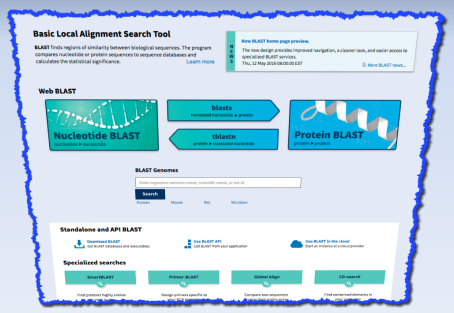
BLAST

Basic Local Alignment Search Tool



NCBI

Today

What is BLAST used for?

Algorithm basics

Hands-on

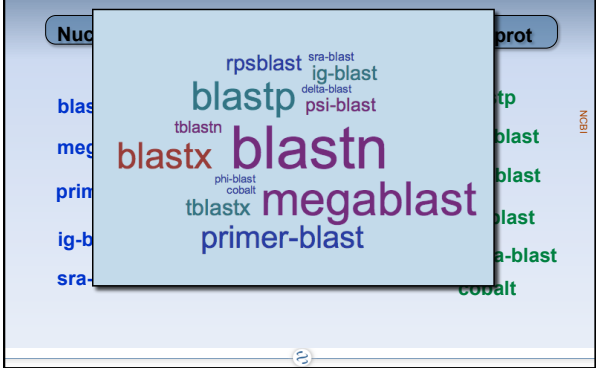
NCBI

BLAST Uses

- **Identification**
  What microbial species is this?
- **Homology**
  - Related sequences
  - Genome annotation
  - Function
- **Location**
  - Is my oligo/primer unique?
  - Where does it align?

NCBI

Family of BLAST Programs



Location Example

Query = TCTAAGTGAAAATCATGACATTTG (24 bases)

Question: Is this 24-mer unique in the human genome?

Database = human genome

(RID = KTY9C75F013)

NCBI

Nucleotide Words

Query: CTTT**TGTATGC**TCATAATCG

7-mer

word size 7

Lookup Table (of words)

CTTTTGT
TTTTGTA
TTTGTAT
TTGTATG
TGTATGC
. . .

NCBI

Basic Algorithm

- Build lookup table
- Find word matches
- Extend
  - no gaps
  - gaps

NCBI

Protein Words

Query: GTQITVEDLFYNIATRRKALKN

GTQ
TQI
QIT
ITV → Neighborhood Words LTV, MTV, ISV, LSV, etc.

Lookup Table

GTQITVEDLFYNIATRRKALKN
TQ+TVED YNI TR KAL+N
CTQLTVEDNFYNIPTRAKALEN

NCBI

Expect Value

E = number of database hits you expect to find by chance, ≥ S

$E = Kmne^{-\lambda S}$ or $E = mn2^{-S'}$

K = scale for search space
λ = scale for scoring system
λS' = bitscore = (λS – lnK)/ln2
m = query length
n = database length

E is dependent on m x n (search space)

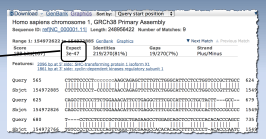
More info: www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html
Videos: http://bit.ly/ZhVPDQ

NCBI

Expect Value Uses

Chance?

Expect 3e-47



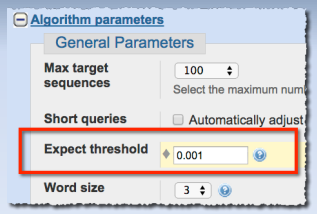
Sorting

Max score	Total score	Query cover	E value	Ident
184	184	59%	4e-51	38%
187	187	56%	5e-51	39%
187	187	60%	7e-51	38%
186	186	55%	1e-50	39%
186	186	55%	1e-50	39%
183	183	56%	9e-50	39%
181	181	56%	2e-49	38%
183	183	56%	2e-49	37%

NCBI

Expect Value Uses – con't

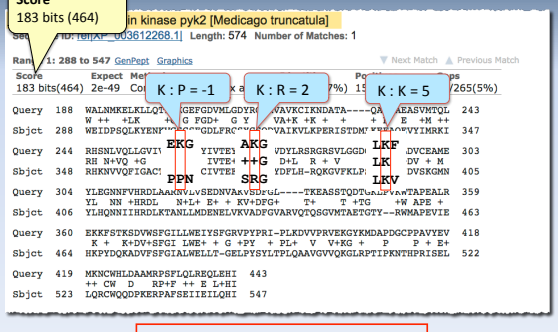
Expect threshold (cutoff)



NCBI

Scores - protein

Score
183 bits (464)

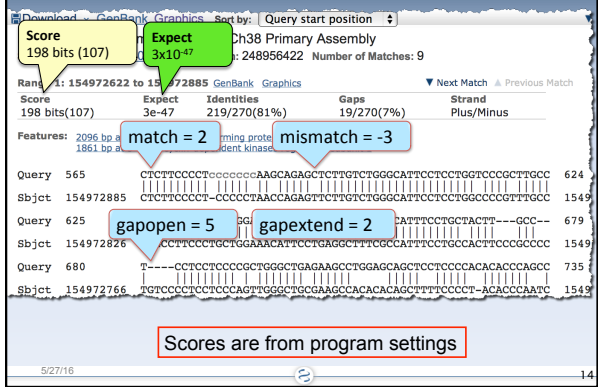


NCBI

Scores from BLOSUM62 matrix

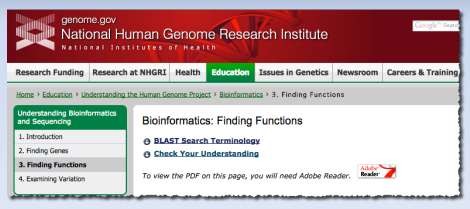
```
#  Matrix made by matblas from blosum62.iij
#  * column uses minimum score
#  BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
#  Blocks Database = /data/blocks_5.0/blocks.dat
#  Cluster Percentage: >= 62
#  Entropy =   0.6979, Expected =  -0.5209
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  Z  X  *
A  4 -1 -2 -2  0 -1 -1  0 -2 -1 -1 -1 -1 -2 -1  1  0 -3 -2  0 -2 -1  0 -4
R -1  5  0 -2 -3  1  0 -2  0 -3 -2  2 -1 -3 -2 -1 -1 -3 -2 -3 -1  0 -1 -4
N -2  0  6  1 -3  0  0  0  1 -3 -3  0 -2 -3 -2  1  0 -4 -2 -3  3  0 -1 -4
D -2 -2  1  6 -3  0  2 -1 -1 -3 -4 -1 -3 -3 -1  0 -1 -4 -3 -3  4  1 -1 -4
C  0 -3 -3 -3  9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1  1  0  0 -3  5  2 -2  0 -3 -2  1  0 -3 -1  0 -1 -2 -1 -2  0  3 -1 -4
E -1  0  0  2 -4  2  5 -2  0 -3 -3  1 -2 -3 -1  0 -1 -3 -2 -2  1  4 -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 -1 -4
H -2  0  1 -1 -3  0  0 -2  8 -3 -3 -1 -2 -1 -2 -1 -2 -2  2 -3  0  0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4  2 -3  1  0 -3 -2 -1 -3 -1  3 -3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4 -2  2  0 -3 -2 -1 -2 -1  1 -4 -3 -1 -4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5 -1 -3 -1  0 -1 -3 -2 -2  0  1 -1 -4
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5  0 -2 -1 -1 -1 -1  1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6 -4 -2 -2  1  3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4  1 -3 -2 -2  0  0  0 -4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5 -2 -2  0 -1 -1  0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11  2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7 -1 -3 -2 -1 -4
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4 -3 -2 -1 -4
B -2 -1  3  4 -3  0  1 -1  0 -3 -4  0 -3 -3 -2  0 -1 -4 -3 -3  4  1 -1 -4
Z -1  0  0  1 -3  3  4 -2  0 -3 -3  1 -1 -3 -1  0 -1 -3 -2 -2  1  4 -1 -4
X  0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -2  0  0 -2 -1 -1 -1 -1 -1 -4
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4  1
```

Scores - BLASTN

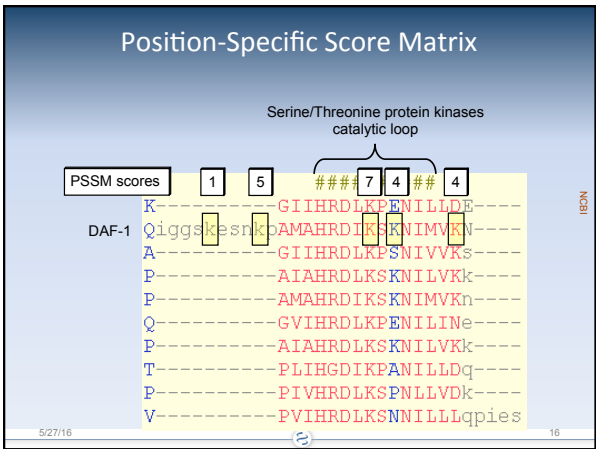


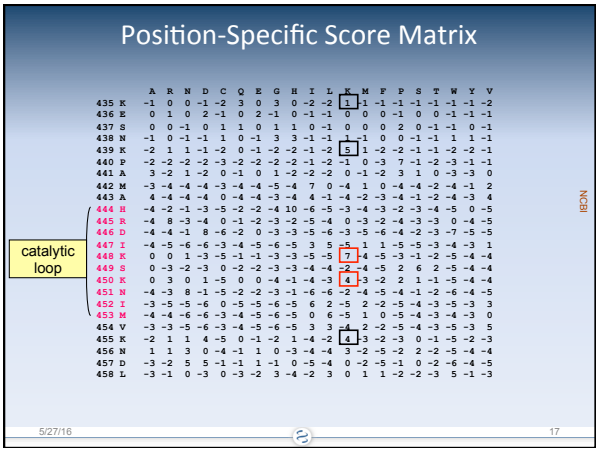
Scores are from program settings

Background Reading ...



http://www.genome.gov/25020002


Position-Specific Score Matrix
Serine/Threonine protein kinases
catalytic loop
PSSM scores
1
5
7
4
4
DAF-1
K----------GIIHRDLKPENILLDE----
Qiggskesnkqo AMAHRDIKSKNIMVKN----
A----------GIIHRDLKPSNIVVKs----
P----------AIAHRDLKSKNILVKk----
P----------AMAHRDIKSKNIMVKn----
Q----------GVIHRDLKPENILINe----
P----------AIAHRDLKSKNILVKk----
T----------PLIHGDIKPANILLDq----
P----------PIVHRDLKSPNLLVDk----
V----------PVIHRDLKSNNILLLqpies
NCBI
5/27/16
16


Position-Specific Score Matrix
catalytic loop
NCBI
5/27/16
17