatttaaggaggccttcaccatcatggaccag 120


Query: 215 aacagggatggcttcatagacaagaatgatctgagggacacctttgctgctcttgggcgt 274
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 121 aacagggatggcttcatagacaagaatgatctgagggacacctttgctgctcttgggcgt 180


Query: 275 gtgaacgtgaaaaatgaggaaattgatgaaatgatcaaggaagctccaggtccaattaac 334
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 181 gtgaacgtgaaaaatgaggaaattgatgaaatgatcaaggaagctccaggtccaattaac 240


Query: 335 tttactgtgttcctaacgatgtttggggagaaacttaaggggcagaccctgaggagccc 394
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| ||
Sbjct: 241 tttactgtgttcctaacgatgtttggggagaaacttaaggggcagaccctgaggagacc 300


Query: 395 gcccttgacgcgttccaagtgtttgaccctgaaggcaaaggggtgctcagggctgattat 454
             ||||  |||||||| ||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 301 atccttaacgcgttcaaagtgtttgaccctgaaggcaaaggggtgctcagggctgattat 360
```

BLASTN Output

```
  Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS,
  or phase 0, 1 or 2 HTGS sequences)
    Posted date:  Sep 4, 2002 12:27 AM
  Number of letters in database: 2,147,257,520
  Number of sequences in database:  1,367,736

Lambda     K      H
    1.37    0.711     1.31

Gapped
Lambda     K      H
    1.37    0.711     1.31


Matrix: blastn matrix:1 -3
Gap Penalties: Existence: 5, Extension: 2
Number of Hits to DB: 3,320,183
Number of Sequences: 1367736
Number of extensions: 3320183
Number of successful extensions: 18698
Number of sequences better than 10.0: 207
length of query: 1447
length of database: 6,442,224,816
effective HSP length: 22
effective length of query: 1425
effective length of database: 6,412,134,624
effective search space: 9137291839200
effective search space used: 9137291839200
T: 0
A: 30
X1: 6 (11.9 bits)
X2: 15 (29.7 bits)
S1: 12 (24.3 bits)
S2: 20 (40.1 bits)
```

BLASTP output

```
BLASTP 2.0.5 [May-5-1998]
Query= human XP-F repair gene          (905 letters)

Database: Non-redundant SwissProt sequences 74,596 sequences; 26,848,718 total letters

B.
```



Distribution of 11 BLAST Hits on the Query Sequence

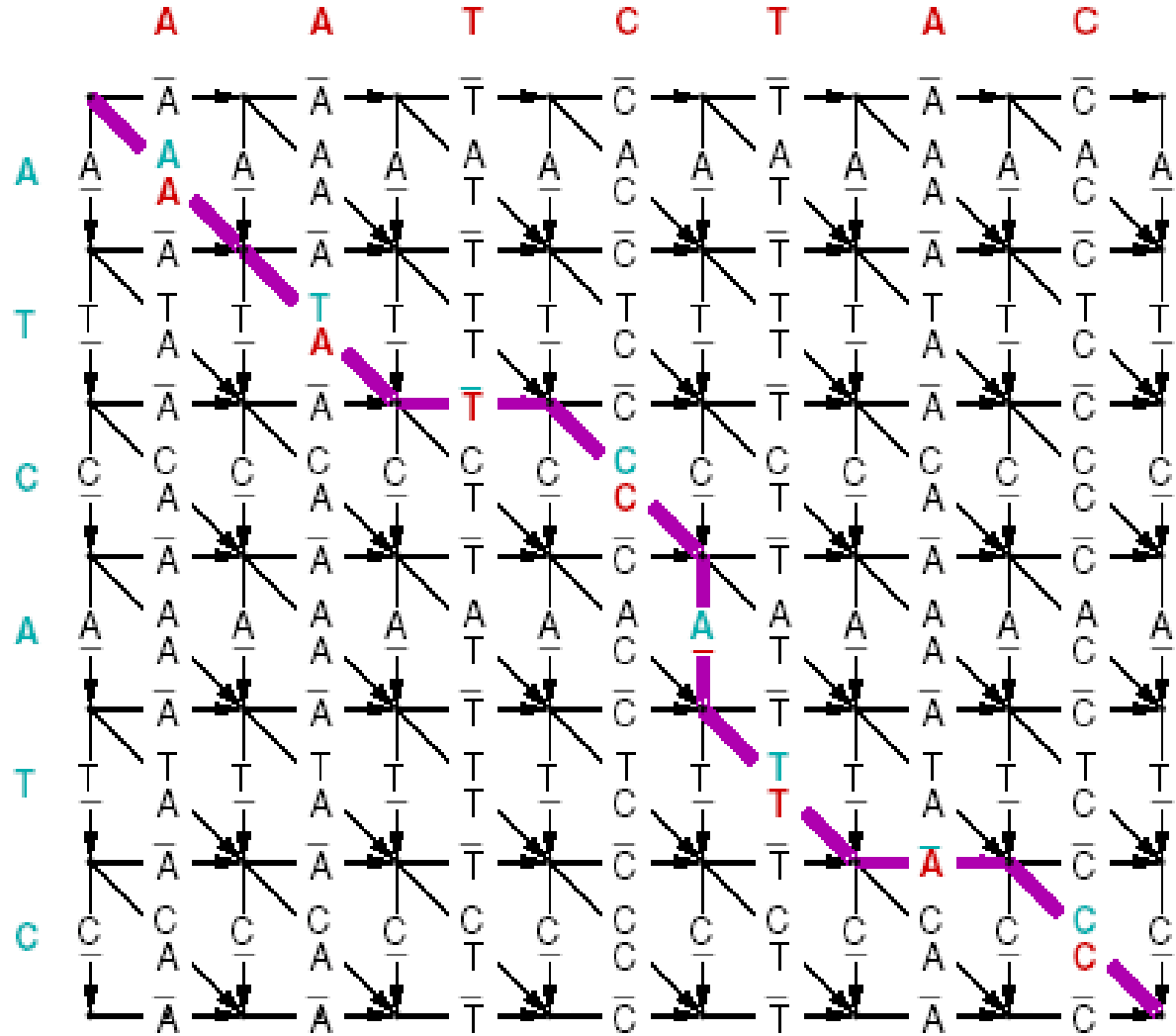
Sequences producing significant alignments:		Score (bits)	E Value
sp|Q92889|XPF_HUMAN	DNA-REPAIR PROTEIN COMPLEMENTING XP-F CELL ...	1659	0.0
sp|P36617|RA16_SCHPO	DNA REPAIR PROTEIN RAD16	485	e-136
sp|P06777|RAD1_YEAST	DNA REPAIR PROTEIN RAD1	231	4e-60
sp|P40562|YIS2_YEAST	PUTATIVE ATP-DEPENDENT RNA HELICASE YIR002C	37	0.17
sp|Q10202|YAXB_SCHPO	PUTATIVE ATP-DEPENDENT RNA HELICASE C13F4.11C	36	0.38

BLASTP

```
Database: Non-redundant SwissProt sequences
Number of letters in database: 26,848,718
Number of sequences in database:  74,596
Lambda      K       H
0.320    0.136    0.394
Gapped
Lambda      K       H
0.270   0.0470    0.230
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 42777291
Number of Sequences: 74596
Number of extensions: 1706128
Number of successful extensions: 4638
Number of sequences better than 10.0: 12
Number of HSP's better than 10.0 without gapping: 4
Number of HSP's successfully gapped in prelim test: 8
Number of HSP's that attempted gapping in prelim test: 4616
Number of HSP's gapped (non-prelim): 16
length of query: 905
length of database: 26848718
effective HSP length: 55
effective length of query: 850
effective length of database: 22745938
effective search space: 19334047300
effective search space used: 19334047300
T: 11
A: 40
X1: 16 ( 7.4 bits)
X2: 38 (14.8 bits)
X3: 64 (24.9 bits)
S1: 41 (21.8 bits)
S2: 68 (30.9 bits)
```

2. Dynamic programming

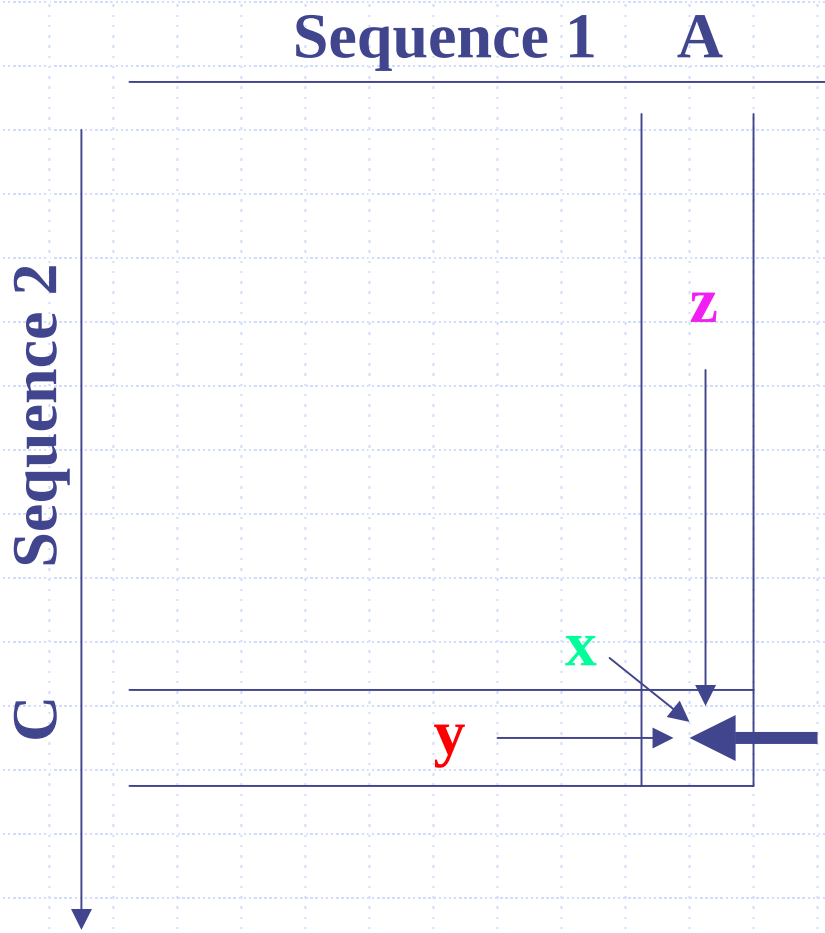
- In early 1950's, mathematician Richard Bellman, who was working at RAND Corporation on optimal decision processes, wanted to concoct an impressive name that would shield his work from U.S. Secretary of Defense Charles Wilson, a man known to be hostile to mathematics research. He figured dynamic programming was "something not even a Congressman could object to".
- The heart of many well-known programs is a dynamic programming algorithm, or a fast approximation of one, such as BLAST, CLUSTAL, HMMER, GENSCAN, MFOLD and PHYLIP.



Alignment corresponding to the colored path:

A T – C A T – C
A A T C – T A C

The DP algorithm



Do we get the best score
1. by aligning C with A and adding the score to the diagonal score x ?
OR
2. by placing a gap either opposite C or A and subtracting the gap penalty from the highest score in row y or column z ?

Mount D, 2002

该算法用代数形式来描述

单元(i,j)的距离可看成三个相邻单元的距离加上相应权重后的最小者，即

$$d(a^i,b^j)=\min\begin{cases}d(a^{i-1},b^j)+w_-(a_i)\\ d(a^{i-1},b^{j-1})+w(a_i,b_j)\\ d(a^i,b^{j-1})+w_+(b_i)\end{cases}$$

且初始条件为

$$d(a^0,b^0)=0$$

$$d(a^0,b^j)=\sum_{k=1}^{j}w_+(b_k)$$

$$d(a^i,b^0)=\sum_{k=1}^{i}w_-(a_k)$$

以两条短序列为例：

A: CTGTATC

B: CTATAATCCC

- 设碱基错配时距离权重为1，引入一个空位时距离权重为3。
- 可以得到多少可能的联配？

	A	C	T	A	T	A	A	T	C	C	C
B	0	3	6	9							
C	3	0	3	6							
T	6	3	0								
G	9	6	3								
T	12	9	6								
A	15	12									
T	18	15									
C	21	18									

	A	C	T	A	T	A	A	T	C	C	C
B	0	3	6	9	12	15	18	21	24	27	30
C	3	0	3	6	9	12	15	19	21	24	27
T	6	3	0	3	6	9	12	15	18	18	24
G	9	6	3	1	4	7	10	13	16	19	22
T	12	9	6	4	1	4	7	10	13	16	19
A	15	12	9	6	4	1	4	7	10	13	16
T	18	15	12	9	6	4	2	4	7	10	13
C	21	18	15	12	9	7	6	3	4	7	10

	A	C	T	A	T	A	A	T	C	C	C
B	0	3	6	9	12	15	18	21	24	27	30
C	3	0	3	6	9	12	15	19	21	24	27
T	6	3	0	3	6	9	12	15	18	18	24
G	9	6	3	1	4	7	10	13	16	19	22
T	12	9	6	4	1	4	7	10	13	16	19
A	15	12	9	6	4	1	4	7	10	13	16
T	18	15	12	9	6	4	2	4	7	10	13
C	21	18	15	12	9	7	6	3	4	7	10

Six possible alignments

CTATAATCCC
CTGTA–TC – –

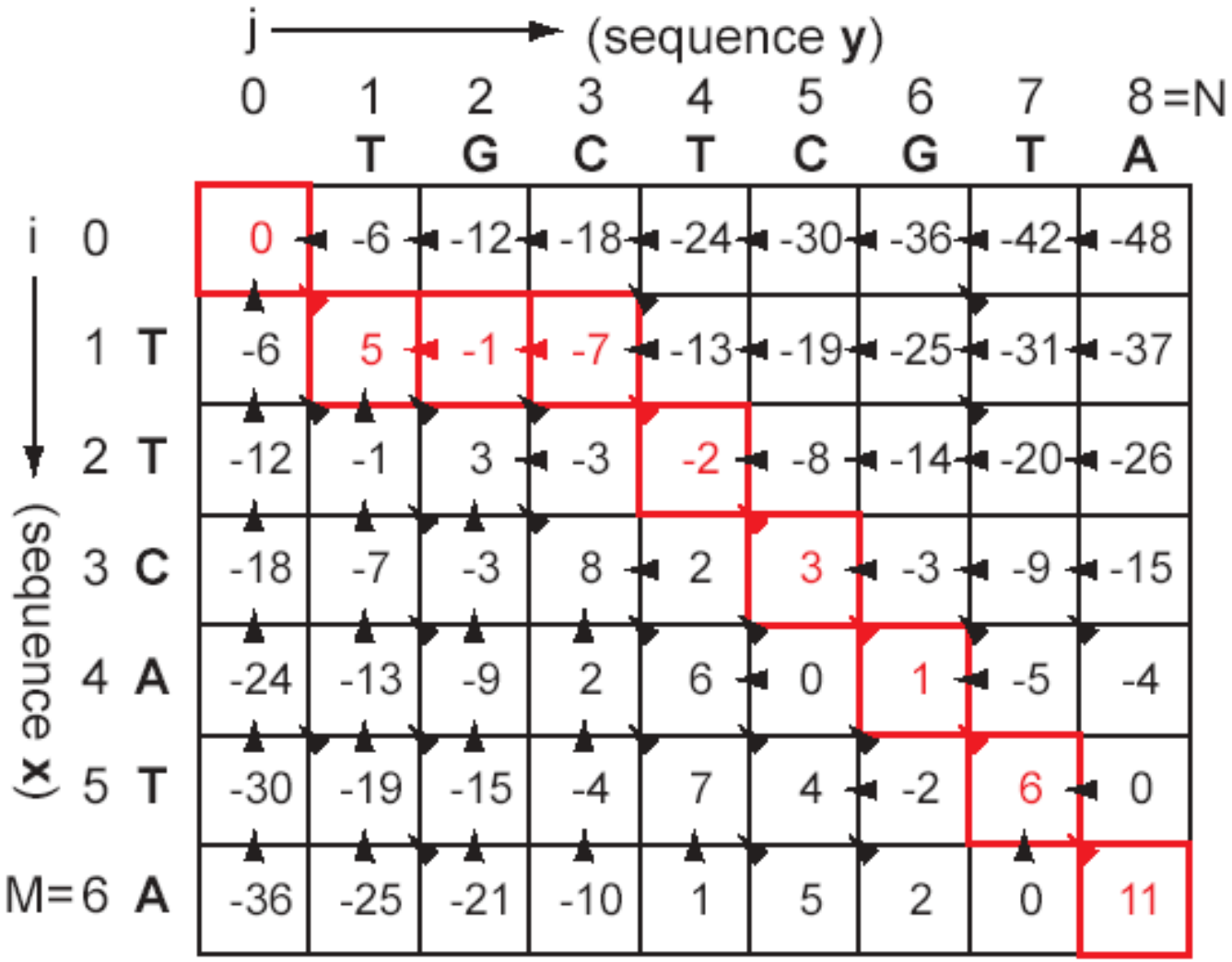
CTATAATCCC
CTGTA –T–C –

CTATAATCCC
CTGTA–T– –C

CTATAATCCC
CTGT–ATC– –

CTATAATCCC
CTGT –AT–C–

CTATAATCCC
CTGT –AT– –C



- +5 for a match, -2 for a mismatch and –6 for each insertion or deletion

optimum alignment scores 11:

T	-	-	T	C	A	T	A
T	G	C	T	C	G	T	A
+5	-6	-6	+5	+5	-2	+5	+5

But what do we really need to know?

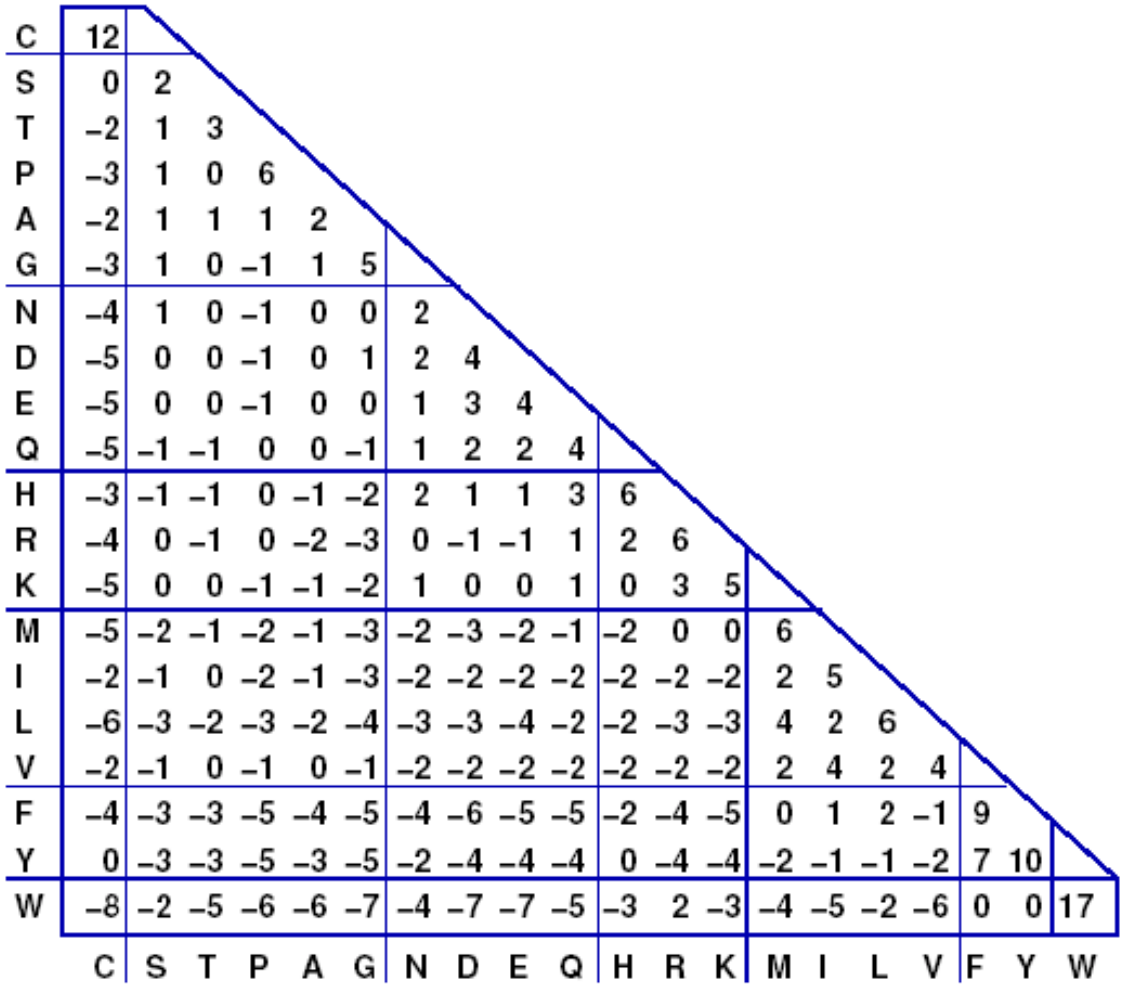
- It is guaranteed to give you a mathematically optimal (highest scoring) solution. Whether that corresponds to the biologically correct alignment is a problem for your scoring system, not for the algorithm;
- The question of when a score is statistically significant is a separate problem, requiring clever statistical theory;
- Dynamic programming is surprisingly computationally demanding. Alternatively, fast approximation, like workhorse BLAST, FASTA and BLAT were used;

- Only certain scoring systems are amenable to dynamic programming. The scoring system has to allow the optimal solution to be broken up into independent parts, or else it can't be dealt with recursively.

3. Scoring system (matrix)

- Almost all alignment methods find the best alignment between two strings under some scoring scheme;
- Back in the good old days, so many things were easier to understand. The first sequence comparisons just assigned –1 per mismatch and insertion/deletion;
- However, we want a system to give the biologically most likely alignment the highest score and to take into account the fact that biological molecules have evolutionary histories etc.

Where did the BLOSUM62 alignment score matrix come from?



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	12																			
S	0	2																		
T	−2	1	3																	
P	−3	1	0	6																
A	−2	1	1	1	2															
G	−3	1	0	−1	1	5														
N	−4	1	0	−1	0	0	2													
D	−5	0	0	−1	0	1	2	4												
E	−5	0	0	−1	0	0	1	3	4											
Q	−5	−1	−1	0	0	−1	1	2	2	4										
H	−3	−1	−1	0	−1	−2	2	1	1	3	6									
R	−4	0	−1	0	−2	−3	0	−1	−1	1	2	6								
K	−5	0	0	−1	−1	−2	1	0	0	1	0	3	5							
M	−5	−2	−1	−2	−1	−3	−2	−3	−2	−1	−2	0	0	6						
I	−2	−1	0	−2	−1	−3	−2	−2	−2	−2	−2	−2	−2	2	5					
L	−6	−3	−2	−3	−2	−4	−3	−3	−4	−2	−2	−3	−3	4	2	6				
V	−2	−1	0	−1	0	−1	−2	−2	−2	−2	−2	−2	−2	2	4	2	4			
F	−4	−3	−3	−5	−4	−5	−4	−6	−5	−5	−2	−4	−5	0	1	2	−1	9		
Y	0	−3	−3	−5	−3	−5	−2	−4	−4	−4	0	−4	−4	−2	−1	−1	−2	7	10	
W	−8	−2	−5	−6	−6	−7	−4	−7	−7	−5	−3	2	−3	−4	−5	−2	−6	0	0	17

- alignment scores are log-odds scores

$$s(a, b) = \frac{1}{\lambda} \log \frac{p_{ab}}{f_a f_b}$$

- target/ background frequencies
- lambda is a scaling factor, which is usually set to something that lets us round off all the terms in the sore matrix to sensible integers.

BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4

Block IPB000409

Q9C728	(780)	PWAKGSPEMFIARNREALESEYVSSHLHDWIDLIFGHKQRG
Q9ZQX5	(3071)	PWARGSVREFIRKHREALESDYVSENLHHWIDLIFGHKQRG
Q9C6Q7	(627)	PWAKGSPEMFIARNREALESEYVSSHLHDWIDLIFGHKQRG
O64634	(2414)	PPWAKNPVDFVHKQRRALESEHVSAHLHEWIDLIFGYKQRG
Q9Z2X9	(3315)	PWARNDPRLFILIHRQALESDHVSQNICHWIDLVFGYKQKG
Q9EPM9	(2458)	PPWAKKPEDFVRINRMALESEFVSCQLHQWIDLIFGYKQRG
Q9ESE1	(2387)	PPWAKTSEEFVRINRLALESEFVSCQLHQWIDLIFGYKQQG
Q9EPN0	(2426)	PPWAKKPEDFVRINRMALESEFVSCQLHQWIDLIFGYKQRG
Q9ESD3	(2387)	PPWAKTSEEFVRINRLALESEFVSCQLHQWIDLIFGYKQQG
Q9WVM9	(239)	PPWAKKPEDFVRINRMALESEFVSCQLHQWIDLIFGYKQRG
Q9EPN1	(2458)	PPWAKKPEDFVRINRMALESEFVSCQLHQWIDLIFGYKQRG
Q9ESD4	(2387)	PPWAKTSEEFVRINRLALESEFVSCQLHQWIDLIFGYKQQG
Q8CHB9	(1)	PWAKGDPREFIRVHREALECDYVSAHLHEWIDLIFGYKQQG
Q8CBR4	(292)	PPWAETSEEFVRINRLALESEFVSCQLHQWIDLIFGYKQQG
Q8C931	(261)	PPWAKKPEDFVRINRMALESEFVSCQLHQWIDLIFGYKQRG
Q8C7M6	(469)	PAWASSPQDFLQKNKDALESGYVSEHLHEWIDLIFGYKQKG
Q8BSM6	(289)	PPWAKTSEEFVRINRLALESEFVSCQLHQWIDLIFGYKQQG
Q8BRX5	(45)	PRWAKSAEDFIYKHRKALESEYVSAHLHEWIDLIFGYKQRG
Q8BQ04	(225)	PWARNDPRLFILIHRQALESDHVSQNICHWIDLVFGYKQKG

- BLOSUM62

$$s_{ij} = 2\log_2(q_{ij}/e_{ij})$$

(half-bit unit)

For example, L/L and W/W pair are 3.8 and 10.5, respectively (p_{LL}=0.0371, p_{WW}=0.0065; f_L-0.099, f_W=0.013; BLOSUM62's original lamda=0.0347)

Making up your own score matrices

- PSI-BLAST: position-specific scoring matrix (PSSM)
- Exercise:

  Make a DNA scoring matrix optimized for finding 88% identify alignments. Assume that all mismatches are equiprobable and background sequences is uniform at 25% for each nucleotide. Scale up with half-bit unit (BLOSUM62) and round off.

4. Entropy: *H*

- How to evaluate the Quality of a matrix (such as BLOSUM62/PAM/PSSM)
- How well BLOSUM or PAM matrix to discriminate real local alignments from chance alignments?

  or how much information content in the scoring matrix?
- Uncertainty and information

Meaning of Uncertainty

Uncertainty is the no. of questions that must be asked to identify the correct choice i.e. the correct base or amino acid in one location of a particular pattern.

Uncertainty is reduced by the information in the scoring matrix

Example: consider 64 cups in a row with an object hidden under one of them. The goal is to find the object with as few questions as possible. (Answer: uncertainty is six and is zero when the object is found)

	1	2	3
G	1.0	0.0	0.25
A	0.0	0.5	0.25
C	0.0	0.0	0.25
T	0.0	0.5	0.25

How much uncertainty is there for: column 1, column 2 and column 3?
For DNA sequences, what is the maximum uncertainty there can be? What is the minimum uncertainty?

Information and Probability

$H = \log N = -\log P, \quad P = 1/N$ (Hartly, 1928)

$H(P) = -\sum p_i \log(p_i)$ *(Shannon, 1948)*

(in units of $\log_2$, called bits)

$$H(P, Q) = H(X, Y) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}$$

(Relative entropy is also called cross-entropy, or Kullback-Liebler distance. It is reviewed as a measure of the distance between two distributions P and Q. The more dissimilar P and Q are, the larger the relative entropy.)

Relative entropy of matrix

$$H = \sum_{i-1}^{20} \sum_{j=1}^{i} q_{ij} \times s_{ij}$$

In information theory, this weighted score is called the average mutual information content per pair, and the sum over all pairs.

In general, all other factors being equal, the higher the value of H for a scoring matrix, the more likely it is to be able to distinguish real from chance alignments.

Calculating Uncertainty and Information Content for PSSM

Uncertainty (H):

$$H = f_G \log_2 (f_G) + f_A \log_2 (f_A) + \ldots.$$

In general, the average amount of uncertainty (H_c) in bits per symbol for column c of the PSSM is given by

$$H_c = - \sum_{All\ i} p_{ic} \log_2(p_{ic})$$

for entire PSSM

$$H = \sum_{All\ columns} H_c$$

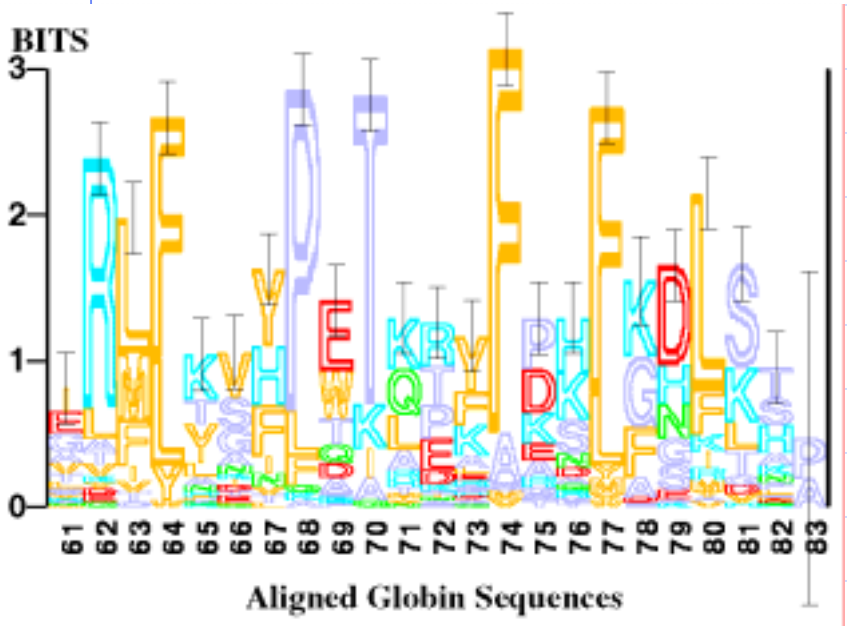
***H* is also known as the entropy of the PSSM position in information theory because the higher the value, the greater the uncertainty.**

Information content (IC):

IC for column = 2 – H (for DNA sequence)

IC for whole PSSM = sum of columns ICs

Sequence Logo



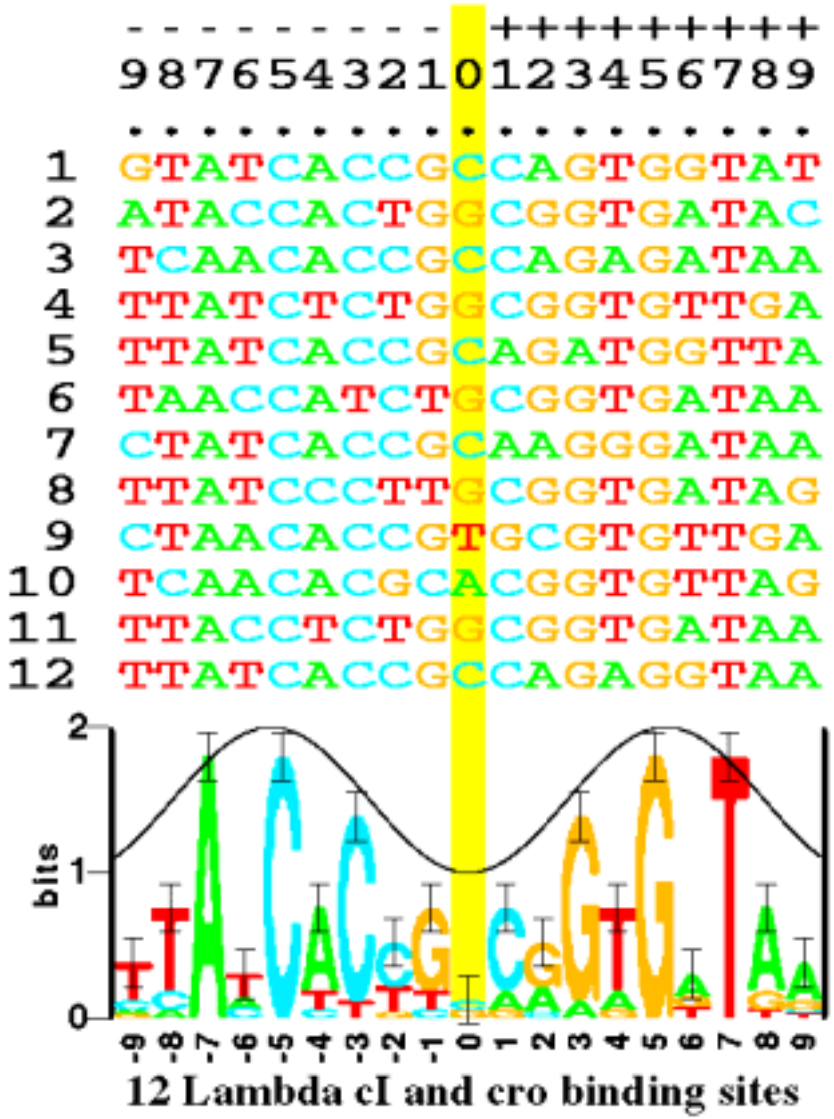


Fig. 1. Some aligned sequences and their sequence logo. At the top of the figure are listed the 12 DNA sequences from the P_L and P_R control regions in bacteriophage lambda. These are bound by both the cI and cro proteins [16]. Each even numbered sequence is the complement of the preceding odd numbered sequence. The sequence logo, described in detail in the text, is at the bottom of the figure. The cosine wave is positioned to indicate that a minor groove faces the center of each symmetrical protein. Data which support this assignment are given in reference [17].

5. Score and E-value

- Bit score

$s = a*MA+b*MM-c*OG-d*EG$

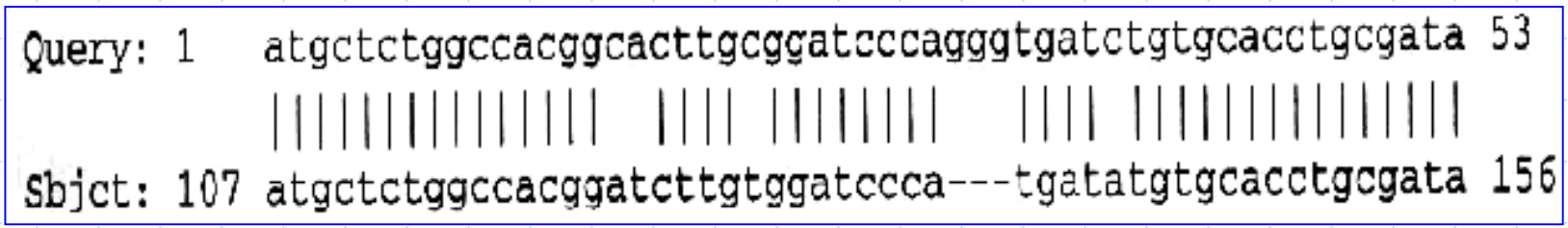
a=1, b=-3, c=5, d=2 (BLASTn)

By normalizing a raw score s using the formula

$$s' = \frac{\lambda s - \ln k}{\ln 2}$$

The bit score (standard unit) is independent to scoring system

- A example: 53 letters BLASTn against GenBank database:



$S = (46)(1)+(-3)(4)-(5)(1)-(3)(2) = 23$

$$S' = \frac{(1.37)(23) - \ln(0.771)}{\ln 2} \approx 46$$

E-value

The expected number of HSPs (high-scoring segment pairs) with score at least *s* is given by the formula

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

(The longest run of matches in a alignment is equivalent to the longest run of heads in the coin-tossing sequence. and it should be possible to use the Erdos and Renyi law to predict the longest run of matches. See Mount's book, pp91-92)

The E-value corresponding to a given bit score is simply

$$E = mn2^{-s'}$$

So that to calculate significance one needs to know in addition only the size of the search space (i.e. *n* and *m*).

Again, the above example:

$$E = mn2^{-s'} = 34 \times 5854611841 \times 2^{-46} = 0.003$$

- P-value

The number of random HSPs with score >=s is described by a Poisson distribution.

Specifically, the chance of finding zero HSPs with score >=s is $P_0 = e^{-x}$ or e^{-E}, so the probability of finding at least one such HSP is

$$P = 1 - e^{-E}$$
$$= 1 - e^{Kmne^{-\lambda s}}$$

The Gumbel extreme value distribution

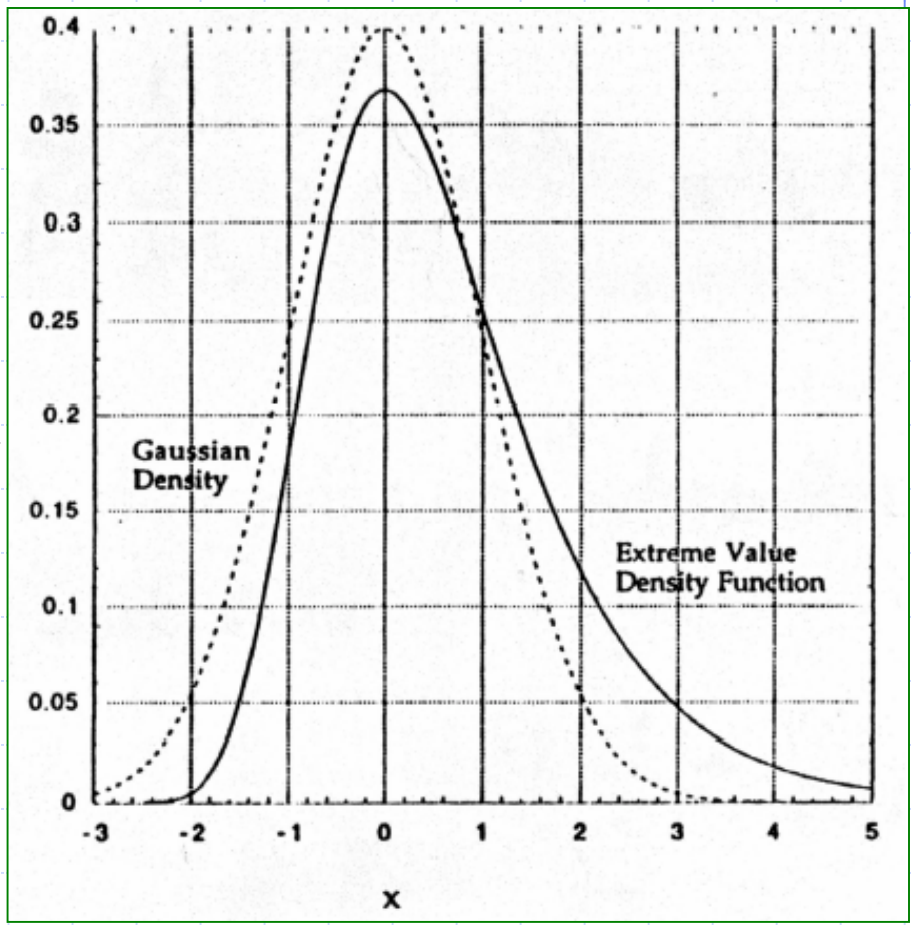
$P(s \geqq x) =$
$1-\exp[-e^{-\lambda(x-\mu)}]$

$\lambda = \pi(\sigma\sqrt{6})$

$\mu = \underline{x} - \gamma/\pi = \underline{x} - 0.450\sigma$
$= (\ln Kmn)/\lambda$

$P(s \geqq x) = 1 - e^{Kmne^{-\lambda s}}$
$\approx Kmne^{-\lambda s}$

$$P(S \geq x) = 1 - \exp[-e^{-x}] \rightarrow e^{-x}$$



6. λ and K

- **Can be thought of simply as the scoring system and natural scales for the search space size;**
- **recall that**

s(a,b) = $\frac{1}{\lambda}\log\frac{p_{ab}}{f_a f_b}$

μ=(lnKmn)/λ

$$P_{ab} = f_a f_b e^{\lambda s_{ab}}$$

$$\sum_{a,b} f_a f_b e^{\lambda s_{ab}} = 1$$

K is a constant that can be calculated from the value of p_i and s_{ij};

It depends on the base composition.