How where contigs assembled?

Sequence alignment

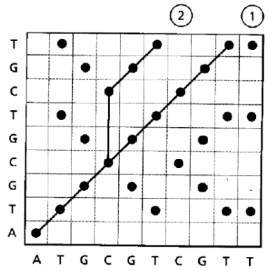
BLAST scores

Demo of NCBI and UCSC

How do we score these alignments?

Which alignment Do we choose?

How do we score gaps?




```
ATGCGTCGTT
|||||||||
ATGCGTCGT
```

```
ATG---CGTCGTT
|||   |||
ATGCGTCGT
```

Basic sequence alignment and database searching:
BLAST nucleotide searches

Sophisticated programming

- Dynamic programming
- Look up tables
  - Database broken up into short segments which are screened first for exact matches (word size). If a "hit" occurs, then try to extend.
- Needle-Wunsch alignments
  - Optimal alignment

Basic Local Alignment Search Tool (BLAST)

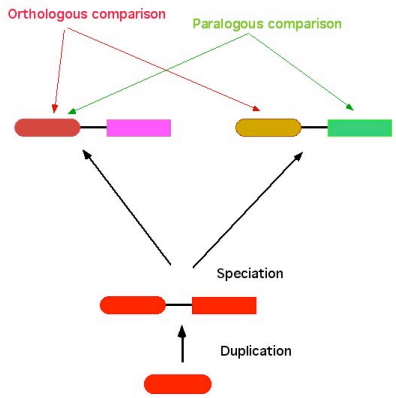
- Compares a QUERY (your sequence) to a database (i.e. GenBank).
- Different programs depending on sequence
  - BLASTN (DNA vs DNA)
  - BLASTP (AA vs AA)
  - BLASTX (Translated DNA vs AA)
  - TBLAST: Translated database searches

Main questions for today

- I have a DNA/protein sequence, is it similar to anything else in the database?
- My sequence is similar to something in the database, is it significant?

What does **statistically significant** similarity mean?

- Common ancestry
  - Homologous (all or none)
    - Orthologous: separated by speciation event: same locus
    - Paralogous: Separated by gene duplication: different locus
- **May** provide clues to similar function



Three scores from BLAST

- Raw score
  - No information unless know scoring matrix
- Bit Score
  - Takes into account scoring matrix
- E-value
  - Easy to interpret statistic, takes into account database search size.

Raw score

- Sum up score for matches
- Subtract penalties for miss-match
- Subtract penalties for gaps (open and extend)

```
C    C    A    A    G    A    C
|    |    |         |    |    |
C    C    A    T    G    A    C
1    1    1   -2    1    1    1

Raw Score = 1+1+1-2+1+1+1 = 4
```

Bit score: take into account the scoring system

- Normalize raw scores
- Two values describe scoring system
  - *K* and *lambda* (λ) (both values determined by program)

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

Bit scores can be converted into E-values, which are easy to understand

- Takes into account the search space
  - m and n (database and query size)

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

- One equation from score to E-value

$$E = Kmn\, e^{-\lambda S}$$

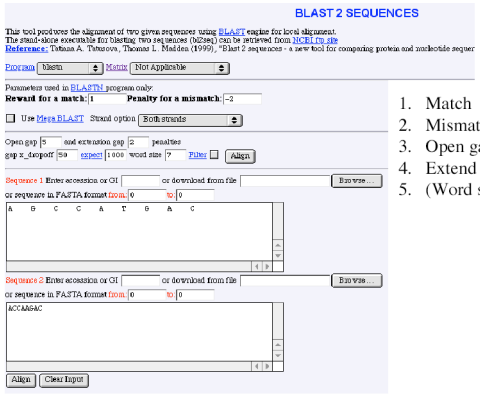
What does E-value mean?

The Expect value (E) is a parameter that describes the number of hits one can "expect" to see just by chance when searching a database of a particular size. Typical cut-offs biologists use are ~< $E = 10^{-5}$

- **An E value of <0.05 is considered statistically significant.**

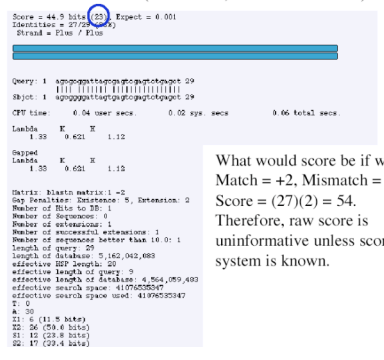
Example of score calculation using BLAST two sequences

- Aligns two sequences.
- Calculates statistics based upon size of Genbank.



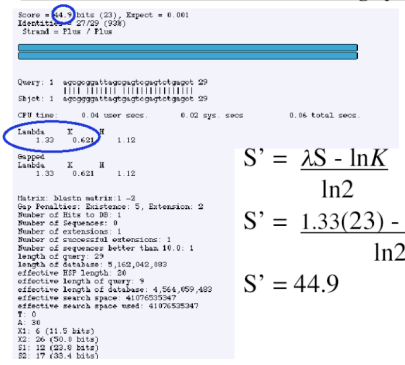
1. Match
2. Mismatch
3. Open gap
4. Extend gap
5. (Word size)

Raw score: sum values (27 matches, 2 miss-match: 27-4) = 23



What would score be if we used Match = +2, Mismatch = 0? Score = (27)(2) = 54. Therefore, raw score is uninformative unless scoring system is known.

Bit score: take into account scoring system

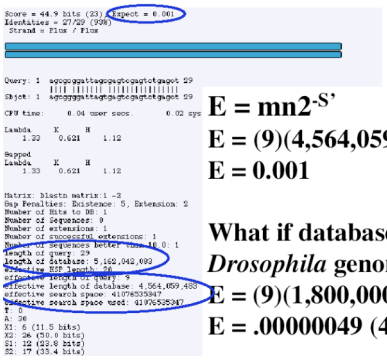


$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

$$S' = \frac{1.33(23) - \ln 0.621}{\ln 2}$$

$$S' = 44.9$$

Expect score: take into account search space



E = mn2$^{-S'}$

E = (9)(4,564,059,483)2$^{-44.9}$

E = 0.001

What if database was the *Drosophila* genome?

E = (9)(1,800,000)2$^{-44.9}$

E = .00000049 (4.9e^{-7})

Summary of statistics

- **Raw score**
  - Fist pass - no information because scoring system unknown
- **Bit Score**
  - Take into account the scoring system
- **E-value**
  - Takes into account the search space. Easy to interpret. ONLY USE E-VALUES!

BLAST against database: Important input parameters

- Database
- Filter
- (Expect)
- (Word size)
- Matrix (proteins)
- Gap open and extend

Affine Gap Penalty

- Usually do not need to change
- Penalty to open Gap
  - Default 5 for nucleotide, 11 for protein
- Penalty to extend gap
  - Default 2 for nucleotide, 1 for protein

Databases

- NR (non-redundant)
  - GenBank+EMBL+DDBJ+PDB
- ESTs
  - Expressed Sequence tags, organism subsets
- HTGS (High-Throughput Genomic Sequence)
  - Low quality, but may find new matches
- PAT (Patent)
- Vector
- Organism specific genomic sequences
- Month
  - Monthly updates to search for new hits

Filter

- Mask out low complexity regions (ATATATATATATATATA)
- Only masks Query sequence
- Can mask out repeats like Lines & Sines
- Mask out “lookup table”
- Masked regions appear as Xs or Ns in Query sequence

What types of searches to perform on protein coding regions?

Protein searches are more sensitive

- More characters to compare.
  - 20 amino acids versus 4 bases
- Substitution matrices.
- No “wobble” codon positions.
- Potentially smaller database size.

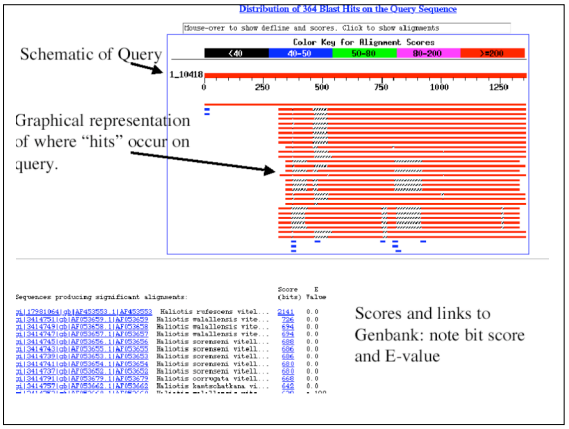
Protein Scoring Matrices

- Multiple matrices of each type optimized for particular divergence. Hiher number for greater divergence.
- PAM (PAM250)
  - Align all sequences of varying divergence and calculate probability of substitution between different amino acids
- BLOSSUM (BLOSSUM62)
  - Align blocks of sequences with varying divergence and calculate probability of substitution between different amino acids

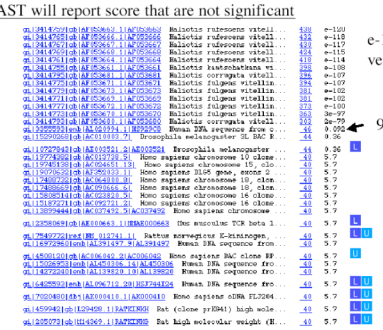
PAM250 substitution matrix
Matrices can be optimized for different distances
Newer matrices are BLOSSUM

		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
Cysteine	C	12																			
Hydrophilic	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														
Acid-amide	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										
Basic	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							
Hydro-phobic	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	−4	−3	−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			
Aromatic	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

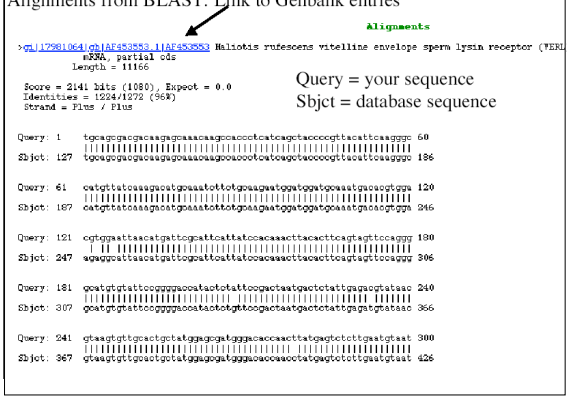
BLAST example



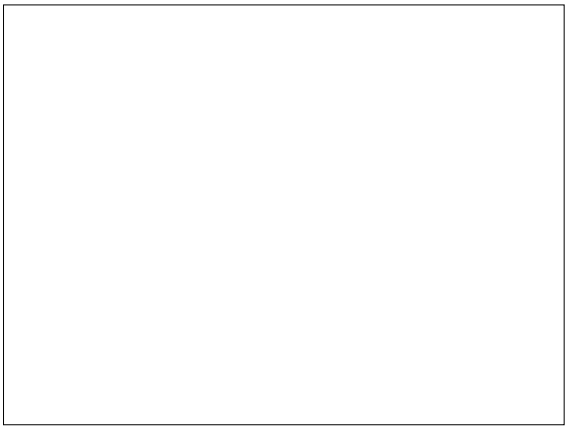
BLAST will report score that are not significant



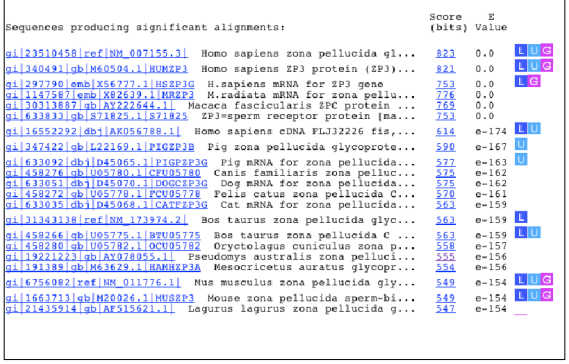
Alignments from BLAST: Link to Genbank entries



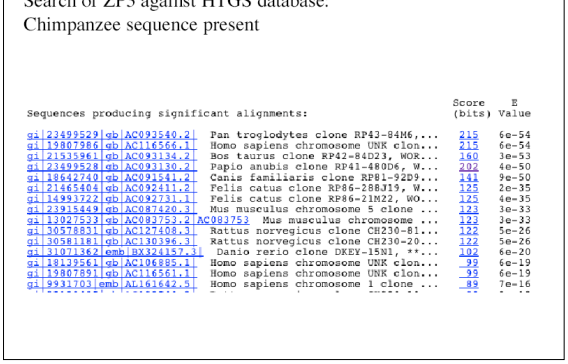
Example entry:

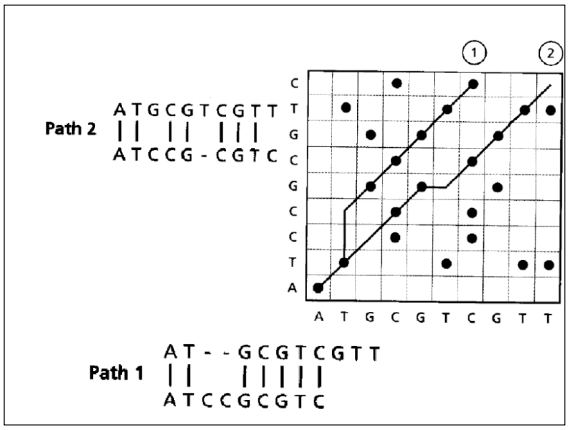


Search of ZP3 against NR database:
Notice no data from chimpanzee - but genome at 4x coverage

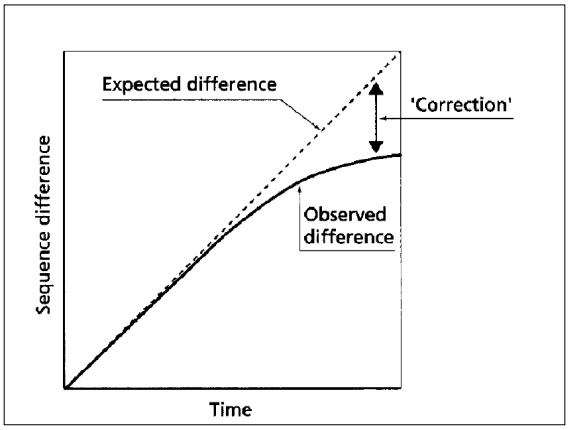


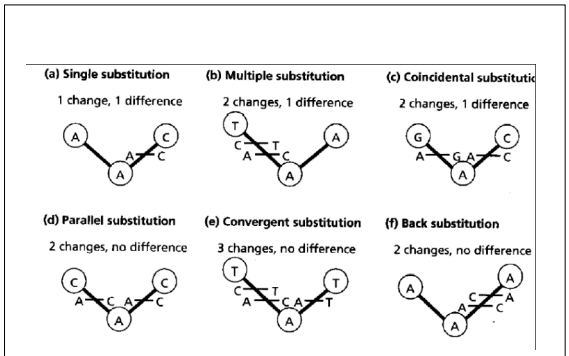
Search of ZP3 against HTGS database:
Chimpanzee sequence present





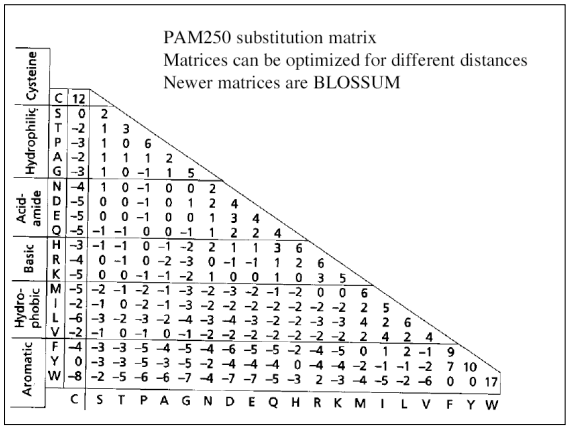
One thing we did not consider was multiple substitutions





More sensitive searches

- Profiles
  - Position Specific Scoring Matrix (PSSM)
- PSI-Blast
  - -automated PSSM
- Hidden Markov Models (HMMs)



Profile (position specific scoring matrix: PSSM)

- Made from alignment from one gene family.
- Substitution matrix for each position in alignment.
- Incorporates the types of changes that occur for that protein family at each position.
- Still uses general substitution matrix since the alignment would not cover all possible changes
- Changes gap penalty based upon alignment (remember ClustalW).

Alignment of sp18.

```
red18     MRSLVLLCVLLMAICAADK--------KSTVSKENAAAMKVAMIKFLDSRTDRFK-KRIE 51
white18   MRSLVLLCVLLMAICAADK--------KTTVSKENAAAMKIAMIKFLDARAGKFK-KRVE 51
asimi18   MRSLVLLCVLLMAICAADK--------KTSVSKENEAAMKVAMMKFLDMKAGVFK-EIIE 51
pink18    MRFLLLLCVLMGAVSQAVCRKRPNVWGKIVVKEKNKAAMKIGFMEYLDAKLVKFKRHWLV 60
green18   MRSLVLLCVLMAVGCVAFD--------DVVVSRQESYVQRGMVNFLDEEMHKLV-KRFR 51
          ** *:*****: . . *          .  *..:: : :: .::::** .   :  . .

red18     KIGYPITPPQYTTLLYYNRERLMDWCHNYVEVSKKIILLGGNKLNKKNFARMGRIIGWKN 111
white18   NMGYPITPPQWTTLLYYNRQRLMEWCHTYVEFSKKIILMGGNKLNKKNFTRMGRIIGWKN 111
asimi18   DMGYPITPPQWTTLLYYNRERLIEFCRSFLALSKKIILLGGNKLNKANFARMGRILGWKS 111
pink18    GANWKLQKFETDEMRYLAIKRLIKVCHGYTIWSQRLIMLKYRPLNEKYFKKVGRYLAWRN 120
green18   DMRWNLGPGFVFLLKKVNRERMMRYCMDYARYSKKILQLKHLPVNKKTLTKMGRFVGYRN 111
             : :       :     :*::  *  :   *:::: :    :*:  : ::** :.::.

red18     QWILKRRQWHMVR---VMRRYKASAIAKKIVAMKVADLPCN------ 149
white18   QWVLKRRQWEMVR---VMRRYKSTAIAKKIVAMKVADLPCN------ 149
asimi18   QWAVRQRQWGMVR---VSRRHTSTAIAKRIVAMKVADLPCN------ 149
pink18    YLIVFRMWIGVLKK--NLKRSEITKPMQKLLDTKDGELPCPVRKIHG 165
green18   YGVIRELYADVFRDVQGFRGPKMTAAMRKYSSKDPGTFPCKNEKRRG 158
```

Sp18 profile: compare to alignment

```
Cons  A   R   C   D   E   P   G   H   I   K   L   M   N   P   Q   R   S   T   V   W   X   Y   Z  Gap  Len  .
 M   -7   0  -9 -30 -20  16 -35 -13  25 -14  20  43 -22 -24 -10 -17 -14  -6  16 -10   0  -2   0 100  100
 E   -6   0 -22  -3   4 -32 -10   6 -24  27 -22 -17   3  -9  15  47  -3  -2 -20 -16   0 -18   0 100  100
 S    1   0  -1  -9  -9   0 -11  -1 -10  -8  -9  -5  -2  -7  -5 -10   7   4  -6 -13   0   0   0 100  100
 L  -12   0 -15 -40 -20  20 -44 -19  28 -21  40  20 -30 -22 -16 -22 -21 -13  10  -7   0   0   0 100  100
 V   -2   0  -4 -32 -21   6 -36 -19  30 -18  24  19 -24 -19 -15 -20 -13  -3  29 -20   0  -7   0 100  100
 L  -12   0 -15 -40 -20  20 -44 -19  28 -21  40  20 -30 -22 -16 -22 -21 -13  10  -7   0   0   0 100  100
 L  -12   0 -15 -40 -20  20 -44 -19  28 -21  40  20 -30 -22 -16 -22 -21 -13  10  -7   0   0   0 100  100
 C    5   0 135 -32 -30  -8 -20 -13 -11 -28 -15  -9 -18 -31 -24 -22   1  -5   0 -10   0  -5   0 100  100
 V    1   0   8 -29 -19   1 -31 -20  31 -17  10  16 -22 -18 -15 -20 -13   8  34 -26   0 -11   0 100  100
 L  -12   0 -15 -40 -20  20 -44 -19  28 -21  40  20 -30 -22 -16 -22 -21 -13  10  -7   0   0   0 100  100
 M   -9   0 -11 -34 -23  17 -38 -15  26 -16  33  36 -25 -23 -12 -19 -16  -8  16  -8   0  -1   0 100  100
 M    5   0  -7 -13 -10 -14   5 -11  -4 -10  -4   6  -8 -13  -7 -11  -1  -3  -2 -16   0 -10   0 100  100
 A   17   0   3 -10  -5 -15  -6 -11   3  -7  -3   0  -8  -3  -5 -10   4   4  16 -33   0 -10   0 100  100
 I   -1   0 -10 -23 -19 -10  -8 -19  12 -16   3   4 -16 -20 -15 -18  -9  -5  12 -26   0 -17   0 100  100
 C    6   0  82 -21 -20 -13 -13  -9 -12 -19 -16 -10 -10 -21 -16 -16   6   0  -2 -16   0  -9   0 100  100
 A    9   0  -4  -7   0 -16 -10  -5   8  -2  -4  -1  -5  -4   2  -4   2   2   6 -30   0 -17   0 100  100
 A   24   0   5  -3   0 -22   5  -8  -8  -4 -12  -7  -3   3  -2  -6  11   6   1 -36   0 -22   0 100  100
 D   -7   0 -15  -1  -5   2 -24  -4  -4 -12  -5  -3  -6 -19  -8 -16  -9  -6  -2 -18   0   0   0 100  100
 K   -1   0  11   6   4 -29 -10   0 -23   6 -24 -17   4 -13   2   4   2  -1 -15 -32   0 -10   0 100  100
 R   -3   0 -29  17  16 -36  -7   5 -26  24 -26 -18  11  -6  13  18   2   8 -20 -40   0 -21   0  24   24
 V    1   0  -4 -18 -13  -6 -25 -13  16 -10   6   8 -11 -11  -9 -13   8   7  17 -27   0 -13   0 100  100
 V    3   0  -1 -16 -10  -9 -21 -12  13  -9   3   5 -10  -9  -8 -12   1   8  18 -29   0 -14   0 100  100
 V    1   0   8 -29 -19   1 -33 -20  31 -17  10  16 -22 -18 -15 -20 -13   8  34 -26   0 -11   0 100  100
 S    6   0  -7   5   4 -29   0   0 -16   9 -21 -14   8   1   5   6  14  11 -33   0 -19   0 100  100
 K   -3   0 -26   8  16 -34  -9   5 -23  24 -23 -16   0  -6  15  26   8   8 -31   0 -11   0 100  100
 H   -1   0 -27  15  23 -33  -9   6 -22  18 -12 -15   8  -4  19  13   1   8 -17 -35   0 -12   0 100  100
 N   -2   0 -21  33  16 -33   0   9 -27   9 -29 -11  29  -7   8   3   6   3 -21 -38   0 -17   0 100  100
 Q    4   0 -10   7  14 -29  -6   3 -16  14 -10 -11   5  -2  14  11   4   1 -12 -34   0 -21   0 100  100
 A   20   0   3   0   0 -24   4  -6 -16  -2 -14  -9   0   3   0  -4  14   8  -2 -35   0 -21   0 100  100
 A   10   0   2 -10  -7  -1  -9   0  -7  -9   8  -5  -6  -7  -6  -9   2  -1  -2 -13   0   7   0 100  100
 M   -4   0  -4 -29 -19  11 -34 -15  26 -14  25  35 -22 -22 -11 -17 -12  -4  21 -14   0  -4   0 100  100
 K   -3   0 -26   6  13 -30 -10   7 -28  26 -19 -12   7  -4  18  23   1   8 -16 -32   0 -19   0 100  100
 I   -4   0 -10 -24 -15  -5 -31 -13  18  -5  10   9 -17 -18  -7  -1 -10  -3  16 -19   0 -11   0 100  100
 G   12   0  -9   0  -4 -39  40 -11 -29  -8 -10 -23   1   8  -6  -8   6  -3 -18 -38   0 -32   0 100  100
 M  -11   0  -8 -34 -25  31 -39  -9  26 -10  25  35 -24 -27 -14 -21 -17 -10  11   3   0  13   0 100  100
 M   -4   0  -6 -31 -21  10 -36 -17  36 -16  25  30 -23 -22 -13 -19 -13  -4  25 -16   0  -5   0 100  100
 K   -2   0 -25  16  17 -34  -5   7 -24  19 -25 -18  17  -6  13  13   3   1 -19 -37   0 -30   0 100  100
 F  -22   0  -7 -40 -35  44 -45   5   5 -29  14  10 -26 -36 -23 -28 -25 -21  -2  37   0  58   0 100  100
 L  -12   0 -15 -40 -28  20 -44 -19  28 -21  40  28 -30 -22 -16 -22 -21 -13  18  -7   0   0   0 100  100
```

Compare the difference between Blast raw score using PAM250 and ProfileSearch raw score

Compute raw score using PAM250 matrix (Blast)

```
Seq 1: Lys Ser Thr Val Ser Lys
Seq 2: Asp Val Val Val Ser Arg
```

Compute raw score using PAM250 matrix (Blast)

```
Seq 1: Lys Ser Thr Val Ser Lys
Seq 2: Asp Val Val Val Ser Arg


PAM 250(Seq. 1 vs 2): 0
```

Compute raw score using PAM250 matrix

```
Seq 1: Lys Ser Thr Val Ser Lys
Seq 2: Asp Val Val Val Ser Arg

PAM 250(Seq. 1 vs 2): 0  -1
```

Compute raw score using PAM250 matrix

```
Seq 1: Lys Ser Thr Val Ser Lys
Seq 2: Asp Val Val Val Ser Arg

PAM 250(Seq. 1 vs 2): 0  -1  0
```

Compute raw score using PAM250 matrix

```
Seq 1: Lys Ser Thr Val Ser Lys
Seq 2: Asp Val Val Val Ser Arg

PAM 250(Seq. 1 vs 2): 0  -1  0  4  2  3  = 8

What is needed to compute bit score?
What is needed to compute E-value?
```

Alignment of sp18.

```
red18     MRSLVLLCVLLMAICAADK--------KSTVSKENAAAMKVAMIKFLDSRTDRPK-KRIE 51
white18   MRSLVLLCVLLMAICAADK--------KTTVSKENAAAMKIAMIKFLDARAGKPK-KRVE 51
asimi18   MRSLVLLCVLLMAICAADK--------KTSVSKENEAAMKVAMMKFLDMKAGVPK-EIIE 51
pink18    MRFLLLLCVLMGAVSQAVCRKRPNVWGKIVVKEKNKAAMKIGFMEYLDAKLVKFKRHWLV 60
green18   MRSLVLLCVLMAVGCVAFD--------DVVVSRQEQSYVQRGMVNFLDEEMHKLV-KRPR 51
          ** *:*****: . . *          .  *..:: : :: .::::** .   :  . .

red18     KIGYPITPPQYTTLLYYNRERLMDWCHNYVEVSKKIILLGGNKLNKKNFARMGRIIGWKN 111
white18   NMGYPITPPQWTTLLYYNRQRLMEWCHTYVEFSKKIILMGGNKLNKKNFTRMGRIIGWKN 111
asimi18   DMGYPITPPQWTTLLYYNRERLIEFCRSFLALSKKIILLGGNKLNKANFARMGRILGWKS 111
pink18    GANWKLQKFETDEMRYLAIKRLIKVCHGYTIWSQRLIMLKYRPLNEKYFKKVGRYLAWRN 120
green18   DMRWNLGPGFVFLLKKVNRERMMRYCMDYARYSKKILQLKHLPVNKKTLTKMGRFVGYRN 111
             : :       :    :*::  *  :   *:::: :    :*:  : ::** :.::.

red18     QWILKRRQWHMVR---VMRRYKASAIAKKIVAMKVADLPCN------ 149
white18   QWVLKRRQWEMVR---VMRRYKSTAIAKKIVAMKVADLPCN------ 149
asimi18   QWAVRQRQWGMVR---VSRRHTSTAIAKKRIVAMKVADLPCN------ 149
pink18    YLIVFRMWIGVLKK--NLKRSEITKPMQKLLDTKDGELPCPVRKIHG 165
green18   YGVIRELYADVFRDVQGFRGPKMTAAMRKYSSKDPGTPPCKNEKRRG 158
```

Sp18 profile: compare to alignment

```
Cons  A   B   C   D   E   F   G   H   I   K   L   M   N   P   Q   R   S   T   V   W   X   Y   Z  Gap  Len  .
 M   -7   0  -5  -30  -20  16  -35  -13  25  -14  20  43  -22  -24  -10  -17  -14  -6  16  -10   0  -2   0  100  100
 R   -6   0  -22  -3   4  -32  -10   6  -24  27  -22  -17   5   -9   15   47   -1  -2  -20  -16   0  -10   0  100  100
 S   1    0  -1   -9   -9   0  -11  -1  -16  -9   -9  -5   -2  -7   -5  -10   7    4  -6  -13   0   0    0  160  100
 L  -12   0  -15  -40  -28  20  -44  -19  28  -21  40  30  -30  -22  -16  -22  -21  -13  10   -7   0   0    0  160  100
 V  -2    0  -4   -32  -21   6  -36  -19  30  -18  24  19  -24  -19  -15  -20  -13  -3   29  -20   0  -7   0  160  100
 L  -12   0  -15  -40  -28  20  -44  -19  28  -21  40  30  -30  -22  -16  -22  -21  -13  10   -7   0   0    0  160  100
 L  -12   0  -15  -40  -28  20  -44  -19  28  -21  40  30  -30  -22  -16  -22  -21  -13  10   -7   0   0    0  160  100
 C   5    0  115  -32  -30  -8  -20  -13  -11  -28  -15  -9  -16  -31  -24  -22   1   -5   0  -10   0  -5   0  160  100
 V   1    0   8   -29  -19   1  -33  -20  31  -17  18  16  -32  -18  -15  -20  -18   8   34  -26   0  -11   0  160  100
 L  -12   0  -15  -40  -28  20  -44  -19  28  -21  40  30  -30  -22  -16  -22  -21  -13  10   -7   0   0    0  160  100
 M   -9   0  -11  -34  -23  17  -38  -15  24  -16  33  36  -25  -23  -12  -19  -16  -8   16   -8   0  -1   0  160  100
 K   5    0  -7   -13  -10  -14   5  -11  -4  -10  -4   6   -8  -15  -7   -11  -1   -3  -2  -26   0  -10   0  160  100
 A   17   0   3   -10   -5  -15  -6  -11   3   -7   -3   0   -8   -3   -5  -10   4    4   16  -33   0  -10   0  160  100
 I   -1   0  -10  -23  -19  -10  -8  -13  12  -16   3    4  -16  -20  -15  -18   -9  -5   12  -26   0  -17   0  160  100
 C   6    0   82  -21  -20  -13  -13   -9  -12  -19  -16  -10  -10  -21  -16  -16   6    0   -2  -16   0   -8   0  160  100
 A   9    0   -4   -7    0  -16  -10   -5   8   -2   -4   -1   -5   -4   2    -4   2    2    6  -30   0  -17   0  160  100
 A   24   0   5    -3    0  -22   5   -8   -8   -4  -12   -7   -3   5    -2   -6   11   -6   1   -36   0  -22   0  100  100
 D   -7   0  -15   -1   -5   2  -24   -4   -6  -12   -5   -3   -6  -19   -8   -16   -9   -6   -2  -18   0   0    0  100  100
 K   -1   0   11    6    4  -29  -10   0  -23    6  -24  -17   4   -13   2    -4   2    -1  -15  -32   0  -18   0  100  100
 k   -3   0  -29   17   16  -36   -7   5  -26   24  -26  -18   12   -6   13   18   2    8   -20  -40   0  -23   0   24   24
 V   1    0   -4   -18  -13   -6  -25  -13  16  -10   6    9   -11  -11   -9   -13   8    7   17  -27   0  -13   0  100  100
 V   3    0   -1   -16  -10   -9  -21  -12  13   -9   3    5   -10   -9   -8   -12   1    8   18  -29   0  -14   0  100  100
 V   1    0   8   -29  -19   1  -33  -20  31  -17  18  16  -22  -18  -15  -20  -18   8   34  -26   0  -11   0  160  100
 S   6    0   -7   5    4  -29   0   0  -18   9  -21  -14   8    1    5    6   16   11  -13  -33   0  -19   0  100  100
 K   -3   0  -26   0   16  -34   -9   5  -23  24  -23  -16   6    -6   15   26   8    8   -18  -31   0  -21   0  100  100
 H   -1   0  -27   15   23  -33   -9   6  -22  18  -22  -15   8    -4   19   13   1    8   -17  -35   0  -12   0  100  100
 N   -2   0  -21   23   16  -33   0   9  -27   9  -29  -21   29   -7   9    3    6    3   -21  -38   0  -17   0  100  100
 Q   4    0  -18   7   14  -29   -6   3  -18  14  -18  -11   5    -2   14   11   4    1   -12  -34   0  -21   0  100  100
 A   20   0   3    0    0  -24   4   -6  -18   -2  -14   -9   0    3    0    -4   14   8   -2  -35   0  -21   0  100  100
 A   10   0   2   -10   -7   -1   -8   0   -7   -9   -8   -5   -6   -7   -6   -9   2    -1   -2  -13   0   7    0  100  100
 M   -4   0   -6  -29  -19   11  -34  -15  26  -14  25   35  -22  -22  -11  -17  -12  -4   21  -14   0   -4   0  100  100
 K   -3   0  -26   6   13  -30  -10   7   -26  26  -19  -12   7    -4   18   23   1    8   -16  -32   0  -19   0  100  100
 I   -4   0  -10  -24  -15   -5  -31  -13  18   -5   10   9   -17  -18   -7   -1   -10  -3   16  -19   0  -11   0  100  100
 G   12   0   -9   0   -4  -39  40  -11  -29   -8  -30  -23   1    -8   -6   -8    6   -3  -18  -38   0  -32   0  100  100
 M   -11  0   -8  -34  -25  31  -39   -9  20  -19  25   35  -24  -27  -14  -21  -17  -10  11    3    0  13    0  160  100
 M   -4   0   -6  -31  -21  10  -36  -17  38  -16  25   30  -23  -22  -13  -19  -13  -4   25  -16   0   -5   0  100  100
 K   -2   0  -25   16   17  -34   -5   7   -24  19  -25  -18  17    -6   13   13   3    1   -10  -37   0  -20   0  100  100
 P   -22  0   -7  -40  -35  64  -48   5   5   -29  14   10  -26  -36  -23  -29  -25  -21  -2   37   0   58   0  160  100
 L   -12  0  -15  -40  -28  20  -44  -19  28  -21  40   30  -30  -22  -16  -22  -21  -13  10   -7   0   0    0  160  100
```

Compute each sequence comparison to Profile

```
Pos. : 1   2   3   4   5   6
Seq 1:Lys Ser Thr Val Ser Lys
Seq 2:Asp Val Val Val Ser Arg

Profile vs. Sequence 1: 24
Profile vs. Sequence 2: 17
```

Compute each sequence comparison to Profile

```
Pos. : 1   2   3   4   5  6
Seq 1:Lys Ser Thr Val Ser Lys
Seq 2:Asp Val Val Val Ser Arg

Profile vs. Sequence 1:24 0
Profile vs. Sequence 2:17 17
```

```
Pos. : 1   2   3   4   5  6
Seq 1:Lys Ser Thr Val Ser Lys
Seq 2:Asp Val Val Val Ser Arg

PAM 250(Seq. 1 vs 2)  : 0  -1 0  4   2  3  = 8
Profile vs. Sequence 1:24  0  8  34 16 24  = 106
Profile vs. Sequence 2:17  17 18 34 16 26  = 128
```

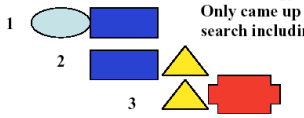
Look at alignment and see why scores are different for the 3 valines

Psi-Blast

- Use BlastP to search protein database.
- Blast makes alignments with all hits with an E-value lower than given value (default e^{-3}).
- Calculates position specific scoring matrix and compares to the database.
- Single gap penalty (different from profiles)
- New "hits" are used to recalculate PSSM and redo the search.
- Searches continue to be done iteratively until no new hits are found.

Problems with iterative searching with Psi-Blast

- You do not make the multiple alignment. There could be errors in the alignment.
- Can pick up significant hits due to shared domains.



No homology between 1 and 3. Only came up from iterative search including 2.

Summary

- Blastp may miss some significant matches.
- Using evolutionary information in the form of a multiple alignment increases sensitivity of database searching.

1. What kind of search does BlastX perform? Why would you use BlastX?
1. Describe a scoring matrix (i.e. PAM250).
2. A blast score has an E-value of e^{-3}, is this statistically significant? Is it biologically significant?
3. I want to search for EST sequences, should I search the DNA nr database?
4. BlastP replaced part of my sequence with X's, why did it do that?
5. Name of few databases you would search using BlastN.
6. Describe the type of search that is performed by BlastN, BlastP, and BlastX?
7. What does an E-value from blast mean?
8. Define homologous, orthologous, and paralogous.
9. You have a protein coding sequence and want to compare it to the database. Should you search the protein or DNA sequence? Why?
10. You are told a sequence alignment has a raw score of 1000. Is this a significant match?
11. What does the filter do in a Blast search?
12. The bit score includes the parameters k and lamda (λ). Why are these parameters in the equation to calculate bit scores?
13. The E-value score includes parameters m and n. Why are these parameters in the equation to calculate E-values?
14. What is a position specific scoring matrix?
15. How does Psi-Blast differ from BLASTP?