



THE BIOTECHNOLOGY
EDUCATION COMPANY®

Edvo-Kit #

340

Edvo-Kit #340

DNA Informatics

Experiment Objective:

In this experiment, students will explore the popular bioinformatics tool BLAST. First they will read sequences from autoradiographs of automated gel runs. The resulting data will then be analyzed using publicly available databases and BLAST to identify genes and gene products.

See page 2 for storage instructions.

Table of Contents

	Page
Experiment Components	2
Experiment Requirements	3
Background Information	4
Experiment Procedures	
Guide to using BLASTN	6
Exercises	11
Study Questions	14
Instructor's Guidelines	15
Answers to Exercises	16
Answers to Study Questions	17

Experiment Components & Requirements

Experiment can be stored at room temperature.

This experiment contains a total of 3 sets of 4 autoradiographs from gel run sequencing. Students can use any sequence database to perform the activities in this lab. For purposes of simplification we have chosen to illustrate with a database offered by the NCBI.

COMPONENTS

- 3 sets of 4 Autoradiographs

REQUIREMENTS

- Computer with Internet access
- White Light Box*

*A white light box is recommended. The EDVOTEK White Light Gel Visualization System (Cat. # 552) is well-suited for this lab. An autoradiograph may also be placed on an overhead projector and shown to the whole class.

EDVOTEK and The Biotechnology Education Company are registered trademarks of EDVOTEK, Inc.



1.800.EDVOTEK • Fax 202.370.1501 • info@edvotek.com • www.edvotek.com

Duplication of any part of this document is permitted for non-profit educational purposes only. Copyright © 2000-2015 EDVOTEK, Inc., all rights reserved. 340.151217

Background Information

BIOINFORMATICS AS A MODERN CORNERSTONE OF BIOLOGY

DNA sequencing technology allows for the analysis of DNA at the nucleotide level. Nucleotides are the monomer building blocks of DNA (Figure 1). Each deoxynucleotide (dNTP) comprises three basic parts: a phosphate group, a deoxyribose sugar, and a nitrogen-containing base (adenine, cytosine, guanine, or thymine). The 3' hydroxyl group on the sugar of one nucleotide forms a covalent bond with the 5' phosphate group of its neighbor. The nature of this bond results in remarkably stable strands with a distinct polarity, making DNA ideal for genetic information storage.

Automated and high-throughput sequencing methods have made generating sequence information vastly more efficient. For example, the human genome, which consists of around three billion nucleotides, initially took over a decade to sequence but can now be completed in less than a day. As a result there is a great emphasis not only on generating sequence data but also on storing and analyzing the data. Bioinformatics represents the intersection of computer science and biology, involving any mathematical or computing approach that advances our understanding of biological processes. One major goal in bioinformatics is to develop computer programs that enable efficient access to and management of large data sets. Other

methods aim to develop new algorithms (mathematical formulas) and statistical measures that assess relationships among members of large data sets. These programs often include pattern recognition, data mining, machine learning, and visualization.

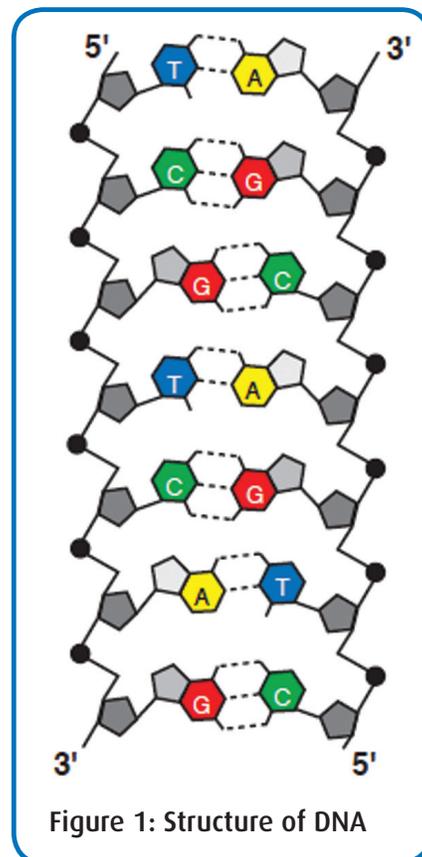


Figure 1: Structure of DNA

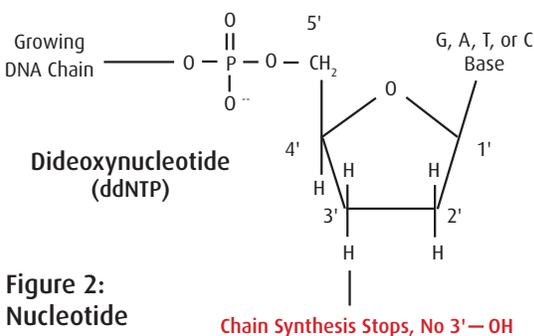
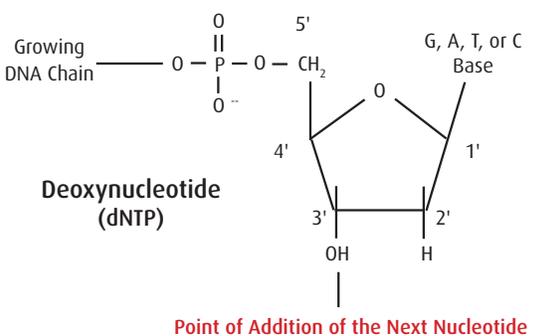


Figure 2:
Nucleotide
Structure

THE SANGER METHOD OF DNA SEQUENCING

Chain termination sequencing, often called Sanger sequencing, allows researchers to generate long DNA reads of a target sequence, also known as the template. The DNA template is combined with DNA primer, the DNA polymerase I enzyme (DNA Pol I), and a mixture of two types of free nucleotides, deoxynucleotides (dNTPs) and dideoxynucleotides (ddNTPs). Importantly, one of the dNTPs is often labeled with radioactive phosphorus-32 or sulfur-35 to allow for visualization of the DNA fragments at later stages of the experiment. During the sequencing reaction, DNA Pol I copies the DNA template by adding dNTPs to the primer to form a complementary strand of DNA. Occasionally, the DNA Pol I will instead add a ddNTP to the DNA strand. This ddNTP

Background Information

lacks a 3' hydroxyl group which makes it impossible for the polymerase to add another nucleotide (Figure 2). As a result the synthesis of that particular strand of DNA is terminated. In Sanger sequencing four separate enzymatic reactions are performed, one for each nucleotide. At the end of the incubation each sample contains a population of molecules each with identical 5' ends and 3' ends that terminate with the same ddNTP base. However, since each reaction terminates randomly the sample will contain a mixture of different sized fragments depending on the site at which point the ddNTP was incorporated. Figure 3 gives an example of the fragments produced by a ddGTP reaction.

Once the four Sanger sequencing reactions have been completed the mixtures of fragments can be analyzed by polyacrylamide gel electrophoresis (PAGE). The samples are added into depressions (or "wells") within a polyacrylamide gel and an electrical current is passed through the gel. Because the sugar-phosphate backbone of DNA has a strong negative charge, the current drives the DNA through the gel towards the positive electrode. At first glance, a polyacrylamide gel appears to be a solid at room temperature. However, on the molecular level the gel contains small channels through which the DNA can pass. Small DNA fragments move through these holes easily, but large DNA fragments have a more difficult time squeezing through the tunnels. Because molecules with dissimilar sizes travel at different speeds, they become separated and form discrete "bands" within the gel. This allows PAGE to separate fragments that differ in size by a single nucleotide. Together, the four sequencing reactions contain DNA fragments that cover the entire length of the DNA's sequence, with each fragment corresponding to a different nucleotide.

After electrophoretic separation is complete, autoradiography is performed. The polyacrylamide gel is placed into direct contact with a sheet of x-ray film. Since the DNA fragments are radiolabeled their position can be detected as dark bands on the x-ray film (Figure 4). Remember, since the smallest DNA fragments migrate further through the gel the sequence must be read from the bottom up. This provides a complementary sequence to the original DNA template for further analysis.

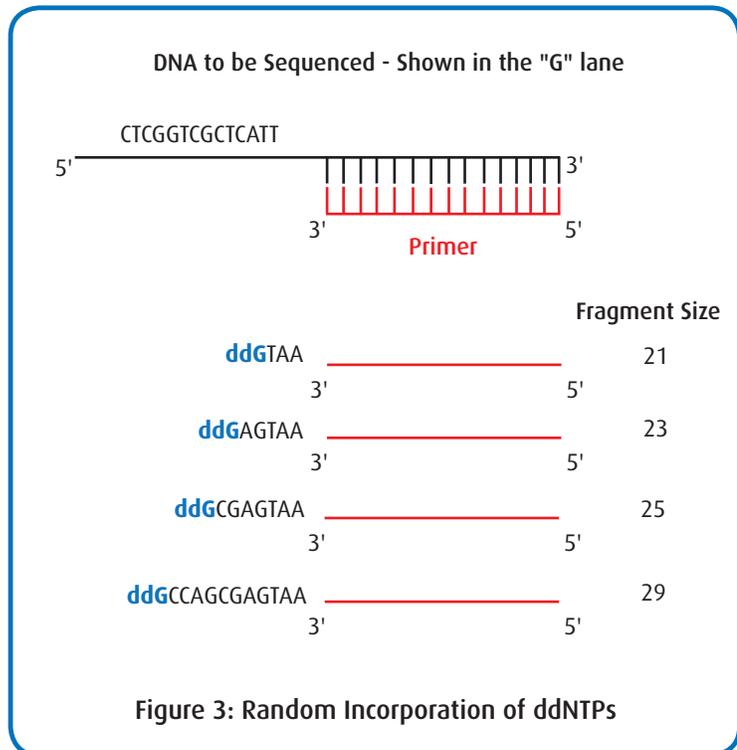


Figure 4: Simulated Sequencing Gel

A	C	G	T	Fragment Size	Deduced Sequence
			T	31	T
		G	T	30	G
			T	29	T
			T	28	A
A				27	A
		G		26	G
A				25	A
			T	24	T
			T	23	T
		G		22	G
			T	21	T
	C			20	C
		G		19	G

Deduced Sequence: GCTGTTAGAATGT

Background Information

SEQUENCE COMPARISON USING BLAST

Data from DNA sequencing is of limited value unless it is converted to biologically useful information, making bioinformatics a critical part of DNA sequencing. After collecting DNA sequence data, molecular biologists will often search public databases for similar sequences. Comparing sequences can lead to the identification of genes and other conserved sequence patterns. In addition, sequence comparison can be used to establish functional, structural, and evolutionary relationships among genes. Another potential benefit is that sequence comparison provides a reliable method for inferring the biological functions of newly sequenced genes.

One of the largest and most influential databases is known as GenBank. This free, open source database contains over a trillion nucleotide bases of publically available sequence data. Each entry in GenBank contains a sequence and a unique accession number, as well as supporting bibliographic and biological annotations such as author references and taxonomic data. The NCBI (National Center for Biotechnology Information) oversees and maintains the entire database, but each entry is submitted directly by individual laboratories. Direct submission has allowed the database to keep pace with the rapid growth in sequence data production. However, it also means that heterogeneity in entry quality exists, especially in the certainty of each nucleotide's identity and in the extent of attached annotation. The impact of this uncertainty can vary depending on the goals of the study, the physical properties of the DNA region(s), and the chosen sequencing method. To address this, GenBank classifies the sequence information based on the sequencing strategy used to obtain the data.

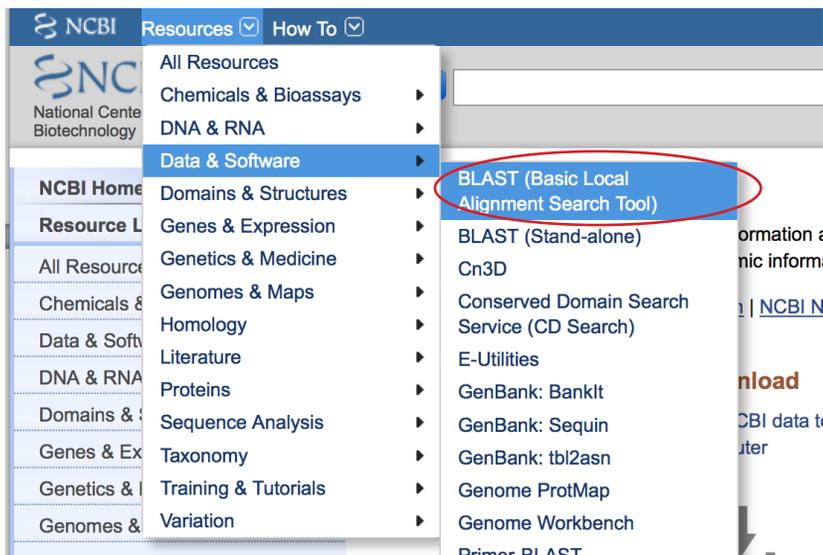
Associated with this database is the Basic Local Alignment Search Tool, or BLAST. The BLAST program finds regions of local similarity between a user's DNA sequence and sequences in the GenBank database. In BLAST terminology, the user's input sequence is known as the query sequence, sequences in the database are known as target sequences, and sequences with similarities to the input sequence are hits. The user can draw inference about the putative molecular function of the query sequence by looking at the hits. Similarities between hits suggest homology, the existence of shared ancestry between the genes. In addition, BLAST can be used to identify unknown species, locate known protein domains, and find potential chromosome locations.

BLAST takes a heuristic approach to the problem of searching through such a mammoth database of target sequences. This means the program takes shortcuts in order to find sequence matches in a reasonable time frame. These shortcuts are based on the assumption that biologically similar sequences will contain short identical stretches of sequences. The entire database is scanned for the presence of these short stretches, and sequences containing two stretches within a preset distance are set aside. These sequences are then locally aligned with the query sequence to see if the nucleotide order beyond the two stretches matches. By searching the GenBank database this way BLAST can return results very quickly, although it sacrifices some accuracy and precision.

This exercise introduces students to bioinformatics through the analysis of sequencing data. In order to gain experience in database searching, students will use the free service offered by the NCBI. At present, GenBank comprises several databases including the GenBank and EMBL (European Molecular Biology Laboratory) nucleotide sequences, the non-redundant GenBank complementary DNA translations (protein sequences), and the EST (expressed sequence tags) database. These exercises will involve using BLASTN to compare nucleotide sequences.

Guide to Using BLASTN

1. Type: **www.ncbi.nlm.nih.gov** to log on to the NCBI web page.
2. On the top left of the screen click on the drop down menu "Resources".
3. Click on "Data and Software", then click on "BLAST" (Basic Local Alignment Search Tool).



4. On the new BLAST Home screen select "nucleotide blast" which is the first option under the Basic BLAST list.

Guide to Using BLASTN, continued

5. On the new screen make sure the tab selected is “blastn”.

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/ BLAST/ blastn suite Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

Enter Query Sequence

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#) [Reset page](#)

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) Query subrange

From

6. Enter the nucleotide sequence into the large box in the “Enter Query Sequence” section; be careful to type the following sequence exactly:

```
ggcaactgcccaaagtgtgatccagcctgtctcaacagaa
```

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

ggcaactgcccaaagtgtgatccagcctgtctcaacagaa

7. Under “Choose Search Set” make sure that “Others (nr etc.)” is selected and that “Nucleotide collection (nr/nt)” is highlighted in the dropdown menu. The remaining entries should be left blank.

Choose Search Set

Database Human genomic + transcript Mouse genomic + transcript Others (nr etc.):

Nucleotide collection (nr/nt)

Organism **Optional** Enter organism name or id—completions will be suggested Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude **Optional** Models (XM/XP) Uncultured/environmental sample sequences

Limit to **Optional** Sequences from type material

Entrez Query **Optional** [YouTube](#) [Create custom database](#)

Enter an Entrez query to limit search

Guide to Using BLASTN, continued

8. Under “Program Selection” select “High similar sequence (megablast)”.

Program Selection

Optimize for

Highly similar sequences (megablast)

More dissimilar sequences (discontiguous megablast)

Somewhat similar sequences (blastn)

Choose a BLAST algorithm

9. Click on the blue “BLAST” query box.

BLAST | Search **database Nucleotide collection (nr/nt)** using **Megablast (Optimize for highly similar sequences)**

Show results in a new window

10. Once the “BLAST” query box has been clicked you will be assigned an ID#. Record this number so you can check your results at a later time.
11. Examine the BLASTN search report. The report includes:
- a. **Search Summary Report** that shows an overview of the BLASTN search parameters.

Your search parameters were adjusted to search for a short input sequence.

[Edit and Resubmit](#) [Save Search Strategies](#) [▶ Formatting options](#) [▶ Download](#)

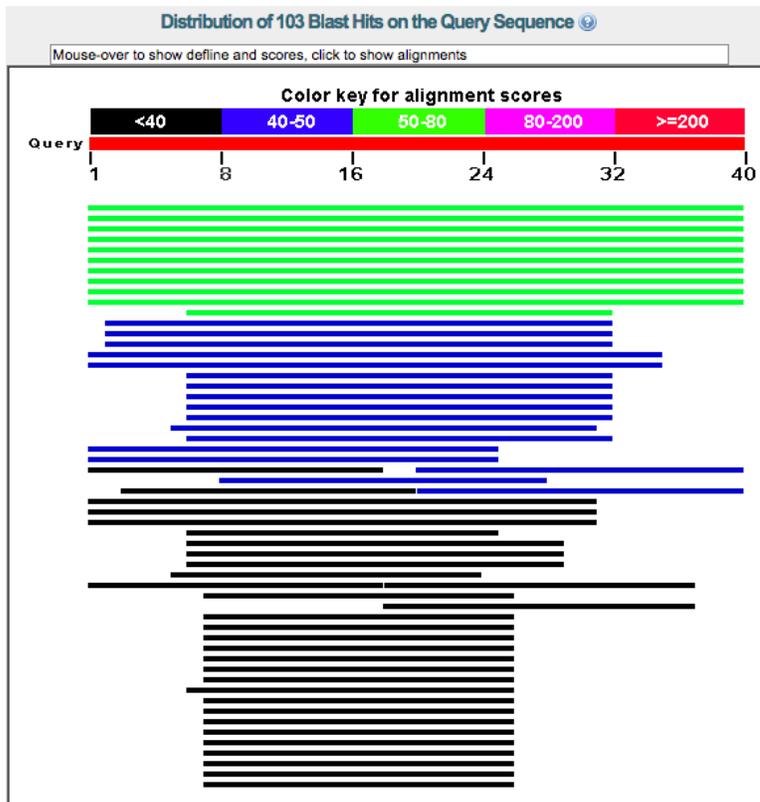
Nucleotide Sequence (40 letters)

<p>RID S38347Y801R (Expires on 06-19 00:29 am)</p> <p>Query ID lcl Query_44773</p> <p>Description None</p> <p>Molecule type nucleic acid</p> <p>Query Length 40</p>	<p>Database Name nr</p> <p>Description Nucleotide collection (nt)</p> <p>Program BLASTN 2.2.31+ ▶ Citation</p>
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------

Other reports: [▶ Search Summary](#) [\[Taxonomy reports\]](#) [\[Distance tree of results\]](#)

Guide to Using BLASTN, continued

- b. **Graphic Summary section** that shows the alignment of database matches to the query sequence. The color of the boxes corresponds to the score of the alignment with red representing the highest alignment scores.



- c. **Description section** that shows all the sequences in the database with significant sequence homology to our sequence. By default the results are sorted according to the E-value but you can click on the column header to sort the results by different categories. Notice that there can be several different entries with identical high scores.

Descriptions

Sequences producing significant alignments:

Select: All None Selected: 0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/> PREDICTED: Bos taurus epidermal growth factor receptor (EGFR), mRNA	79.8	79.8	100%	2e-12	100%	XM_002896890.3
<input type="checkbox"/> PREDICTED: Bos taurus epidermal growth factor receptor (EGFR), mRNA	79.8	79.8	100%	2e-12	100%	XM_592211.7
<input type="checkbox"/> PREDICTED: Bubalus bubalis epidermal growth factor receptor (EGFR), mRNA	79.8	79.8	100%	2e-12	100%	XM_006069632.1
<input type="checkbox"/> PREDICTED: Bos mutus epidermal growth factor receptor (EGFR), partial mRNA	79.8	79.8	100%	2e-12	100%	XM_005901147.1
<input type="checkbox"/> Bos taurus isolate HNBT1007153 epidermal growth factor receptor (EGFR) mRNA, complete cds	79.8	79.8	100%	2e-12	100%	HM749883.1
<input type="checkbox"/> PREDICTED: Bison bison epidermal growth factor receptor (EGFR), mRNA	71.9	71.9	100%	4e-10	98%	XM_010842153.1
<input type="checkbox"/> PREDICTED: Pantheropsis hodgsonii epidermal growth factor receptor (EGFR), mRNA	63.9	63.9	100%	1e-07	95%	XM_005975627.1
<input type="checkbox"/> PREDICTED: Capra hircus epidermal growth factor receptor (EGFR), mRNA	63.9	63.9	100%	1e-07	95%	XM_005695500.1
<input type="checkbox"/> PREDICTED: Ovis aries musimon epidermal growth factor receptor (EGFR), transcript variant X2, mRNA	56.0	56.0	100%	2e-05	93%	XM_012170695.1
<input type="checkbox"/> PREDICTED: Ovis aries epidermal growth factor receptor (EGFR), transcript variant X1, mRNA	56.0	56.0	100%	2e-05	93%	XM_012099280.1
<input type="checkbox"/> PREDICTED: Orcinus orca epidermal growth factor receptor (EGFR), mRNA	52.0	52.0	65%	4e-04	100%	XM_004283006.1
<input type="checkbox"/> PREDICTED: Balaenoptera acutorostrata scammoni epidermal growth factor receptor (EGFR), transcript variant X3, mRNA	46.1	46.1	77%	0.024	94%	XM_007187664.1
<input type="checkbox"/> PREDICTED: Balaenoptera acutorostrata scammoni epidermal growth factor receptor (EGFR), transcript variant X2, mRNA	46.1	46.1	77%	0.024	94%	XM_007187663.1

Guide to Using BLASTN, continued

- d. **Alignment section** that shows alignment blocks for each BLAST hit. Each alignment block begins with a summary that includes the Max score and expected value, sequence identity, the number of gaps in the alignment, and the orientation of the query sequence relative to the subject sequence.

The image displays three screenshots of BLASTN alignment results, each showing a predicted sequence and its alignment with a query sequence. Each screenshot includes a table with columns for Score, Expect, Identities, Gaps, and Strand, followed by the query and subject sequences with vertical bars indicating alignment positions.

Screenshot 1:
 Download GenBank Graphics
 PREDICTED: *Bos taurus* epidermal growth factor receptor (EGFR), mRNA
 Sequence ID: [ref|XM_002696890.3|](#) Length: 8412 Number of Matches: 1
 Range 1: 711 to 750 GenBank Graphics Next Match Previous Match

Score	Expect	Identities	Gaps	Strand
79.8 bits(40)	2e-12	40/40(100%)	0/40(0%)	Plus/Plus

 Query 1 GGCAACTGCCCAAAGTGTGATCCAGCCTGTCTCAACAGAA 40
 Sbjct 711 GGCAACTGCCCAAAGTGTGATCCAGCCTGTCTCAACAGAA 750

Screenshot 2:
 Download GenBank Graphics
 PREDICTED: *Bos taurus* epidermal growth factor receptor (EGFR), mRNA
 Sequence ID: [ref|XM_592211.7|](#) Length: 8414 Number of Matches: 1
 Range 1: 711 to 750 GenBank Graphics Next Match Previous Match

Score	Expect	Identities	Gaps	Strand
79.8 bits(40)	2e-12	40/40(100%)	0/40(0%)	Plus/Plus

 Query 1 GGCAACTGCCCAAAGTGTGATCCAGCCTGTCTCAACAGAA 40
 Sbjct 711 GGCAACTGCCCAAAGTGTGATCCAGCCTGTCTCAACAGAA 750

Screenshot 3:
 Download GenBank Graphics
 PREDICTED: *Bubalus bubalis* epidermal growth factor receptor (EGFR), mRNA
 Sequence ID: [ref|XM_00609632.1|](#) Length: 8425 Number of Matches: 1
 Range 1: 715 to 754 GenBank Graphics Next Match Previous Match

Score	Expect	Identities	Gaps	Strand
79.8 bits(40)	2e-12	40/40(100%)	0/40(0%)	Plus/Plus

 Query 1 GGCAACTGCCCAAAGTGTGATCCAGCCTGTCTCAACAGAA 40
 Sbjct 715 GGCAACTGCCCAAAGTGTGATCCAGCCTGTCTCAACAGAA 754

12. Select a sequence to focus on in-depth. You can do this by clicking on a colored bar in the graphic section, clicking on the sequence name in the description section, or scrolling down to the alignment section. Then click on the sequence ID. This brings up additional information about the subject sequence, including the gene name, the genus and species of origin, and articles written about the gene. After performing this search, one of the top hits should be *Bos taurus* epidermal growth factor receptor (EGFR), mRNA. Sequence ID: [ref|XM_002696890.3|](#). If top hit does not match, try re-entering the sequence. Be sure to double check the search parameters before the BLASTN searching.

Exercises

EXERCISE 1

Familiarize yourself with the autoradiograph by reading the DNA sequence for sample #1.

1. Start at the arrow and read up the gel for 20 nucleotides. Record the DNA sequence. Submit it to NCBI using the BLASTN program.
2. Start at the arrow and read up the gel for 30 nucleotides. Record the DNA sequence. Submit it to NCBI using the BLASTN program.

A few notes on reading a sequencing gel:

- Either enter the sequence directly into the query box or write the sequence down on a piece of paper and then enter it.
- It is critical that you do not confuse lanes when reading the sequence. The gel contains the A, C, G, T lanes from left to right.
- Reading a sequence gel requires that you read the nucleotides in the 5'→3' direction. This can be accomplished by reading "up" the gel (from the bottom of the gel to the top).
- Notice that for the most part that the spacing and intensity of most of the bands is fairly constant. Ignore lightly colored bands and choose only the darker ones. Occasionally the sequence will be dark and all four lanes will be of relatively similar intensity. This is called a DNA sequencing compression and is common when there are stretches of G and C's. This type of pattern should be treated as an ambiguous position (see next note).
- If an exact band at a position is ambiguous, you can enter an N which denotes that it could be either A, C, G or T.

Results to be obtained for autoradiograph 1:

- a. Do the BLASTN results for the first and second search resemble each other?
- b. What is the name of this gene?
- c. What organism's DNA was likely sequenced in this exercise?

Exercises, continued

EXERCISE 2

Now that you have familiarity with the entry and submission process, read the DNA sequence analysis from the autoradiograph corresponding to number 2. Notice that it is sometimes difficult to judge the spacing and strongest intensity of the band in each lane and therefore you need to use your best judgment.

1. Begin the exercise by reading the DNA sequence for sample number 2 approximately 6 cm from the bottom of the strip. The first 12 nucleotides should be: 5'...GGACGACGGTAT...3'.
2. Submit the sequence to NCBI using the BLASTN program.
3. After receiving the BLASTN results, scroll down to the alignment section and look at the entries that have nucleotide matches with your query sequence.

Some more notes on reading a sequencing gel and interpreting BLAST results:

- Remember that DNA sequence is always entered in the 5'→3' direction.
- DNA is double stranded and contains a top (5'→3') and bottom (3'→5') strand (sometimes this corresponds to the coding and noncoding strands). When a query sequence is searched against the database both strands of the query are examined. The entered sequence is known as the plus strand and reverse complement of this sequence is known as the minus strand.
- As a general rule, identical nucleotide sequence spanning greater than 21 bp between two samples usually indicates that the sequences are related or identical.

Results to be obtained for autoradiograph 2:

- a. What is the name of this gene?
- b. What strand does the matched query sequence represent? What strand does the hit sequence represent?

EXERCISE 3

1. Read the DNA sequence from sample number 3. Start at the bottom of the strip and record the DNA sequence.
2. Submit the sequence to NCBI using the BLASTN program.
3. Click on the GenBank accession number of a sequence hit to access further information about the DNA sequence and/or gene.

Results to be obtained for autoradiograph 3:

- a. What is the name of this gene?
- b. Approximately how many base pairs does this gene have?

Exercises, continued

EXERCISE 4

This section demonstrates the interaction of two proteins encoded by two genes. Protein-protein interactions play a fundamental role in virtually every process in a living cell.

For example, signals from the exterior of a cell are mediated to the inside of that cell by the protein-protein interactions of the signaling molecules. This process, called signal transduction, is of central importance in many biological processes such as cell division and cytoskeleton formation. In this exercise, we will use DNA sequences to characterize two human genes.

1. Read the DNA sequence obtained from sample number 4. Start at the bottom of the strip and record the DNA sequence for around 30 base pairs.
2. Next, move approximately a third of the way up the strip (~14cm) and read a portion of this section of the DNA sequence.
3. For this exercise limit the database search to human genes. To do this go to "Choose Search Set" and select "Human genomic + transcript". See Guide to Using BLASTN step 7.
4. Submit each sequence section individually to the BLASTN program.
5. Once you have identified the name of these two genes, perform a general Internet search to collect more information about the two proteins.

Results to be obtained for autoradiograph 4:

- a. This sample contains two DNA sequences (from bottom section and starting from the middle section). What are the corresponding names of these genes?
- b. What are the functions of the two proteins encoded by these genes?
- c. How do these two proteins interact in a living cell?

Study Questions

1. What is DNA sequencing?
2. What does each band in the autoradiograph represent?
3. What is BLAST? Why is it considered a bioinformatics tool?

Instructor's Guide

Below are the answers obtained by performing a BLASTN search with the sequences provided with this kit as of its publication (2015). Because the GenBank database is constantly being revised, we recommend that you run the below sequences through GenBank prior to performing this lab with your students. If you observe difference in the answers, please contact Edvotek® Technical Support at 1-800-EDVOTEK or info@edvotek.com

#1 sequence (20 nucleotides) 5'....ACAAATAGTTACCTTGAAC.....3'

#1 sequence (30 nucleotides) 5'.... ACAAATAGTTACCTTGAACATCAAACGTG....3'

#2 sequence 5'...GGACGACGGTATGGAATAGAGAGGAAGTTCCTC...3'

#3 sequence 5'...GATTTGTAATGTAAGTGAATAAGGAATACCATCTA...3'

#4 sequence (bottom) 5'... CAGCTGGGTGGTCATATGGCCATGGAG...3'

#4 sequence (middle) 5'...GAGACGGAGCTGTAGGTAAACTTGCTACTGATCAGTT...3'

EDVO-TECH Service

1.800.EDVOTEK

Mon. - Fri. 8am-5:30pm EST



Please Have the Following Info:

- Product Number & Description
- Lot Number on Box
- Order/Purchase Order #
- Approx. Purchase Date

Fax: 202.370.1501 • info@edvotek.com • www.edvotek.com

www.edvotek.com

- Online Catalog
- Order Products
- Experiment Protocols
- Tech Support
- Resources!



1.800.EDVOTEK • Fax 202.370.1501 • info@edvotek.com • www.edvotek.com

Duplication of any part of this document is permitted for non-profit educational purposes only. Copyright © 2000-2015 EDVOTEK, Inc., all rights reserved. 340.151217



**Please refer to the kit
insert for the Answers to
Study Questions**