Advanced BLAST

Words Scores Expect values

Tips for Web BLAST

Alternatives to Web BLAST

Wayne Matten
matten@ncbi.nlm.nih.gov

NCBI Discovery Workshops

Nucleotide Words

Query: GTACTGGACATGGACCCTACAGGAA

11-mer

word size

Lookup Table

```
GTACTGGACAT
 TACTGGACATG
  ACTGGACATGG
   CTGGACATGGA
    TGGACATGGAC
     GGACATGGACC
      . . .
```

words

NCBI Discovery Workshops

Basic Algorithm

word (seed)

word size = 7

```
CACTTTTGTATGCTCATAATCGATCACCTGTAGAG
||||  |||||||  |||      ||| |   |||
CACTACTGTATGCCGATA------CACTTCCGGAG
```

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

NCBI Discovery Workshops

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 Discovery Workshops

Scores - protein

Score
183 bits (464)

K : P = -1 K : R = 2 K : K = 5

```
Sequence: ...in kinase pyk2 [Medicago truncatula]
ID: ref|XP_003612268.1| Length: 574 Number of Matches: 1

Range 1: 288 to 547 GenPept Graphics        ▼ Next Match  ▲ Previous Match
Score          Expect  Methods                          Identities  Positives  Gaps
183 bits(464)  2e-49   Co...k a...7%) 15...  ...(265(5%)

Query  188  WALNMKELKLQV...GEFGDVMLGDYR...VAVKCIKNDATA----QA...EASVMTQL  243
            W ++  +LK   +...G FGD+  G Y ...VA+K +K +   +   + ...E  +M ++
Sbjct  288  WEIDPSQLKYENY...GDLFR...VAIKVLKPERISTDM...VYIMRKI  347

Query  244  RHSNLVQLLGVI...YIVTE...VDYLRSRGRSVLGGD...DVCEAME  303
            RH N+VQ +G      IVTE...++G  D+L  R + V   LK  DV + M
Sbjct  348  RHKNVVQFIGAC...CIVTE...YDFLH-RQKGVFKLP...DVSKGMN  405

Query  304  YLEGNNFVHRDL...SEDNVA...----TKEASSTQDT...TAPEALR  359
            YL  NN +HRDL   N+L+ E+ + KV+DFG+    T+    T +TG    +W APE +
Sbjct  406  YLHQNNIIHRDL...ANLLMDENELVKVADFGVARVQTQSGVMTAETGTY--RWMAPEVIE  463

Query  360  EKKFSTKSDVWSFGILLWEIYSFGRVPYPRI-PLKDVVPRVEKGYKMDAPDGCPPAVYEV  418
             K +  K+DV+SFGI LWE+ + G +PY  + PL+  V  V+KG +   P    P + E+
Sbjct  464  HKPYDQKADVFSFGIALWELLT-GELPYSYLTPLQAAVGVVQKGLRPTIPKNTHPRISEL  522

Query  419  MKNCWHLDAAMRPSFLQLREQLEHI  443
            ++ CW  D   RP+F ++ E L+HI
Sbjct  523  LQRCWQQDPKERPAFSEIIEILQHI  547
```

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

Score
198 bits (107)

Expect
$3x10^{-47}$

match = 2

mismatch = -3

gapopen = 5

gapextend = 2

Scores are from program settings

7

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

8

Expect Value Uses

Chance?

Expect
3e-47

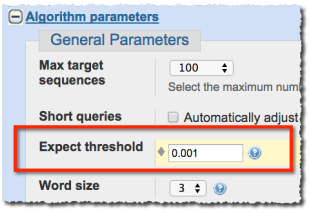
Sorting

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%

9

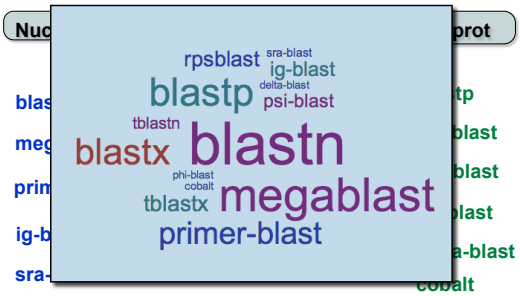
Expect Value Uses

Expect threshold (cutoff)



NCBI Discovery Workshops

Family of BLAST Programs



NCBI Discovery Workshops

Web BLAST Tips

Error: CPU usage limit was exceeded, resulting in SIGXCPU (24).

To BLAST Pages

Example failed search: RID = 0MBDBT7Y014

NCBI Discovery Workshops

Alternatives to Web BLAST

Error: CPU usage limit was exceeded, resulting in SIGXCPU (24).

Still?

1) Blast+ package (dalone, command-line)
2) Local SRA b
3) Blast in the cloud
4) BLAST URL API

To BLAST Help Pages

NCBI Discovery Workshops