Harvard-MIT Division of Health Sciences and Technology
HST.508: Genomics and Computational Biology

Problem Set 2

Please make sure to show your work and calculations and state any assumptions you make in answering the following questions. Include the names of the people you worked with at the top of your problem set.

Problem 1: Genome sizes and data storage (35 points total)

The NIH 's National Center for Biotechnology Information (NCBI) provides a huge repository and a multitude of databases for biological information. NCBI Entrez's Genome page (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome)is a good starting place for resources on genome projects. Many biology textbooks also commonly discuss genomic size and its biological basis.

1 (a) Find the approximate size of the West Nile viral genome, the microbial *Escherichia coli* K12 genome, the *Caenorhabditis elegans* haploid genome, the haploid human genome and the *Amoeba dubia* genome in base pairs. (1 pt each, total of 5)

West Nile virus: 10,962 bp
http://www.ncbi.nlm.nih.gov:80/PMGifs/Genomes/vis.html

Escherichia coli K12: 4.6 Mbp
http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/framik?db=Genome&gi=115

Caenorhabditis elegans: 97 Mbp
http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map_search?chr=celegans.inf

Human: 3.2 Gbp
http://www.ncbi.nlm.nih.gov/genome/seq/page.cgi?F=HsProgress.shtml&ORG=Hs

Amoeba dubia: 670 Gbp
http://gnn.tigr.org/articles/02_01/Sizing_genomes.shtml

1(b) Find out the estimated total number of genes in each of the above organisms. (1 pt each, total of 5)

NOTE: While the size of the genome for *Amoeba dubia* is known, an accurate estimate of the number of genes is not. For this case of 1(b), draw your conclusions based on a comparison of the complexity of the morphology of *Amoeba dubia* to the other organisms relative to the number of genes.

West Nile virus: 9-11 genes
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=nucleotide&list_uids=11497619&dopt=GenBank
http://mbe.library.arizona.edu/data/1992/0904/8watt.pdf

Escherichia coli K12: 4288 genes
http://www.ncbi.nlm.nih.gov/entrez/utils/qmap.cgi?uid=97426617&form=6&db=m&Dopt=r

Caenorhabditis elegans: 19,000 genes
http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map_search?chr=celegans.inf

Human: 30,000-40,000 genes
http://www.nature.com/genomics/human/papers/409860a0_fs_1.html

Amoeba dubia: Unknown

Due to estimates that vary with different sources, any estimates that approximate the above results is an acceptable answer. Because *Amoeba dubia* is unknown (as indicated in the note), full credit is given for either an “unknown” or a number supported by a literature reference.

Is the size of genome proportional to the total number of genes? Give at least one reason why this is or is not the case.(4 points)

Not precisely. Higher eukaryotes generally need larger genomes to accommodate the extra genes, but this is not precise (e.g., yeast, with a genome size of 12Mb or 0.004 times the size of the human genome, would be expected to have 0.004 * 80000 human genes = 320 genes, rather than it's 6000 genes, if its ratio of gene number to genome size was the same as that of humans). Ultimately the variation in genome sizes is due in large part to the amount of non-coding sequence, in the form of introns and intergenic sequences especially long, genome-wide repeats, in the genome. So genome size can be seen as related to how "tightly packed" the coding sequence is within the genome.

Is it always true that the more complex the organism, the large genome it has? Give an example if your answer is no and explain why.(4 points)

No. The morphology of *Amoeba dubia*, a unicellular organism, is far simpler than that of a human, which has many cell types, organs, etc. Nevertheless, *Amoeba dubia* has a genome ~200 times larger. There is no strict correlation between complexity and genome size, a concept sometimes referred to as the C-Value Paradox. A reason for this may be the presence of large quantities of non-coding or “junk” regions within the genome of organisms such as *Amoeba dubia*.

1(c) What is the minimum number of bytes required to store the genomes listed above? To store the human genome in its diploid rather than haploid form? Show your calculations! (6 points)

The minimum number of bytes would require each basepair to be encoded as two bits (rather than storage as an ASCII character). A byte would thus store 4 basepairs:

West Nile virus:	10962 bp / 4bp/byte	= 2.7 KB
Escherichia coli K12:	4.6 x 10^6 bp / 4bp/byte	= 1.15 MB
Caenorhabditis elegans:	95 x 10^6 bp / 4bp/byte	= 23.8 MB
Amoeba dubia:	670 x 10^9 bp / 4bp/byte	= 167 GB
Human (haploid):	3.2 x 10^9 bp / 4bp/byte	= 0.8 GB

Human (diploid): 3.2 x 10^9 bp / 4bp/byte X 2 = 1.6 GB

1(d) What is the minimum number of bytes needed to store all human genomes? All such genomes can be represented as a single individual's genome plus the variations, or polymorphisms, seen in all other human genomes. Assume that the human population is ~ 6 billion, which was the population reached in October 1999, and that polymorphic sites tend to be simple single nucleotide polymorphisms (SNPs) such as "A" in one genome and "C" in another) and occur about once every 3 kb (4pts).

The reference haploid genome can be stored with 0.8 GB, as above. So we know we need 0.8GB + space required to store the polymorphisms of 12billion (minus one) other haploid genomes.

There are two ways to think about the polymorphisms: (i) they all occur at the exact same place; (ii) between any two genomes that you compare (say, the reference genome and every other), there's a polymorphism every 3Kbp. Technically the first interpretation is contradicted by "A" in one genome and "C" in the other, which implies a comparison of two genomes.

(i) If you assume that all polymorphisms occur at the same place, you need to store a million locations (only once, since they're all the same) plus 12 billion (minus one) strings of a million bases each, in addition to the 0.8GB reference genome. For each location, you need 4 bytes (log 2 3,200,000,000 = 31.6 bits à 32bits = 4 bytes); one million locations would require 4MB. For storing the actual polymorphic base strings (the 12 billion strings of million bases), 12e9 * 1e6 bases * 0.25 bytes/base = 3e15 bytes. So in all:

8e8 bytes + 4e3 bytes + 3e15bytes = 3e15bytes (3.0000000800004e15 bytes)
= 3 petabytes

(ii) If you've realized that, in comparison with the reference genome, every genome has one million differences. For each difference in every genome, you need its location (32 bits) and the base (2 bits), or 34 bits. For 12 billion haploid genomes, that's 12e9 genomes* 1e6 polymorphisms/genome * 34bits/polymorphism * 0.125 bytes/bit = 51e15.

8e8 bytes + 51e15 bytes = 51E15 bytes = 51 petabytes

Full credit for giving the (ii) answer, or half credit for (i).

1(e) How many double-sided DVDs would it take to store the genomes listed above given your bit conversions above? How many 80GB hard disks would it take to store all human genomes in the world, again given your calculations above (4pts)?

2pt: For 8GB DVDs, it would take only 1 DVD to store any of the genomes listed, except for *Amoeba dubia*, which would require 21 DVDs. Because there are differences in DVD formats, full credit is given for a correct calculation based upon an identified and real DVD storage format that the student specifies.

2pt: It would take (i) 37,500 80 GB hard disks or (ii) 187,500 20GB hard disks to store all human genomes, using the storage methods described in (d).

Full credit given for properly worked-out answers regardless of whether the calculations from 1(c) and 1(d) had been correct.

1(f) Some nucleotide sequence data have to be stored at more than 2 bits/base. Could you think of a reason why this would be the case? (3 points)

There might be ambiguity in the sequence. For example, if people can't decide whether a position in a sequence is an "A" or a "G", it's usually denoted as "R", "C" or "T" is denoted as "Y" and "A" or "G" or "C" or "T" is denoted as "N", etc. The total possible letters representing one base will be more than four, therefore it has to be stored at more than 2 bits/base.

Another explanation is that most searching and string functions operate on ASCII data, which is 8 bits/base. The perl programs used in this class are generally written in this way.

Problem 2: Sequence occurrences (30 points total)

2 (a) At how many *sites* would you expect "CG" to occur in 4.6 Mbp (mega bp) in a *double-stranded* genome? How about "CTAG"? And "GATTACA"? Assume all nucleotides have an equal probability of occurring. (2 pts for each, total of 6 pts)

> **Hint: "CG" is a palindromic sequence. When it occurs on one strand, it also occurs on the complement.**
>
> 5'-*CG*-3'
> 3'-*GC*-5'
>
> At this site, you find 2 occurrences of "CG." (For palindromic sequences, you find 2 occurrences of the sequence at 1 site.)
>
> "GATTACA" is not palindromic.
>
> 5'- GATTACA -3'
> 3'- GATTACA -5'
>
> For non-palindromic sequences, you find 1 occurrence of the sequence at each site.

CG in 4.6 Mbp: (.25**2) * ~4.6e6 = 287 500
CTAG in 4.6 Mbp: (.25**4) * ~4.6e6 = 17 969
GATTACA in 4.6 Mbp: 2 x (.25**7) * ~4.6e6 = 562

Why is the predicted incidences of GATTACA multiplied by 2 (relative to the approach used for CG)? CG and CTAG are palindromic with respect to the complimentary sequence on the other strand; that is, 5'-CG-3' pairs with 3'-GC-5' (which is really the same sequence, 5'-CG-3'). GATTACA is NOT complimentary in this fashion. While the predicted incidence of CG equals the incidence on both strands (of the 4.6e6 bp, you have to sum the incidence of and its reverse compliment (TGTAATC) on the single strand of sequence you are analyzing in order to find the incidence rate of GATTACA on both strands. This is equivalent to multiplying the incidence by two. To summarize the basic idea with an example: all incidences of CG in sense and antisense strands are found by identifying all incidences of CG in the sense strand; however all incidences of GATTACA in the sense and antisense strands are NOT found by identifying all incidences of GATTACA in the sense strand.

2 (b) Run the following perl program, parse.pl, with the ~4.6 Mbp E. coli K12 *genomic sequence as input. (Obtain the genomic sequence file, "E.coli_K12.txt," from the course web page, or download the sequence from ncbi, http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/framik?db=Genome&gi=115. If you use the ncbi site, go to "NCBI FTP site" and download the "U00096.fna" file.)*

```
#!/usr/local/bin/perl
undef $/;
$text = <>;
$text =~ s/\>.+?\n//g;
$text =~ s/\n//g;
$string = $temp = "ctag";
$match = $text=~ s/$string//gi;
```

```
$a = $match * ($temp =~ s/a//gi) + ($text =~ s/a//gi);
$c = $match * ($temp =~ s/c//gi) + ($text =~ s/c//gi);
$g = $match * ($temp =~ s/g//gi) + ($text =~ s/g//gi);
$t = $match * ($temp =~ s/t//gi) + ($text =~ s/t//gi);
$n = $match * ($temp =~ s/n//gi) + ($text =~ s/n//gi);
print "$string: $match\na: $a\nc: $c\ng: $g\nt: $t\nn: $n\n";
```

Step by step instructions for students with fas accounts:

1. Send the *E. coli K12* genome sequence to your fas account and save it in the proper directory.
2. 2. Login to fas.harvard.edu.
3. 3. Type "pico" at the command line. This brings up a text editor called pico.
4. 4. Paste or type the program (above) into the pico window.
5. 5. Press “Ctrl-x” to exit pico. Type “y.” Type “parse.pl” to name the program. Press “return.”

2(b) What is the program output? (1 pt)

```
ctag: 887
a: 1142136
c: 1179433
g: 1176775
t: 1140877
n: 0
```

2(c) What is the ratio of the observed incidence of CTAG in the E. coli K12 genome to the expected value? (1pt)

887 / 17,969 = 0.0494

2(d) In part a, you should have assumed that each nucleotide has an equal probability of occurring, but you know from part c that this is not actuality the case. Based on the output from part c, re-calculate the expected incidence of CTAG in the E. coli K12 genome. (2 pts)

Total bp = 1142136 + 1179433 + 1176775 + 1140877 = 4639221
P(a) = 1142136 / 4639221 = 0.2461913
P(c) = 1179433 / 4639221 = 0.2542308
P(g) = 1176775 / 4639221 = 0.2536578
P(t) = 1140877 / 4639221 = 0.2459199

Expected frequency of ctag
 = P(ctag) * 4639221 = P(a) * P(t) * P(a) * P(g) * 4639221 = 18,112

Depending on the precision used in the calculation, the calculation might vary slightly, so any value approximating 18,112 is acceptable.

2(e) Re-calculate the ratio of observed to expected values for CTAG. You can exclude N's from the length of the sequence. (2 pts)

Ratio = 887 / 18,112 = 0.0489

The idea here is to calculate a new expected incidence of the sequence, using probabilities of the individiual nucleotides which are determined from the actual incidence of these nucleotides in the given sequence (i.e., predicted P(A) = actual #A/total # nucleotides). These new ("adjusted") predicted incidence rates are then used in calculating a new observed: expected ratio.

2(f) Why might the observed incidence be so different from the expected? Speculate what this may mean in a biological sense. (2pts)

Some sequences are involved in regulation (i.e. transcription regulation, RNA splicing, etc.) and/or stability of DNA. Thus, certain sequences may be selected for or against because of their biological manifestations (i.e. the signals they convey or the structural features they impart). Various answers acceptable.

2(g) This version of the E. coli K12 genomic sequence did not contain any N's (Note that an N represents any nucleotide at that particular position). If we were analyzing a sequence with N's, could we simply remove them from the sequence at the beginning of the program ($text = <>; $text =~s/n//ig;) ? (2pts)

No, because this would give us incorrect false positives (e.g., CNG would become CG, resulting in a hit).

2(h) Explain what the program does. Use any of the recommended Perl resources or texts to explain what each line does (12 pts).
Listing:

```
 1: #!/usr/local/bin/perl
 2: undef $/;
 3: $text = <>;
 4: $text =~ s/\>.+?\n//g;
 5: $text =~ s/\n//g;
 6: $string = $temp = "ctag";
 7: $match = $text=~ s/$string//gi;
 8: $a = $match * ($temp =~ s/a//gi) + ($text =~ s/a//gi);
 9: $c = $match * ($temp =~ s/c//gi) + ($text =~ s/c//gi);
10: $g = $match * ($temp =~ s/g//gi) + ($text =~ s/g//gi);
11: $t = $match * ($temp =~ s/t//gi) + ($text =~ s/t//gi);
12: $n = $match * ($temp =~ s/n//gi) + ($text =~ s/n//gi);
13: print "$string: $match\na: $a\nc: $c\ng: $g\nt: $t\nn: $n\n";
```

Analysis:

1: On UNIX, indicates a script interpreter. No operation on Windows.
2: Undefines the record separator, causing us to proceed to read input in "slurp mode" (i.e., we'll get the entire file content as a string rather than reading it a line at a time).
3: Slurps standard input into the scalar variable $text (allowing the user to utilize the < operator on the command line to specify a file to give the program as input, or otherwise type it in by hand).
4: Deletes the contents of $text up to the > and newline character. This is the file header, which might otherwise contain acgt characters which would through off our analysis.
5: Removes newline characters from the input (if we didn't do this, we might not read a proper sequence, e.g., CT\nAG wouldn't be recognized as CTAG).
6: Stores the value "ctag" in the scalar variables $string and $temp.
7: Stores the number of successful occurrences of $string within $text in the scalar variable $match (it does this by removing, via a case-insensitive string substitution, of $string in $text).
8-12: Finds the count of the occurrences of a, c, g, t and n (in each line respectively) by summing the occurrences of the character in the remaining undeleted sequence ($text) with the number of occurrences of that character already detected in line 7 (based on multiplying the times the string was found already times the number of occurrences of the character within the search string itself).
13: Produces output indicating the number of matches for $string, a, b, c and d.

2(i) Modify the program, parse.pl, so that it counts the number of times the string "TCAGGACT" occurs in the E. coli K12 genome. Find occurrences on both the sense and the antisense strands. Show your code and the output. (2 pts)

Code:

```
#!/usr/local/bin/perl
undef $/;
$text = <>;
$text =~ s/\>.+?\n//g;
$text =~ s/\n//g;

$string = $temp = "tcaggact";   # string to search for

# Generate the antisense version of $string with transliteration
$anti   = $string;
$anti =~ tr/acgtACGT/tgcaTCGA/;

# Now detect whether $anti is palindromic to $string.  If it is,
# we can skip $anti and just look for string.  Otherwise, the sum
# of the occurrences of both $anti and $string indicate the
# total occurrences of the string in a double-stranded genome.

$match_anti = 0;
if( $anti eq (reverse $string))
```

```
{
        print "Searching for palindromic string '$string'\n";
}
else
{
        # This is not a palindromic string
        print "Searching for '$string' and antisense '$anti'\n";
        $text_anti = $text;
        $match_anti = $text_anti=~ s/$anti//gi;
}

$match = $text=~ s/$string//gi;
$total_match = $match + $match_anti;

$a = $match * ($temp =~ s/a//gi) + ($text =~ s/a//gi);
$c = $match * ($temp =~ s/c//gi) + ($text =~ s/c//gi);
$g = $match * ($temp =~ s/g//gi) + ($text =~ s/g//gi);
$t = $match * ($temp =~ s/t//gi) + ($text =~ s/t//gi);
$n = $match * ($temp =~ s/n//gi) + ($text =~ s/n//gi);
print "$string: $total_match\na: $a\nc: $c\ng: $g\nt: $t\nn: $n\n";
```

Output:

```
Searching for 'tcaggact' and antisense 'agtcctga'
tcaggact: 78
a: 1142136
c: 1179433
g: 1176775
t: 1140877
n: 0
```

Problem 3: Sequence Alignments (35 points total)

3(a) Briefly describe the differences between global and local alignment and between pairwise and multiple sequence alignment. Bioinformatics (Mount, 2001) covers these in detail. (4pts)

Pairwise and multiple sequence alignment (2pts)

Clearly, pairwise alignment methods align two sequences, multiple sequence alignment more than two. The basic goal of pairwise and multiple sequence alignment is the same: to bring the greatest number of similar characters into register in the same column of the alignment. However, while pairwise alignment methods can accomplish this by scoring matches, mismatches, and gaps in alignments of two sequences to find the optimal alignment, msa methods must not only generate an alignment but also determine a cumulative score for the substitutions in multiple sequences. There a various approaches to msa, including progressive global alignment, starting by aligning the most similar sequences, alignments starting from locally conserved patterns, and iterative alignments.

Global and local (pairwise) alignment (2pts)

Global:

- optimally aligns the full sequences,

- end gaps are penalized as much as center gaps
- no minimum score

Local:

- optimally aligns part of the sequences
- gaps at ends effectively have no penalty
- minimum score is 0 due to the underlying algorithm; one way of thinking about this is that local alignment methods give higher weight to finding islands of strong similarity or identity rather than extending the alignment to find more but weaker neighboring similarity
- better for aligning sequences which are similar along some of their lengths but dissimilar in others (e.g., sequences that have a common conserved domain) and sequences that have significantly different lengths

3(b)(i) Compare BLAST to the Smith-Waterman algorithm. What are the advantages and disadvantages of BLAST? (3pts).

Advantage: speed. BLAST is much faster than Smith-Waterman algorithm and therefore is suitable for database search.

Disadvantage: BLAST is a heuristic algorithm. Therefore, unlike the Smith-Waterman algorithm, it doesn't necessarily give the optimal alignment.

ii. There are two major implementations of the BLAST algorithm initially developed by Altschul et al (J. Mol. Biol, 1998) , NCBI BLAST and WU-BLAST (Washington University). While WU-BLAST is commonly used for searching genome sequences, NCBI BLAST is the more widely used of the two. Here you will become familiar with NCBI BLAST .

Perform a standard nucleotide BLAST search (http://www.ncbi.nlm.nih.gov/BLAST/) with the following sequence, using the default settings. Describe the output (1pts) and explain the meaning of the associated measures associated with the output alignments (score and E-value) (2pts). What gene do you think this sequence is from and why (2pts)? What possible homologs are there in other species and why (2pts)? (7pts total)

```
>Unknown sequence
gagtgcttgg gttgtggtga aacattggaa gagagaatgt gaagcagcca ttcttttcct
gctccacagg aagccgagct gtctcagaca ctggcatggt gttgggggag ggggttcctt
ctctgcaggc ccaggtgacc cagggttgga agtgtctcat gctggatccc cacttttcct
cttgcagcag ccagactgcc ttccgggtca ctgccatgga ggagccgcag tcagatccta
gcgtcgagcc ccctctgagt caggaaacat tttcagacct atggaaactg tgagtggatc
cattggaagg gcaggcccac caccccgacc ccaaccccag ccccctagca gagacctgtg
ggaagcgaaa attccatggg actgactttc tgctcttgtc tttcagactt cctgaaaaca
acgttctggt aaggacaagg gttgggctgg ggacctggag ggctgggggg ctggggggct
gaggacctgg tcctctgact gctcttttca cccatctaca gtcccccttg ccgtcccaag
caatggatga tttgatgctg tccccggacg atattgaaca atggttcact gaagacccag
```

gtccagatga agctcccaga atgccagagg ctgctccccg cgtggcccct gcaccagcag
ctcctacacc ggcggcccct gcaccagccc cctcctggcc cctgtcatct tctgtccctt
cccagaaaac ctaccagggc agctacggtt tccgtctggg cttcttgcat tctgggacag
ccaagtctgt gacttgcacg gtcagttgcc ctgaggggct ggcttccatg agacttcaat
gcctggccgt atccccctgc atttcttttg tttggaactt tgggattcct cttcaccctt
tggcttcctg tcagtgtttt tttatagttt acccacttaa tgtgtgatct ctgactcctg
tcccaaagtt gaatattccc cccttgaatt tgggctttta tccatcccat cacaccctca
gcatctctcc tggggatgca gaacttttct ttttcttcat ccacgtgtat tccttggctt
ttgaaaataa gctcctgacc aggcttggtg gctcacacct gcaatcccag cactctcaaa
gaggccaagg caggcagatc acctgagccc aggagttcaa gaccagcctg ggtaacatga
tgaaacctcg tctctacaaa aaaatacaaa aaattagcca ggcatggtgg tgcacaccta
tagtcccagc cactcaggag gctgaggtgg gaagatcact tgaggccagg agatggaggc
tgcagtgagc tgtgatcaca ccactgtgct ccagcctgag tgacagagca agaccctatc

2pts: The sequence is from the Homo sapiens tumor suppressor protein p53 (P53) gene, as indicated by the output below (E = 0).

2 pts: Tree shrews (*Tupaia belangeri chinensis*), rhesus monkey (*Macaca mulatta*) and the African green monkey (*Cercopithecus aethiops*) appear to have high similarity.based on their low E-Values.

2 pts: Bit scores – raw scores are the sum of the scores of the high scoring sequence pairs composing the alignment but can't generally be compared between different alignments because the scoring matrices may be different. Bit scores are therefore calculated from raw scores; they are raw scores that have undergone conversion for m log base of the scoring matrix to a standard log base2, thereby allowing bit scores to be compared across alignments. E-value – the expectation, the approximate expected occurrence of an alignment with the associated bit score (i.e., the number of times you would expect to get an alignment with that score by chance). E-values of 0.1 to 0.05 are typically used as cut-offs for significant alignments.

1 pt for explaining that the output contains a list of organisms with significant alignments, as ranked by bit scores and E-Values. The following is the output generated by this search:

```
                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

gi|4732144|gb|AF136270.1|HOMOTSP1  Homo sapiens tumor suppre...  1423   0.0
gi|4731629|gb|AF135120.1|HSM059JP1  Homo sapiens tumor suppr...  1423   0.0
gi|3041866|gb|U94788.1|HSU94788  Human p53 (TP53) gene, comp...  1402   0.0    L
gi|35213|emb|X54156.1|HSP53G  Human p53 gene for transformat...  1402   0.0    L
gi|2443802|gb|AF020387.1|AF020387  Homo sapiens deletion joi...  1039   0.0
gi|1177472|emb|X92659.1|HSP53I4  H.sapiens intron 4 from p53...   825   0.0
gi|189467|gb|M22884.1|HUMP53A04  Human phosphoprotein p53 ge...   609   e-171
gi|13097806|gb|BC003596.1|BC003596  Homo sapiens, tumor prot...   555   e-155  L U
gi|11066969|gb|AF307851.1|AF307851  Homo sapiens p53 protein...   555   e-155  U
gi|8400737|ref|NM_000546.2|  Homo sapiens tumor protein p53 ...   555   e-155  L U
gi|506452|emb|X60020.1|HSP53011  Human mRNA for mutated p53 ...   555   e-155  L
gi|506450|emb|X60019.1|HSP53010  Human mRNA for mutated p53 ...   555   e-155  L
gi|506446|emb|X60017.1|HSP53008  Human mRNA for mutated p53 ...   555   e-155  L
gi|506442|emb|X60015.1|HSP53006  Human mRNA for mutated p53 ...   555   e-155  L
```

```
gi|506440|emb|X60014.1|HSP53005  Human mRNA for mutated p53 ...   555   e-155
gi|506438|emb|X60013.1|HSP53004  Human mRNA for mutated p53 ...   555   e-155
gi|506436|emb|X60012.1|HSP53003  Human mRNA for mutated p53 ...   555   e-155
gi|506434|emb|X60011.1|HSP53002  Human mRNA for mutated p53 ...   555   e-155
gi|506432|emb|X60010.1|HSP53001  Human mRNA for mutated p53 ...   555   e-155
gi|7595311|gb|AF192534.1|AF192534  Expression vector Ad5CMV-...   547   e-152
gi|189478|gb|K03199.1|HUMP53T  Human p53 cellular tumor anti...   547   e-152
gi|506448|emb|X60018.1|HSP53009  Human mRNA for mutated p53 ...   547   e-152
gi|506444|emb|X60016.1|HSP53007  Human mRNA for mutated p53 ...   547   e-152
gi|35209|emb|X02469.1|HSP53  Human mRNA for p53 cellular tum...   547   e-152
gi|189453|gb|M13114.1|HUMP5304  Homo sapiens phosphoprotein ...   545   e-152
gi|339813|gb|M14694.1|HUMTP53A  Human p53 cellular tumor ant...   539   e-150
gi|6653198|gb|AF175893.1|AF175893  Tupaia belangeri chinensi...   531   e-148
gi|339815|gb|M14695.1|HUMTP53B  Human p53 cellular tumor ant...   531   e-148
gi|409391|gb|L20442.1|MACP53A  Rhesus monkey p53 mRNA, compl...   428   e-117
gi|4691417|dbj|AB018045.1|  Homo sapiens HSP70-1 gene for he...   424   e-115
gi|2689464|gb|U48956.1|U48956  Macaca mulatta p53 gene, comp...   420   e-114
gi|22795|emb|X16384.1|CAP53  African Green Monkey mRNA for p...   420   e-114
gi|2689466|gb|U48957.1|U48957  Macaca fascicularis p53 gene,...   412   e-112
```

iii. Repeat the search as a translated nucleotide query of all protein databases (BLASTX). What differences do you observe in the output and why (2pts)? Which organisms have possible homologs to the above sequence given this search (1pt)? Which search would you use, the standard nucleotide BLAST or the translated BLAST, if searching for possible homology in future searches (1pt)? (4pts total)

2 pts: More sequences are returned across a wider range of organisms. See the results below for specific differences. These differences are the result of searching for similarity to the translated sequence rather than the sequence itself; the amino acid sequence is going to be more conserved than the nucleotide sequence, given the degenerate nature of the genetic code.

1 pts: Possible homologs - various monkeys, dogs, cats, Chinese hamster all have scores well above the reasonable E-Value cutoff.

1 pt for any reasonable rationale demonstrating an understanding of the differences between standard nucleotide BLAST and translated BLAST.

Blast output:

```
                                                                Score    E
Sequences producing significant alignments:                     (bits) Value

gi|8400738|ref|NP_000537.2|  (NM_000546) tumor protein p53 [...   156   6e-37
gi|4731632|gb|AAD28535.1|AF135121_1  (AF135121) tumor suppre...   156   6e-37
gi|13097807|gb|AAH03596.1|AAH03596  (BC003596) tumor protein...   156   6e-37
gi|506453|emb|CAA42635.1|  (X60020) p53 transformation  supp...   156   6e-37
gi|506443|emb|CAA42630.1|  (X60015) p53 transformation  supp...   156   6e-37
gi|339814|gb|AAA61211.1|  (M14694) p53 antigen [Homo sapiens]     156   6e-37
gi|506441|emb|CAA42629.1|  (X60014) p53 transformation  supp...   156   6e-37
gi|506439|emb|CAA42628.1|  (X60013) p53 transformation  supp...   156   6e-37
gi|506451|emb|CAA42634.1|  (X60019) p53 transformation  supp...   156   6e-37
gi|506435|emb|CAA42626.1|  (X60011) p53 transformation  supp...   156   6e-37
```

```
gi|506433|emb|CAA42625.1|  (X60010) p53  transformation supp...   156   6e-37

gi|189479|gb|AAA59989.1|  (K03199) p53 cellular tumor antige...   155   8e-37
gi|129369|sp|P04637|P53_HUMAN  Cellular tumor antigen p53 (T...   155   8e-37

gi|506449|emb|CAA42633.1|  (X60018) p53 transformation  supp...   155   8e-37

gi|506445|emb|CAA42631.1|  (X60016) p53 transformation  supp...   155   8e-37

gi|339816|gb|AAA61212.1|  (M14695) p53 antigen [Homo sapiens]     155   8e-37

gi|386994|gb|AAA59987.1|  (M13121) phosphoprotein p53 [Homo ...   155   8e-37
gi|10720194|sp|Q9TTA1|P53_TUPGB  Cellular tumor antigen p53 ...   155   1e-36
gi|129367|sp|P13481|P53_CERAE  Cellular tumor antigen p53 (T...   147   3e-34
gi|3024332|sp|P56424|P53_MACMU  Cellular tumor antigen p53 (...   147   4e-34
gi|3024331|sp|P56423|P53_MACFA  Cellular tumor antigen p53 (...   145   1e-33
gi|10720190|sp|O36006|P53_MARMO  Cellular tumor antigen p53 ...   114   4e-24
gi|2842741|sp|Q95330|P53_RABIT  Cellular tumor antigen p53 (...   104   3e-21
gi|18997097|gb|AAL83290.1|AF475081_1  (AF475081) P53 [Delphi...   102   8e-21
gi|10720197|sp|Q9WUR6|P53_CAVPO  Cellular tumor antigen p53 ...   100   5e-20
gi|728838|sp|P39195|ALU8_HUMAN  Alu subfamily SX sequence co...    65   8e-20
gi|1890327|emb|CAA70109.1|  (Y08901) p53 tumour suppressor [...    99   9e-20
gi|2499428|sp|O09185|P53_CRIGR  Cellular tumor antigen p53 (...    99   9e-20
gi|728831|sp|P39188|ALU1_HUMAN  Alu subfamily J sequence con...    69   2e-19
gi|7440008|pir||JC6176  tumor suppressor protein p53 - Chine...    98   3e-19
gi|10720186|sp|Q9TUB2|P53_PIG  Cellular tumor antigen p53 (T...    98   3e-19
gi|473579|gb|AAB41344.1|  (U07182) tumor supressor p53 [Meso...    97   4e-19
gi|129370|sp|Q00366|P53_MESAU  Cellular tumor antigen p53 (T...    97   4e-19

gi|10437485|dbj|BAB15056.1|  (AK025047) unnamed protein prod...    68   5e-18
gi|2842672|sp|Q64662|P53_SPEBE  Cellular tumor antigen p53 (...    92   1e-17
gi|21748687|dbj|BAC03469.1|  (AK090511) unnamed protein prod...    63   2e-17
gi|1709531|sp|P51664|P53_SHEEP  Cellular tumor antigen p53 (...    91   4e-17
gi|6841071|gb|AAF28891.1|AF124298_1  (AF124298) p53 protein ...    91   4e-17

gi|8923273|ref|NP_060219.1|  (NM_017749) hypothetical protei...    61   6e-17
gi|21748935|dbj|BAC03508.1|  (AK090720) unnamed protein prod...    66   1e-16
gi|1171969|sp|P41685|P53_FELCA  Cellular tumor antigen p53 (...    88   2e-16
gi|728837|sp|P39194|ALU7_HUMAN  Alu subfamily SQ sequence co...    59   3e-16
gi|4996230|dbj|BAA78379.1|  (AB020761) P53 [Canis familiaris]      86   8e-16
gi|9280152|dbj|BAB01630.1|  (AB046048) unnamed portein produ...    65   1e-15
gi|10437569|dbj|BAB15071.1|  (AK025116) unnamed protein prod...    62   3e-15
gi|6093639|sp|Q29537|P53_CANFA  Cellular tumor antigen p53 (...    84   3e-15
gi|2465420|gb|AAB72093.1|  (AF021816) chimeric tumour suppre...    80   4e-14
gi|16550580|dbj|BAB71008.1|  (AK055769) unnamed protein prod...    61   4e-14

gi|14771451|ref|XP_045396.1|  (XM_045396) similar to KIAA150...    45   5e-14
gi|21758970|dbj|BAC05428.1|  (AK098835) unnamed protein prod...    57   2e-13

gi|10433567|dbj|BAB13989.1|  (AK022217) unnamed protein prod...    78   3e-13

gi|6755881|ref|NP_035770.1|  (NM_011640) transformation rela...    77   4e-13

gi|11493463|gb|AAG35505.1|AF130117_38  (AF130079) PRO2852 [H...    52   5e-13
gi|728836|sp|P39193|ALU6_HUMAN  Alu subfamily SP sequence co...    57   8e-13

gi|2961247|gb|AAC05704.1|  (AF051368) tumor suppressor p53 [...    76   8e-13

gi|200201|gb|AAA39882.1|  (M13873) p53 [Mus musculus]              76   1e-12
gi|223827|prf||1001197A  antigen p53,tumor [Mouse cytomegalo...    76   1e-12

gi|5081783|gb|AAD39535.1|AF151353_1  (AF151353) tumor suppre...    76   1e-12

gi|20881811|ref|XP_126235.1|  (XM_126235) transformation rel...    76   1e-12

gi|15375072|gb|AAK94783.1|  (AY044188) transformation relate...    76   1e-12
gi|575528|dbj|BAA03927.1|  (D16460) p53 protein [Felis catus]      75   1e-12
gi|21756961|dbj|BAC04988.1|  (AK097266) unnamed protein prod...    75   2e-12

gi|11342599|emb|CAC17147.1|  (AJ297973) transformation relat...    75   2e-12
gi|21758113|dbj|BAC05246.1|  (AK098160) unnamed protein prod...    65   3e-12
gi|481535|pir||S38824  cellular tumor antigen p53, minor spl...    74   3e-12
gi|13365926|dbj|BAB39337.1|  (AB056812) hypothetical protein...    69   5e-12

gi|53571|emb|CAA25323.1|  (X00741) p53 [Mus musculus]              74   5e-12
gi|13359175|dbj|BAB33321.1|  (AB051438) KIAA1651 protein [Ho...    56   6e-12
gi|10121865|gb|AAG13405.1|AF285159_1  (AF285159) topoisomera...    50   7e-12
gi|21753371|dbj|BAC04333.1|  (AK094331) unnamed protein prod...    58   8e-12
gi|7770139|gb|AAF69605.1|AF119917_13  (AF119851) PRO1722 [Ho...    72   1e-11
gi|18027310|gb|AAL55737.1|AF289553_1  (AF289553) unknown [Ho...    61   1e-11
gi|2829679|sp|P79892|P53_HORSE  Cellular tumor antigen p53 (...    72   1e-11
gi|1836145|gb|AAB46899.1|  (S83123) sequence-specific transc...    72   1e-11

gi|11494110|gb|AAG35765.1|AF209191_1  (AF209191) p53 alterna...    72   1e-11
gi|129372|sp|P10361|P53_RAT  Cellular tumor antigen p53 (Tum...    72   1e-11
gi|13591878|ref|NP_112251.1|  (NM_030989) tumor protein p53 ...    72   1e-11
gi|21756629|dbj|BAC04924.1|  (AK096998) unnamed protein prod...    72   2e-11

gi|18595461|ref|XP_088648.1|  (XM_088648) similar to PRO2822...    72   2e-11

gi|8923452|ref|NP_060312.1|  (NM_017842) hypothetical protei...    52   2e-11
gi|21753365|dbj|BAC04331.1|  (AK094327) unnamed protein prod...    57   2e-11

gi|10438620|dbj|BAB15291.1|  (AK025947) unnamed protein prod...    63   2e-11
gi|12698155|dbj|BAB21904.1|  (AB055280) hypothetical protein...    51   2e-11
gi|21751050|dbj|BAC03893.1|  (AK092450) unnamed protein prod...    72   2e-11
gi|2499426|sp|Q29628|P53_BOVIN  Cellular tumor antigen p53 (...    71   3e-11
gi|1729419|dbj|BAA08629.1|  (D49825) p53 gene product [Bos p...    71   3e-11
gi|10434925|dbj|BAB14424.1|  (AK023140) unnamed protein prod...    71   3e-11

gi|22044024|ref|XP_170931.1|  (XM_170931) similar to hypothe...    71   3e-11
gi|1000577|gb|AAB42022.1|  (S77819) p53 [Canis familiaris]        71   3e-11
gi|16266760|dbj|BAB69969.1|  (AB033632) p53 [Meriones unguic...    71   3e-11
gi|21749185|dbj|BAC03549.1|  (AK090929) unnamed protein prod...    46   4e-11

gi|18554512|ref|XP_087329.1|  (XM_087329) similar to hypothe...    67   4e-11
gi|21928603|dbj|BAC05890.1|  (AB065664) seven transmembrane ...    59   4e-11
gi|1310770|pdb|1TSR|A  Chain A, P53 Core Domain In Complex W...    70   4e-11
gi|2781308|pdb|1YCS|A  Chain A, P53-53bp2 Complex                  70   4e-11
gi|20809854|gb|AAH28935.1|  (BC028935) Unknown (protein for ...    70   6e-11

gi|1938365|gb|AAB80959.1|  (U90328) mutant p53 [Rattus norve...    70   6e-11
gi|6690223|gb|AAF24043.1|AF090928_1  (AF090928) PRO0470 [Hom...    50   6e-11
gi|1619833|gb|AAB16961.1|  (U62133) p53 [Canis familiaris]         70   8e-11

gi|18552162|ref|XP_087124.1|  (XM_087124) similar to hypothe...    70   8e-11
```

iv. Repeat the search in iii using the PAM30 matrix rather than BLOSUM62. Describe any differences in output that you observe. *(3 points)* In BLASTX, you have 5 matrices to choose from: PAM30, PAM70, BLOSUM45, BLOSUM62 and BLOSUM80. If you want to find more divergent sequences, which two should you use and why?*(3 points)*

3 pts: PAM70 and BLOSUM45 will produce more divergent results.

3 pts BLAST output:

```
Sequences producing significant alignments:                      (bits) Value

gi|189479|gb|AAA59989.1|  (K03199) p53 cellular tumor antige...   161   2e-66

gi|8400738|ref|NP_000537.2|  (NM_000546) tumor protein p53 [...   161   2e-66
gi|129369|sp|P04637|P53_HUMAN  Cellular tumor antigen p53 (T...   161   2e-66

gi|4731632|gb|AAD28535.1|AF135121_1  (AF135121) tumor suppre...   161   2e-66

gi|13097807|gb|AAH03596.1|AAH03596  (BC003596) tumor protein...   161   2e-66

gi|506453|emb|CAA42635.1|  (X60020) p53 transformation  supp...   161   2e-66

gi|506449|emb|CAA42633.1|  (X60018) p53 transformation  supp...   161   2e-66

gi|506445|emb|CAA42631.1|  (X60016) p53 transformation  supp...   161   2e-66

gi|506443|emb|CAA42630.1|  (X60015) p53 transformation  supp...   161   2e-66

gi|339814|gb|AAA61211.1|  (M14694) p53 antigen [Homo sapiens]     161   2e-66

gi|339816|gb|AAA61212.1|  (M14695) p53 antigen [Homo sapiens]     161   2e-66

gi|506441|emb|CAA42629.1|  (X60014) p53 transformation  supp...   161   2e-66

gi|386994|gb|AAA59987.1|  (M13121) phosphoprotein p53 [Homo ...   161   2e-66

gi|506439|emb|CAA42628.1|  (X60013) p53 transformation  supp...   161   2e-66
gi|10720194|sp|Q9TTA1|P53_TUPGB  Cellular tumor antigen p53 ...   161   2e-66

gi|506451|emb|CAA42634.1|  (X60019) p53 transformation  supp...   161   2e-66

gi|506435|emb|CAA42626.1|  (X60011) p53 transformation  supp...   161   2e-66

gi|506433|emb|CAA42625.1|  (X60010) p53  transformation supp...   161   2e-66
gi|3024332|sp|P56424|P53_MACMU  Cellular tumor antigen p53 (...   141   2e-59
gi|129367|sp|P13481|P53_CERAE  Cellular tumor antigen p53 (T...   141   2e-59
gi|3024331|sp|P56423|P53_MACFA  Cellular tumor antigen p53 (...   138   2e-58
gi|10720190|sp|O36006|P53_MARMO  Cellular tumor antigen p53 ...   102   4e-42
gi|2842741|sp|Q95330|P53_RABIT  Cellular tumor antigen p53 (...   110   4e-40
gi|18997097|gb|AAL83290.1|AF475081_1  (AF475081) P53 [Delphi...   101   9e-37
gi|10720197|sp|Q9WUR6|P53_CAVPO  Cellular tumor antigen p53 ...    84   2e-32
gi|1171969|sp|P41685|P53_FELCA  Cellular tumor antigen p53 (...    96   2e-29
gi|728831|sp|P39188|ALU1_HUMAN  Alu subfamily J sequence con...    87   1e-28
gi|2842672|sp|Q64662|P53_SPEBE  Cellular tumor antigen p53 (...    99   3e-27
gi|728838|sp|P39195|ALU8_HUMAN  Alu subfamily SX sequence co...    77   8e-27

gi|10437485|dbj|BAB15056.1|  (AK025047) unnamed protein prod...    80   2e-24
gi|2465420|gb|AAB72093.1|  (AF021816) chimeric tumour suppre...   115   2e-24

gi|8923273|ref|NP_060219.1|  (NM_017749) hypothetical protei...    70   6e-24
gi|21748935|dbj|BAC03508.1|  (AK090720) unnamed protein prod...    76   9e-24
gi|21748687|dbj|BAC03469.1|  (AK090511) unnamed protein prod...    70   1e-23
gi|728837|sp|P39194|ALU7_HUMAN  Alu subfamily SQ sequence co...    69   1e-22

gi|15375072|gb|AAK94783.1|  (AY044188) transformation relate...    93   2e-21

gi|6755881|ref|NP_035770.1|  (NM_011640) transformation rela...    93   2e-21

gi|200201|gb|AAA39882.1|  (M13873) p53 [Mus musculus]             93   2e-21
gi|223827|prf||1001197A  antigen p53,tumor [Mouse cytomegalo...    93   2e-21

gi|2961247|gb|AAC05704.1|  (AF051368) tumor suppressor p53 [...    93   2e-21

gi|5081783|gb|AAD39535.1|AF151353_1  (AF151353) tumor suppre...    93   2e-21

gi|20881811|ref|XP_126235.1|  (XM_126235) transformation rel...    93   2e-21
gi|481535|pir||S38824  cellular tumor antigen p53, minor spl...    93   2e-21
gi|6841071|gb|AAF28891.1|AF124298_1  (AF124298) p53 protein ...   103   1e-20
gi|1709531|sp|P51664|P53_SHEEP  Cellular tumor antigen p53 (...   103   1e-20

gi|11342599|emb|CAC17147.1|  (AJ297973) transformation relat...    93   1e-20
gi|10437569|dbj|BAB15071.1|  (AK025116) unnamed protein prod...    71   1e-20
gi|10720186|sp|Q9TUB2|P53_PIG  Cellular tumor antigen p53 (T...   102   3e-20
gi|9280152|dbj|BAB01630.1|  (AB046048) unnamed portein produ...    76   4e-20
gi|2781308|pdb|1YCS|A  Chain A, P53-53bp2 Complex                 101   5e-20
gi|1310770|pdb|1TSR|A  Chain A, P53 Core Domain In Complex W...   101   5e-20
gi|2829679|sp|P79892|P53_HORSE  Cellular tumor antigen p53 (...    98   7e-20
gi|1836145|gb|AAB46899.1|  (S83123) sequence-specific transc...    98   7e-20
gi|1890327|emb|CAA70109.1|  (Y08901) p53 tumour suppressor [...   100   9e-20
gi|2499428|sp|O09185|P53_CRIGR  Cellular tumor antigen p53 (...   100   9e-20
gi|473579|gb|AAB41344.1|  (U07182) tumor supressor p53 [Meso...   100   9e-20
gi|129370|sp|Q00366|P53_MESAU  Cellular tumor antigen p53 (T...   100   9e-20
gi|1729419|dbj|BAA08629.1|  (D49825) p53 gene product [Bos p...   100   1e-19
gi|2499426|sp|Q29628|P53_BOVIN  Cellular tumor antigen p53 (...   100   1e-19
gi|1813451|gb|AAB41831.1|  (U48616) p53 [Mastomys natalensis]      96   2e-19
gi|21730310|pdb|1GZH|C  Chain C, Crystal Structure Of The Br...    99   3e-19
gi|21730308|pdb|1GZH|A  Chain A, Crystal Structure Of The Br...    99   3e-19
gi|20151154|pdb|1KZY|A  Chain A, Crystal Structure Of The 53...    99   3e-19
gi|16266760|dbj|BAB69969.1|  (AB033632) p53 [Meriones unguic...    98   4e-19
gi|129372|sp|P10361|P53_RAT  Cellular tumor antigen p53 (Tum...    98   4e-19

gi|13591878|ref|NP_112251.1|  (NM_030989) tumor protein p53 ...    98   4e-19

gi|11494110|gb|AAG35765.1|AF209191_1  (AF209191) p53 alterna...    98   4e-19

gi|22069934|ref|XP_170805.1|  (XM_170805) similar to OK/SW-C...    53   8e-19
gi|7440008|pir||JC6176  tumor suppressor protein p53 - Chine...    97   9e-19
gi|1000577|gb|AAB42022.1|  (S77819) p53 [Canis familiaris]         96   1e-18
gi|4996230|dbj|BAA78379.1|  (AB020761) P53 [Canis familiaris]      96   1e-18
gi|6093639|sp|Q29537|P53_CANFA  Cellular tumor antigen p53 (...    96   1e-18
gi|1619833|gb|AAB16961.1|  (U62133) p53 [Canis familiaris]         96   1e-18
gi|575528|dbj|BAA03927.1|  (D16460) p53 protein [Felis catus]      96   2e-18
gi|21930134|gb|AAM82163.1|  (AF525302) tumor suppressor p53 ...    95   3e-18
gi|21104464|dbj|BAB93502.1|  (AB062477) OK/SW-CL.41 [Homo sa...    48   3e-18

gi|1938365|gb|AAB80959.1|  (U90328) mutant p53 [Rattus norve...    95   4e-18

gi|22044024|ref|XP_170931.1|  (XM_170931) similar to hypothe...    94   7e-18

gi|53571|emb|CAA25323.1|  (X00741) p53 [Mus musculus]              93   1e-17

gi|11493463|gb|AAG35505.1|AF130117_38  (AF130079) PRO2852 [H...    53   2e-17
gi|18027310|gb|AAL55737.1|AF289553_1  (AF289553) unknown [Ho...    74   4e-17
gi|13365926|dbj|BAB39337.1|  (AB056812) hypothetical protein...    84   5e-17

gi|8923452|ref|NP_060312.1|  (NM_017842) hypothetical protei...    56   7e-17

gi|14189960|gb|AAK55521.1|AF305818_1  (AF305818) PRO0764 [Ho...    55   7e-17
gi|9929935|dbj|BAB12124.1|  (AB047600) hypothetical protein ...    75   7e-17
gi|728836|sp|P39193|ALU6_HUMAN  Alu subfamily SP sequence co...    60   3e-16

gi|10438620|dbj|BAB15291.1|  (AK025947) unnamed protein prod...    74   3e-16
```

```
gi|1177607|emb|CAA63219.1|  (X92485) pva1 [Plasmodium vivax]       54   4e-16
gi|7542422|gb|AAF63442.1|  (AF210310) p53 tumor suppressor p...     88   4e-16
gi|21749185|dbj|BAC03549.1|  (AK090929) unnamed protein prod...     53   7e-16
gi|21757878|dbj|BAC05206.1|  (AK097965) unnamed protein prod...     86   2e-15
gi|7107285|gb|AAF36354.1|AF209128_1  (AF209128) tumor suppre...     86   2e-15
gi|21754870|dbj|BAC04580.1|  (AK095583) unnamed protein prod...     52   2e-15
gi|21758113|dbj|BAC05246.1|  (AK098160) unnamed protein prod...     69   3e-15
gi|13359175|dbj|BAB33321.1|  (AB051438) KIAA1651 protein [Ho...     58   7e-15
gi|6690223|gb|AAF24043.1|AF090928_1  (AF090928) PRO0470 [Hom...     63   9e-15
gi|21757775|dbj|BAC05191.1|  (AK097877) unnamed protein prod...     52   1e-14
gi|16550580|dbj|BAB71008.1|  (AK055769) unnamed protein prod...     58   2e-14
gi|18027740|gb|AAL55831.1|AF318324_1  (AF318324) unknown [Ho...     83   2e-14
gi|21756629|dbj|BAC04924.1|  (AK096998) unnamed protein prod...     83   2e-14
```

(c) Global alignment with Needlman-Wunsch (6pts total).

Back in 1969, S. Needleman and C. Wunsch came up with an efficient method of obtaining an optimal global alignment (Needleman, S. B., Wunsch, C. D., *J. Mol. Biol.* (1970) 48:443-453). Although not known to the authors, the proposed algorithm followed the principle of dynamic programming introduced some 12 years earlier by Richard Bellman, and later popular alignment algorithms (such as Smith-Waterman local alignment) were based on this Needleman-Wunsch method. The Needleman-Wunsch algorithm can be written as follows:

$$S_{ij} = \max \{ S_{i-1,\, j-1} + s(a_i b_j),\quad \max (S_{i-x,\, j} - w_x),\quad \max(S_{i,\, j-y} - w_y) \}$$

or, to simplify:

$$S_{ij} = \max \{ S_{i-1,\, j-1} + w(\text{match, mismatch}),\quad S_{i-1,\, j} + w(\text{gap}),\quad S_{i,\, j-1} + w(\text{gap}) \}$$

Although the Needleman-Wunsch algorithm was initially used to align amino acid sequences, the algorithm can be applied to strings of an arbitrary alphabet. For the purposes of this question, we will consider only DNA sequences. In *E. coli* promoter sequences, the -35 signal TTGACAT is well-known to have functional significance. *Align the -35 signal TTGACAT to the sequence GTTGTACTT* using the Needleman-Wunch algorithm with the following weight function for the DNA sequence alignments in the problems below:

□·•w(match) =2
□·•w(mismatch) =-1
□·•w(gap) =-3

Show the optimal global alignment(s) as well as the matrix that scores all possible alignments.

Alignment matrix:

	0	T	T	G	A	C	A	T
0	0	-3	-6	-9	-12	-15	-18	-21
G	**-3**	-1	-4	-4	-7	-10	-13	-16
T	-6	**-1**	1	-2	-5	-8	-11	-11
T	-9	-4	**1**	0	-3	-6	-9	-9
G	-12	-7	-2	**3**	0	-3	-6	-9
T	-15	-10	-5	**0**	2	-1	-4	-4
A	-18	-13	-8	-3	**2**	1	1	-2
C	-21	-16	-11	-6	-1	**4**	1	0
T	-24	-19	-14	-9	-4	1	**3**	3
T	-27	-22	-17	-12	-7	-2	0	**5**

Optimal global alignment
-TTG-ACAT
GTTGTACTT

3 pts for demonstrating understanding of the algorithm plus 3 pts for a correct answer.

(d) How would you (or Smith and Waterman) modify the Needleman-Wunsch algorithm to yield optimal local rather than global alignments (2pts)? Perform your alignment again with the local alignment algorithm. Show the optimal local alignment as well as the matrix that scores all possible alignments. (3 pts)

2 pts: Add zero, such that the maximum of the diagonal (match/mismatch), vertical, horizonal (gaps), AND 0, is taken:

$$S_{ij} = \max \{ S_{i-1,j-1} + s(a_i b_j), \\ \max(S_{i-x,j} - w_x), \\ \max(S_{i,j-y} - w_y), \\ 0 \\ \}$$

(Smith-Waterman,1981)

or, following the above simplified format:

$$S_{ij} = \max \{ S_{i-1,j-1} + w(\text{match, mismatch}), \\ S_{i-1,j} + w(\text{gap}), \\ S_{i,j-1} + w(\text{gap}), \\ 0 \\ \}$$

3 points :

Alignment matrix:

	T	T	G	A	C	A	T
G	0	0	2	0	0	0	0
T	**2**	2	0	1	0	0	2
T	2	**4**	1	0	0	0	2
G	0	1	**6**	3	0	0	0
T	2	2	**3**	5	2	0	2
A	0	1	1	**5**	4	4	1
C	0	0	0	2	**7**	4	3
T	2	2	0	0	4	**6**	6
T	2	4	1	0	1	5	**8**

Local alignment
TTG-ACAT
TTGTACTT