

## Lesson 6 -Key

### 1. BLASTP search

B. Examine the BLAST output:

#### **BLAST output:**

```
Sequences producing significant alignments: (bits) Value
sp|P48013|RAD9_SCHOT DNA REPAIR PROTEIN RAD9 106 8e-23
sp|P26306|RAD9_SCHPO DNA REPAIR PROTEIN RAD9 101 3e-21
sp|P09627|PMA1_SCHPO PLASMA MEMBRANE ATPASE 1 (PROTON PUMP) 31 3.3
sp|Q63504|NRD2_RAT ORPHAN NUCLEAR RECEPTOR NR1D2 (REV-ERB-B... 31 5.7
sp|P35658|N214_HUMAN NUCLEAR PORE COMPLEX PROTEIN NUP214 (... 31 5.7
sp|O02755|CEBB_BOVIN CCAAT/ENHANCER BINDING PROTEIN BETA (C... 30 9.9
```

1. Go back to Netscape.
2. How many hits are significant? **2**
3. Click on the name of each of the best hits and save the resulting sequence file to your diskette.
4. What biological function does our unknown human protein perform? **It is a DNA radiation damage repair protein.**
5. What biochemical function does our unknown human protein have? **The blast search reveals nothing about the protein's biochemical function.**

D. Compare the alignment given by Bestfit to the alignment given with the BLAST program.

#### **Blast alignment:**

```
sp|P48013|RAD9_SCHOT DNA REPAIR PROTEIN RAD9
Length = 432
```

```
Score = 106 bits (262), Expect = 8e-23
Identities = 81/309 (26%), Positives = 143/309 (46%), Gaps = 37/309 (11%)
```

```
Query: 1 MKCLVTGGNVKVLGKAVHLSRIGDELYLEPLEDGLSLRTVNSSRSAYACFLFAPLFFQQ 60
M+ +V+ N++ L + +LSRI D + E +D L L T+NSSRS + FF +
Sbjct: 1 MEFVVSNTNLRDLSRIFLNLRSRIDDAVNWEINKDQLILTTLNSSRSGFGKVTTLTKKFFDK 60

Query: 61 YQAATPGQDL-----LRCKILMKSFLSVFRS-----LAMLEKTVEKC 97
+ L +R +K LS+FR+ E + +K
Sbjct: 61 FTFHPDTLFLTGFVSPTVRLSTQIKPILSIFRNKIFESTLLVNNNLNTNAGAAESSKKN 120
```

Query: 98 CISLN-----GRSSRLVVQLHCKFGVRKTHNLSFQDCESLQAVFDPASCPHMLRAPAR 150  
+ N G+ R++ + +CK GV KT+ +++++ ++L AVFD ASC + + ++  
Sbjct: 121 VVVENIQMQITSGKECRVIFKFNCKHGVVVKTYKIAYEQTQTLHAVFDKASCHNNWQINSK 180

Query: 151 VLGEAVLPFSPALAEVTXXXXXXXXXXXXXXXXSYHEE---EADSTAKAMVTEMCLGEEDFQQL 207  
+L + + F E+T S+ EE D + T + + ++F+Q+  
Sbjct: 181 ILKDLIEHFGQKTEELT-IQPVQGRVLLTSFTEEVVHNKDVLKQPTQTTVSIDGKEFEQV 239

Query: 208 QAQEGVAITFCLKEFRGLLSFAESANLNLSIHFDAPGRPAIFTI---KDSLDDGHFVLAT 264  
EG+ IT LKEFR + AES +++ ++ G+PA+FT K ++ F+LAT  
Sbjct: 240 SLNEGIIITLSLKEFRAAVLLAESLGTSIASYYSVSGKPALFTFNKGKFMEIEAQFILAT 299

Query: 265 LSDTDSHSQ 273  
+ D +  
Sbjct: 300 VMGPDDDFDE 308

**Bestfit alignment (with default parameters):**

```

1 MKCLVTGGNVKVLGKAVHSLSRIGDELYLEPLEDGLSLRTVNSSRSAYAC 50
|. .|. |.: | : .|||| | . | .| | | |.||||| :
1 mefvvsntnlrdlsriflnlsriddavnweinkdqlilttl nssrsgfgk 50
.
51 FLFAPLFFQOY.....QAATPG..QDLLRCKILMKSFLSVFRSLAMLEKT 93
|| :. : | .| .| ||:|. . | |
51 vtltkffdkftfhpdtlfltgvspvrlstqikpilsifrn.kifest 99
.
94 .....VEKCCISL.NGRSSRLVVQLHCKFGVR 119
|| . : .|: |: . .|| ||
100 llvnnlnltnagaaessskknvveniqmqitsgkecrvifkfnckhgv 149
.
120 KTHNLSFQDCESLQAVFDPASCPHMLRAPARVLGEAVLPFSPALAEVTLG 169
||: :.: : .| |||| ||| . . :.: : | |. |:
150 ktykiayeqtqtlhavfdkaschnnwqinskilkdliehfggkteeltiq 199
.
170 IGRGRRVILRSYHEE...EADSTAKAMVTEMCLGEEDFQQLQAQEGVAIT 216
.| ||:| |: || | . | . : .:|. | |: ||
200 pvqg.rvlltsfteevvhnkdvlkqptqttsidgkefeqvslnegiit 248
.
217 FCLKEFRGLLSFAESANLNLSIHFDAPGRPAIFTI...KDSLDDGHFVLA 263
||||| . ||| .: .: |:|:| | | :. |:|
249 lslkefraavllaeslgtstiasyysvsgkpalftfnkgkfmeieaqfila 298
.
264 TLSDTDSHSQDLGSPERHQVPVQLQAHSTPHPDFA...NDDIDSYMIAM 310
|. | : | . | | .: | |
299 tvmgpddfdesslgarwqsgtansllvpentsaapaleneapsasigw 348
.
311 ETTIGNEGSRVLPISLSPGPQPPKSPGPHSEEEDEAEPTVPGTP 356
:| | ||. | | : | . :. | | . | |
349 qtngdaetsrmfhstldiprneepaakpsrqttddeenhpflflegmp 394

```

1. If the two alignments are different, why? **They are different because Gapped Blast does not do a full Smith-Waterman alignment.**
2. What does this imply about the accuracy of BLAST2 alignments? **See answer to question 1.**
3. What should be done to improve upon BLAST2 alignments? **A full Smith-Waterman alignment should be performed with Bestfit with optimized parameters.**
4. Further test the statistical significance of the match using PRSS. Is it a match?

**PRSS3 output:**

PRSS3 - 1000 shuffles; uniform shuffle  
 unshuffled s-w score: 417; Z(417,432): 512.8 p(417) < 9.36335e-27  
 For 1000 sequences, a score >= 417 is expected 9.363e-24 times

$$E=D*P = 84,629*9.36335e-27 = 7.92411E-22$$

*As the E is very low, the match is statistically significant. Note that this more sophisticated estimate of E is very closely approximated by the Blast estimate.*

E. Repeat the search against the nonredundant protein database and save the output as above.

**BLAST output:**

Sequences producing significant alignments:	(bits)	Value
ref NP_004575.1   RAD9 (S. pombe) homolog >gi 1765956 gb AA...	723	0.0
ref NP_035367.1   RAD9 homolog (S. pombe) >gi 3869272 gb AA...	601	e-171
gb AAD31691.1 AF124502_1 (AF124502) DNA repair protein Rad9...	129	3e-29
sp P48013 RAD9_SCHOT DNA REPAIR PROTEIN RAD9 >gi 1085808 pi...	106	3e-22
sp P26306 RAD9_SCHPO DNA REPAIR PROTEIN RAD9 >gi 101067 pir...	101	2e-20
pir  S26143 rad9 protein (allele rad9-192) - fission yeast ...	100	3e-20

1. Does the nonredundant database help you identify the biochemical function of the protein? **No. The best hit is really only a slightly different version of our query sequence from the same lab. None of the statistically significant hits have a known biochemical function.**
2. Do the same proteins have the same E() values? If not explain why? **No, they don't. E values depend on database size, and since there are more residues (and sequences) in the nonredundant database than in Swissprot, the corresponding E() values are higher in the nonredundant database than in Swissprot.**

## 2. BLASTN search

A. Compare the nucleic acid sequence to the nonredundant nucleic acid database using the BLASTN program.

Score	E		(bits)	Value
Sequences producing significant alignments:				
ref NM_004584.1		Homo sapiens RAD9 (S. pombe) homolog (RAD...	2109	0.0
gb U53174.1 HSU53174		Human cell cycle checkpoint control pr...	2109	0.0
ref NM_011237.1		Mus musculus RAD9 homolog (S. pombe) (Rad...	753	0.0
gb AF045663.1 AF045663		Mus musculus radio-resistance/chemo-...	753	0.0
gb S57501.1 S57501		protein phosphatase type 1 catalytic sub...	367	9e-99
ref NM_002708.1		Homo sapiens protein phosphatase 1, catal...	359	2e-96
emb X70848.1 HSPH1CAT		H.sapiens mRNA for phosphatase 1 cata...	359	2e-96
gb M63960.1 HUMPRPHOS1		Human protein phosphatase-1 catalyti...	355	4e-95
gb AF045662.1 AF045662		Mus musculus cell cycle checkpoint c...	170	1e-39
dbj D00859.1 RATPP1ACS		Rattus norvegicus PP-1 alpha mRNA fo...	159	4e-36
dbj D90163.1 RATPP1AA		Rattus norvegicus mRNA for catalytic ...	159	4e-36
dbj AB005817.1 AB005817		Homo sapiens DNA containing CpG is...	157	2e-35
gb S78215.1 S78215		protein phosphatase 1 alpha [rats, stria...	147	2e-32
emb Y00701.1 OCCPP1		Rabbit mRNA for protein phosphatase-1 ca...	105	5e-20
emb X07798.1 OCCPP1A		Rabbit mRNA for type-1 protein phospho...	105	5e-20
dbj D17240.1 HUMD4F12M3		Human HepG2 3' region MboI cDNA, cl...	105	5e-20
gb AF167082.2 AF167082		Homo sapiens neuronal potassium chan...	44	0.17
gb AC004030.1 AC004030		Homo sapiens DNA from chromosome 19,...	44	0.17
gb AC005233.2 AC005233		Homo sapiens PAC clone RP5-1198021 f...	40	2.7
gb AC004923.2 AC004923		Homo sapiens PAC clone RP5-901A4, co...	40	2.7
gb AE002050.1 AE002050		Deinococcus radiodurans R1 section 1...	40	2.7
gb AC007977.11 AC007977		Drosophila melanogaster, chromosome...	40	2.7
gb AC005962.1 AC005962		Homo sapiens chromosome 17, clone hR...	40	2.7
gb AC004757.1 AC004757		Homo sapiens chromosome 17, clone hC...	40	2.7
emb Z82076.1 CEW07G1		Caenorhabditis elegans cosmid W07G1, c...	40	2.7
emb Z99569.1 HS106H14		Human DNA sequence from PAC 106H14 on...	40	2.7
dbj D00521.1 ABCALDH		Acetobacter polyoxogenes gene for alde...	40	2.7
gb M81255.1 PH2HEADTL		Bacteriophage 21 head gene operon	40	2.7
emb Y08696.1 AEAHDH		Acetobacter europaeus gdhA, aldF, aldG ...	40	2.7

**Here is a sample alignment with a phosphatase**

gb|S57501.1|S57501 protein phosphatase type 1 catalytic subunit [human, mRNA, 1400 nt]  
 Length = 1400

Score = 367 bits (185), Expect = 9e-99  
 Identities = 206/209 (98%), Gaps = 3/209 (1%)  
 Strand = Plus / Minus

Query: 1876 gagaatccagctttgacctttattcaagagaccagatggggttgccccaggatccgg-ctg 1934  
 |||  
 Sbjct: 1388 gagaatccagctttgacctttattcaagagaccagatggggttgccccaggatccgggctg 1329

```

Query: 1935 ccagccctgaggccaagcacggctggagacccacgacctggcctgccggtgcctgagct 1994
          ||||||| |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1328 ccagcc-tgaggccaagcacggctggagacccacgacctggcctgccggtgcctgagct 1270

Query: 1995 gcagcctcggccccaggatcctgctcacagtcaccgcaggtgcaggcaggaagcagccct 2054
          ||||||||||||||||||||||||||||||| |||||||||||||||||||||||||||
Sbjct: 1269 gcagcctcggccccaggatcctgctcacagt-accgcaggtgcaggcaggaagcagccct 1211

Query: 2055 gggggactggacgctgctattgattcatt 2083
          |||||||||||||||||||||||||||
Sbjct: 1210 gggggactggacgctgctattgattcatt 1182

```

1. What categories do the statistically significant hits fall into? ***They are either rad9s (cell cycle checkpoint control proteins) or phosphatases.***
2. Note the reading sense of the phosphatase genes. Since the query nucleic acid sequence is based on mRNA, it represents the correct reading sense. ***The reading frames of the phosphatases are in reverse.***
3. Is our protein a phosphatase? ***No, the gene for a phosphatase, READ IN THE REVERSE DIRECTION, coexists with our query sequence gene, READ IN THE FORWARD DIRECTION, over some of its length.***

**3. TBLASTN search**

Compare the protein sequence to the nonredundant nucleic acid database with TBLASTN. Does this search tell you anything new?

Score	E		(bits)	Value
Sequences producing significant alignments:				
ref NM_004584.1		Homo sapiens RAD9 (S. pombe) homolog (RAD...	723	0.0
gb U53174.1 HSU53174		Human cell cycle checkpoint control pr...	723	0.0
ref NM_011237.1		Mus musculus RAD9 homolog (S. pombe) (Rad...	601	e-170
gb AF045663.1 AF045663		Mus musculus rad9 gene	601	e-170
gb AF045662.1 AF045662		Mus musculus cell cycle checkpoint c...	132	8e-33
gb AF124502.1 AF124502		Drosophila melanogaster DNA repair p...	129	1e-28
gb AF076846.1 AF076846		Drosophila melanogaster Rad9-like pr...	129	1e-28
emb AL136235.1 SPAC664		S.pombe chromosome I cosmid c664	77	1e-12
emb X64648.1 SPRAD9G		S.pombe rad9 gene	77	1e-12
emb X58231.1 SPRAD9		S.pombe rad9 gene	77	1e-12
emb X77276.1 SMRAD9		S.malidevorans rad9 gene	77	1e-12
emb X65864.1 SPRADG		S.pombe gene rad9-192	76	2e-12
emb X77277.1 SORAD9		S.octosporus rad9 gene	46	3e-08
gb AC002350.1 AC002350		Homo sapiens 12q24 PAC RPCI3-424M6 (...)	48	6e-04
emb AL031633.1 CEY39A1A		Caenorhabditis elegans cosmid Y39A1...	43	0.021

**No it doesn't. Only previously found genes are found. Accession numbers** AC002350 and CEY39A1A are statistically significant but are raw genomic DNA which has not been split up into genes.

#### 4. BLASTX search

Compare the nucleic acid sequence to the nonredundant protein database. Does this search tell you anything new?

Sequences producing significant alignments:	(bits)	Value
ref NP_004575.1   RAD9 (S. pombe) homolog >gi 1765956 gb AA...	723	0.0
ref NP_035367.1   RAD9 homolog (S. pombe) >gi 3869272 gb AA...	601	e-171
gb AAD31691.1 AF124502_1 (AF124502) DNA repair protein Rad9...	129	7e-29
sp P48013 RAD9_SCHOT DNA REPAIR PROTEIN RAD9 >gi 1085808 pi...	106	7e-22
sp P26306 RAD9_SCHPO DNA REPAIR PROTEIN RAD9 >gi 101067 pir...	101	3e-20
pir S26143 rad9 protein (allele rad9-192) - fission yeast ...	100	7e-20
emb CAA18514.1  (AL022374) hypothetical protein SC5B8.06 [S...	40	0.063
ref NP_032914.1   paired mesoderm homeobox 2b >gi 6093753 s...	40	0.082
ref NP_003915.1   paired mesoderm homeobox 2b >gi 1841338 d...	40	0.082

**No it doesn't. Only previously found genes are found. The homeobox genes would require lower E values to be credible,**

#### 6. BLASTP search against C. elegans database. No significant hits.

*There is no new information available from the more sensitive search methods in this case. However, they are still worth learning how to do. Note that in the Fasty3 search all of the apparent hits (upon visual inspection) contain significant low-complexity regions.*

Key to written problems:

$$1. E = KmNe^{-s}$$

$$= .219$$

$$K = .082$$

$$s = 103$$

$$m = 100$$

$$s = .219 \times 103 = 22.6$$

$$e^{-s} = 1.6 \times 10^{-10}$$

$$Kme^{-s} = .082 \times 100 \times 1.6 \times 10^{-10} = 1.3 \times 10^{-9}$$

$$\text{For Swissprot, } N = 2.1 \times 10^7$$

$$E = KmNe^{-s} = 2.1 \times 10^7 \times 1.3 \times 10^{-9} = .028$$

For nonredundant,  $N = 1.33 \times 10^8$

$$E = KmNe^{-s} = 1.33 \times 10^8 \times 1.3 \times 10^{-9} = .17$$

$$2. \quad S = - \sum_{i=1}^N \frac{n_i}{L} \log_2 \frac{n_i}{L}$$

$$L = 5$$

$$n(i) = 1$$

$$n(s) = 1$$

$$n(a) = 2$$

$$n(c) = 1$$

$$\begin{aligned} S &= -[1/5 \log_2(1/5) + 1/5 \log_2(1/5) + 2/5 \log_2(2/5) + 1/5 \log_2(1/5)] \\ &= -[3/5 \log_2(1/5) + 2/5 \log_2(2/5)] \\ &= -[.6 \log_2(.2) + .4 \log_2(.4)] \\ &= -[.6(-2.32) + .4(-1.32)] \\ &= -[-1.39 - .53] \\ &= -[-1.92] \\ &= 1.92 \end{aligned}$$