Example Blast Hit: What is the expected value?

EXAMPLE BLAST HIT

```
 >gi|14211319|gb|AF375323.1| Hylurgops sp. HLH04
cytochrome oxidase I (CO1) gene, partial cds;
mitochondrial gene for mitochondrial product Length
= 1348

Score = 92.5 bits (46), Expect = 9e-17 Identities =
53/56 (94%) Strand = Plus / Plus

Query: 4   ggggcaggtactggttgaayagtttatcccccgttagcttcaaatatttcccatga 59
            |||||||| |||||||||| |||||||||||| ||||||||||||||||||||||
Sbjct: 223 ggggcaggaactggttgaacagtttatccccctttagcttcaaatatttcccatga 278
```

BLAST Algorithm

1) Break the query sequence into words

```
AGTACTAATACAAAATTT
AGTA
 GTAC
  TACT
```

2) Search for EXACT word matches

```
          AGTA
CGCGAGCTGTAGCAAGTACTATTAC. . .
```

3) Extend the match until …when exactly?

```
          AGTACTAATACAAAT
CGCGAGCTGTAGCAAGTACTATTACCCCGGG. . .
```

Getting worse, but does computer know when to stop?

DNA Scoring Matrices

BLAST

	A	T	C	G
A	5	-4	-4	-4
T	-4	5	-4	-4
C	-4	-4	5	-4
G	-4	-4	-4	5

```
Query: A G C T A
         5
Subj : A A C T T
```

SCORE: 5

DNA Scoring Matrices

BLAST

	A	T	C	**G**
A	5	-4	-4	**-4**
T	-4	5	-4	-4
C	-4	-4	5	-4
G	-4	-4	-4	5

Query: A G C T A

5 -4

Subj : A A C T T

SCORE: 1

DNA Scoring Matrices

BLAST

	A	T	**C**	G
A	5	-4	-4	-4
T	-4	5	-4	-4
C	-4	-4	5	-4
G	-4	-4	-4	5

Query: A G C T A

5-4+5

Subj : A A C T T

SCORE: 6

DNA Scoring Matrices

BLAST

	A	T	C	G
A	5	-4	-4	-4
T	-4	5	-4	-4
C	-4	-4	5	-4
G	-4	-4	-4	5

Query: A G C T A

5-4+5+5-4

Subj : A A C T T

FINAL SCORE: 7

Test Yourself

Seq1: AGTTGATGA

Seq2: AAGAGTTTTAAG

WORD size = **3** Threshold: **5**

Stop when score goes below threshold.

	A	**T**	**C**	**G**
A	5	-4	-4	-4
T	-4	5	-4	-4
C	-4	-4	5	-4
G	-4	-4	-4	5

BLAST Algorithm

1) Break the query sequence into words

```
RILYPATVIGCT
RILY
 ILYP
  LYPA
   YPAT
```

2) Search for EXACT word matches

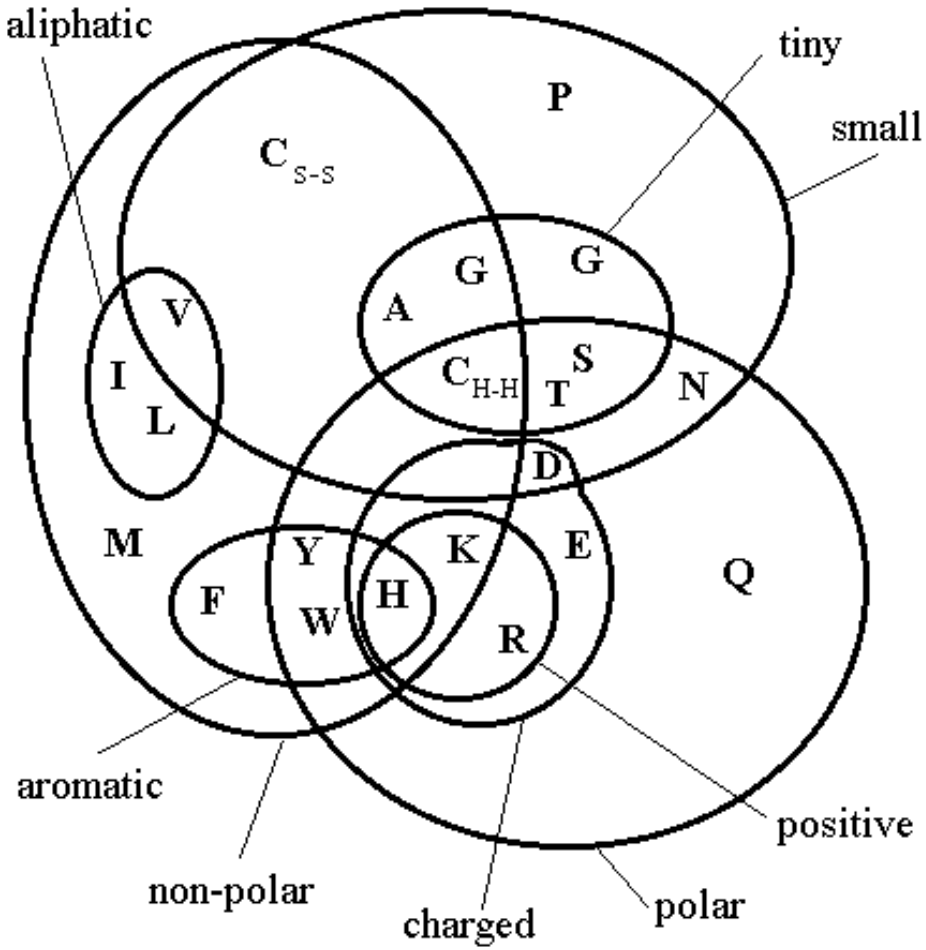
```
          ILYP
MVQGWALYDFLKCRAILYPGTVLMRWW…
```

3) Extend the match until it falls below a fixed threshold

```
          ←RILYPATVIG→
 MVQGWALYDFLKCRAILYPGTVLMRWW…
```

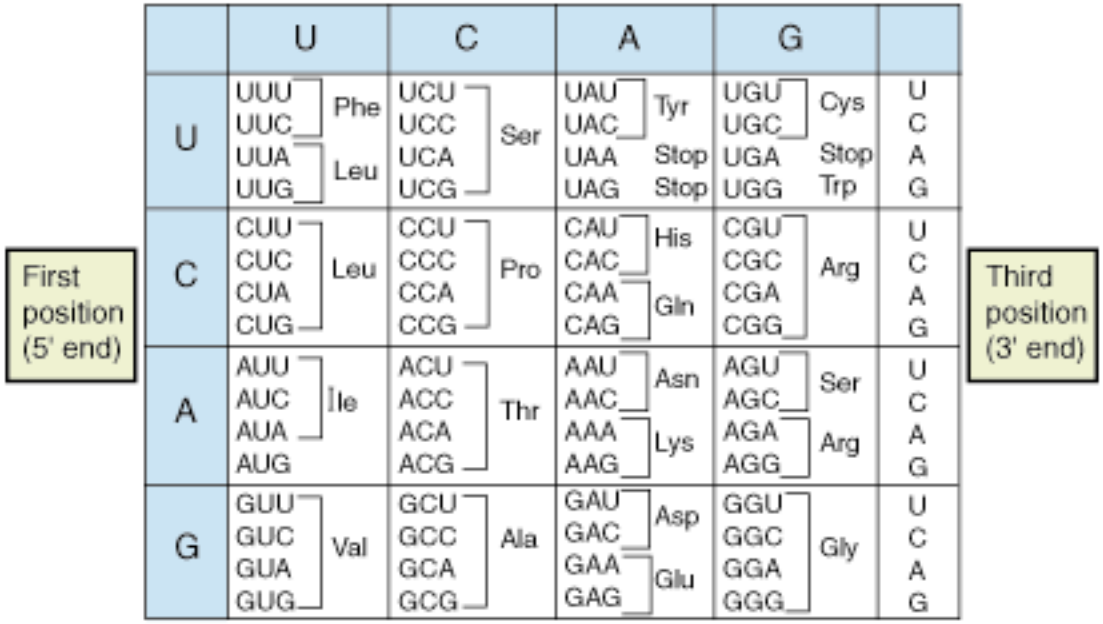
What kind of scoring scheme would you use?

Amino Acid Properties



PAM pre-exercise

Genetic Code



First position (5' end)	U	C	A	G	Third position (3' end)
U	UUU Phe UUC UUA Leu UUG	UCU Ser UCC UCA UCG	UAU Tyr UAC UAA Stop UAG Stop	UGU Cys UGC UGA Stop UGG Trp	U C A G
C	CUU Leu CUC CUA CUG	CCU Pro CCC CCA CCG	CAU His CAC CAA Gln CAG	CGU Arg CGC CGA CGG	U C A G
A	AUU Ile AUC AUA AUG	ACU Thr ACC ACA ACG	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG	U C A G
G	GUU Val GUC GUA GUG	GCU Ala GCC GCA GCG	GAU Asp GAC GAA Glu GAG	GGU Gly GGC GGA GGG	U C A G

Amino acid names:

Ala = alanine
Arg = arginine
Asn = asparagine
Asp = aspartate
Cys = cysteine
Gln = glutamine
Glu = glutamate
Gly = glycine
His = histidine
Ile = Isolevcine
Leu = leucine
Lys = lysine
Met = methionine
Phe = phenylalanine
Pro = proline
Ser = serine
Thr = threonine
Trp = tryptophan
Tyr = Tyrosine
Val = valine