IB Biology 3.1 Genes/3.2 Chromosomes: Bioinformatics Database Practice Activity

Application: Use of a database to determine differences in the base sequence of a gene in two species.

GenBank Database Procedure:
1. Go to the Entrez Gene website http://www.ncbi.nlm.nih.gov/gene and search for "human actin."
2. Click on the first link that appears and scroll WAY down to the section "NCBI Reference Sequences."
3. Under "mRNA and Proteins," click on the first file name. It will be named "NM_000177.4" or something similar. These standardized numbers make cataloging sequence files easier. Do not worry about the file number for now.
4. Just below the gene title click on "FASTA." This is the name for a particular format for displaying sequences.
5. The nucleotide sequence displayed is that of the actin gene in humans.
6. On the far right, under Analyze this Sequence, click Run BLAST
7. Give the query a title in the box provided if you plan on saving it for later. For example - Actin 1
8. Under "Choose Search Set," "Database" select other to compare against all organisms available.
9. Under "Program Selection," choose "highly similar sequences"
10. Click BLAST.
11. Select the first response from the list - *Homo sapiens*. The **identities** section indicates the number of identical nucleotides in the sequence to humans. Since this is a human, it should be 100% identical.

Homo sapiens gelsolin (GSN), transcript variant 1, mRNA
Sequence ID: ref|NM_000177.4| Length: 2719 Number of Matches: 1

Range 1: 1 to 2719 GenBank Graphics   ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
5022 bits(2719)	0.0	2719/2719(100%)	0/2719(0%)	Plus/Plus

```
Query  1   ACTTAAGGTCGGCGACCCGAGGCCGCGGCTGCCGACTGGGTCCCCTGCCGCTGTCGCCAC  60
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1   ACTTAAGGTCGGCGACCCGAGGCCGCGGCTGCCGACTGGGTCCCCTGCCGCTGTCGCCAC  60
```
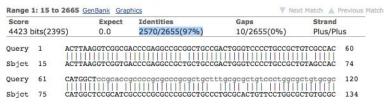
Related Information

Gene - associated gene details
GEO Profiles - microarray expression data
Map Viewer - aligned genomic context

14. Select another species at random from further down the list.

☐ PREDICTED: Pan troglodytes gelsolin (GSN), transcript variant X12, mRNA	4421	4421	89%	0.0	99% XM_001161601.4
☐ PREDICTED: Pan troglodytes gelsolin (GSN), transcript variant X11, mRNA	4421	4421	89%	0.0	99% XM_009457206.1
☐ PREDICTED: Pan troglodytes gelsolin (GSN), transcript variant X10, mRNA	4421	4421	89%	0.0	99% XM_009457205.1
☐ PREDICTED: Pan troglodytes gelsolin (GSN), transcript variant X9, mRNA	4421	4421	89%	0.0	99% XM_001160857.4
☐ PREDICTED: Pan troglodytes gelsolin (GSN), transcript variant X8, mRNA	4421	4421	89%	0.0	99% XM_009457204.1
☐ PREDICTED: Pan troglodytes gelsolin (GSN), transcript variant X7, mRNA	4421	4421	89%	0.0	99% XM_009457203.1
☐ PREDICTED: Pan troglodytes gelsolin (GSN), transcript variant X6, mRNA	4421	4421	89%	0.0	99% XM_009457202.1
☐ PREDICTED: Pan troglodytes gelsolin (GSN), transcript variant X5, mRNA	4421	4421	89%	0.0	99% XM_009457201.1
☐ PREDICTED: Pan troglodytes gelsolin (GSN), transcript variant X4, mRNA	4421	4421	89%	0.0	99% XM_009457200.1

15. The **identities** section indicates the number of identical nucleotides in the sequence (as compared to humans)

PREDICTED: Macaca fascicularis uncharacterized LOC101864919 (LOC101864919), transcript variant X1, mRNA
Sequence ID: ref|XM_005580904.1| Length: 2672 Number of Matches: 1

Range 1: 15 to 2665 GenBank Graphics   ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
4423 bits(2395)	0.0	2570/2655(97%)	10/2655(0%)	Plus/Plus

```
Query  1   ACTTAAGGTCGGCGACCCGAGGCCGCGGCTGCCGACTGGGTCCCCTGCCGCTGTCGCCAC   60
           ||||||||||||| ||||||||||||| || |||||||||||||||||||||||| ||||
Sbjct  15  ACTTAAGGTCGGTGACCCGAGGCCGCGCTGCTGCCGACTGGGTCCCCTGCCGCTGTAGCCAC  74

Query  61  CATGGCTccgcaccgccccgcgcccgcgctgctttgcgcgctgtccctggcgctgtgcgc  120
           |||||||||||||| |||||||||||||||||| ||||||||||||||||| |||||||||
Sbjct  75  CATGGCTCCGCATCGCCCCGCGCCCGCGCTGCCCTGCGCACTGTTCCTGGCGCTGTGCGC  134
```

Related Information

Gene - associated gene details
Map Viewer - aligned genomic context

16. Repeat steps 1-15 for a gene of your choice. Suggestions include: catalase, myosin or keratin.
17. Record your selected gene and the results for 4 species in the table below.

Genus species	GENE:	
	Number of identical nucleotides	Similarity (%)
Homo sapiens		100

Questions:

1. What is the function of the gene you investigated? Explain whether or not the function is the same in all four species?

2. Explain the benefits of the GenBank database for the scientific community.

3. Deduce what it means for the evolutionary ancestry of the species, if there is a high percentage of identical nucleotide sequences.

4. Transcribe and translate the following DNA gene sequence:

   Antisense DNA: TAC   ACT   TAA   GGT   CGG   CGA   CCC   GAG   GCC   GCG

   mRNA:

   polypeptide:

6. If a base substitution mutation occurred in the 5th mRNA codon and the second C was replaced with an A, explain the effect on the resulting polypeptide sequence.

7. If a base substitution mutation occurred in the 5th mRNA codon and the G was replaced with a C, explain the effect on the resulting polypeptide sequence.

Application: Use of an online database to identify the locus of a human genes and its protein product.

Online Mendelian Inheritance in Man Database Procedure:
1. Go to the Online Mendelian Inheritance in Man website http://www.omim.org/
2. Enter "gene map" into the search box and hit search.
3. Type HBB in the search box and hit search. HBB is the beta subunit of hemoglobin.
4. Record the cytogenetic location in the table below.
5. Repeat steps 3-5 for TDF, the gene that causes a fetus to develop into a male.
6. Read the "How Scientists Record Gene Location" bullet points below and use the information to complete the table below.

Gene	Cytogenetic Location	Chromosome #	Arm	Position of Band	Sub-Band
HBB					
TDF					

How Scientists Record Gene Location - http://ghr.nlm.nih.gov/handbook/howgeneswork/genelocation
- The unique combination of numbers and letters provide a gene's "address" on a chromosome.
- The first number or letter used to describe a gene's location represents the chromosome.
- Each chromosome is divided into two sections (arms) based on the location of a narrowing (constriction) called the centromere. By convention, the shorter arm is called p, and the longer arm is called q. The chromosome arm is the second part of the gene's address. For example, 5q is the long arm of chromosome 5 and Xp is the short arm of the X chromosome.
- The position of the gene on the p or q arm. The position of a gene is based on a distinctive pattern of light and dark bands that appear when the chromosome is stained in a certain way. The position is usually designated by two digits (representing a region and a band), which are sometimes followed by a decimal point and one or more additional digits (representing sub-bands within a light or dark area). The number indicating the gene position increases with distance from the centromere. For example: 14q21 represents position 21 on the long arm of chromosome 14. 14q21 is closer to the centromere than 14q22.
- Sometimes, the abbreviations "cen" or "ter" are also used to describe a gene's cytogenetic location. "Cen" indicates that the gene is very close to the centromere. Ter" stands for terminus, which indicates that the gene is very close to the end of the p or q arm.

Chromosomal location of a gene

chromosome # 7

short arm p

Centromere

long arm q

long arm q

Example gene:
CFTR

Chromosomal location:
7 q 3 1 . 2

band # 1
sub-band # 2

region # 3

U.S. National Library of Medicine

11p15.4 11p14.3 11p12 11q12.1 11q13.2 11q14.1 11q22.1 11q23.2 11q24.3

11p15.2 11p14.1 11p11.12 11q12.3 11q13.4 11q14.3 11q22.3 11q24.1