EDVOTEK®

The Biotechnology Education Company ®

EDVO-Kit #

340

DNA Informatics

EXPERIMENT OBJECTIVE:

In this experiment, DNA sequences will be submitted to databank searches using the World Wide Web to identify genes and gene products.

EDVOTEK, Inc. • 1-800-EDVOTEK • www.edvotek.com

EVT 2010_0911AM

Table of Contents

Experiment Components

This experiment contains a total of 3 sets of 4 autorads of automated gel run sequence printouts.

- Three sets of four autorads

Requirements

- Computer with Internet access
- Autorad visualization source (White light box)

Background Information

Bioinformatics is a new field of biotechnology that is involved in the storage and manipulation of DNA sequence information from which one can obtain useful biological information. Almost routinely, data from DNA sequence analysis is submitted to Data bank searches using the World Wide Web (WWW) to identify genes and gene products.

For sequence analysis, four separate enzymatic reactions are performed, one for each of the four nucleotides. Each reaction contains the DNA Polymerase, the single-stranded DNA template to be sequenced to which a synthetic DNA primer has been hybridized, the four deoxyribonucleotide triphosphates (dATP, dGTP, dCTP, dTTP), often an isotopically labeled deoxynucleotide triphosphate, such as 32P or 35S dATP, and the appropriate DNA sequencing buffer. The reactions contain the dideoxytriphosphate reactions as follows: the "G" reaction contains dideoxyGTP, the "C" reaction dideoxyCTP, the "A" reaction dideoxyATP, and the "T" reaction dideoxyTTP. The small amounts of dideoxynucleotide concentrations are carefully adjusted so they are randomly and infrequently incorporated into the growing DNA strand.

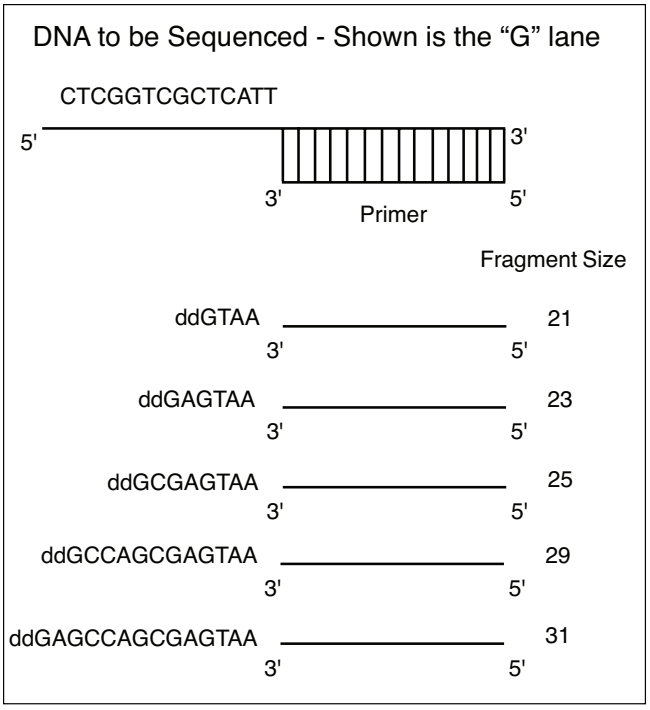


Figure 1

Once a dideoxynucleotide is incorporated into a single strand, DNA synthesis is terminated since the modified nucleotide does not have a free 3'hydroxyl group on the ribose sugar which is the site of the addition of the next nucleotide in the DNA chain. The incorporation of the dideoxynucleotide allows the generation of the nested DNA fragments and makes possible to determine the position of the various nucleotides in DNA. A particular reaction will contain millions of growing DNA strands, and therefore "nested sets" of fragments will be obtained. Each fragment is terminated at a different position corresponding to the random incorporation of the dideoxynucleotide.

As an example, in "nested sets" of fragments produced for a hypothetical sequence in the "G" reaction contains dATP, dCTP, dGTP, dTTP, DNA polymerase, DNA sequencing buffer, 32P-labeled dATP and dideoxyGTP. (Fig. 1) As can be seen, ddGTP (dideoxyGTP) incorporation randomly and infrequently will produce a "nested set" of fragments which terminate with a ddGTP. The "nested set" is complimentary to the region being sequenced. Similar "nested sets" are produced in the separate "A", "T", and "C" reactions. For example, the "A" "nested set" would terminate with a ddATP.

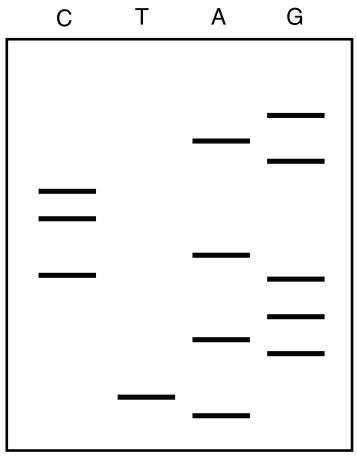
It should be readily apparent that together the "G, A, T, C" "nested sets" contain radioactive 32P-labeled fragments ranging in size successively from 19 to 31 nucleotides for the hypothetical sequence in Figure 2.

Background Information

As shown in the figure, the "G" reaction contains fragments of 21, 23, 25, 29 and 31 nucleotides in length. Seventeen of these nucleotides are contained in the synthetic DNA sequencing primer. The rest are added during de novo DNA synthesis.

The products from the G, A, T, and C reactions are separated by a vertical DNA polyacrylamide gel. Well # 1 contains the "G" reaction; well # 2 the "A" reaction; well # 3 the "T" reaction; and well # 4 the "C" reaction. It is important to note that the strand being sequenced will have the opposite Watson/Crick base. As an example, the G reaction in tube one will identify the C nucleotide in the template being sequenced. 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 labeled with 32P, their position can be detected by a dark exposure band on the sheet of x-ray film. In addition to 32P or 35S-deoxynucleotide triphosphates used in DNA sequencing, non-isotopic methods of using fluorescent dyes and automated DNA sequencing machines are beginning to replace the traditional isotopic methods. For a given sample well, the horizontal "bands" appear in vertical lanes from the top to the bottom of the x-ray film. Generally, a single electrophoretic gel separation can contain several sets of "GATC" sequencing reactions.

Figure 2 shows an autoradiograph which would result from analysis of the hypothetical sequence in Figure 1. The dark bands are produced by exposure of the x-ray film with 32P which has been incorporated into the dideoxy-terminated fragments during DNA synthesis. The sequence deduced from the autoradiogram will actually be the complement of the DNA strand contained in the singled-stranded DNA template. This DNA sequencing procedure is called the Sanger "dideoxy" method named after the scientist who developed the procedure.



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

Figure 2

Data from DNA sequencing is of limited use unless it can be converted to biologically useful information. Bioinformatics therefore is a critical component of DNA sequencing. It evolved from the merging of computer technology and biotechnology. The widespread use of the internet has made it possible to easily retrieve information from the various genome projects. In a typical analysis, as a first step, after obtaining DNA sequencing data a molecular biologist will search for DNA sequence similarities using various data banks on the WWW. Such a search may lead to the identification of the sequenced DNA or identify its relationship to related genes. Protein coding regions can also be easily identified by the nucleotide composition. Likewise, noncoding regions can be identified by interruptions due to stop codons. The functional significance of new DNA sequences will continue to increase and become more important as sequence information continue to be added and more powerful search engines become readily accessible.

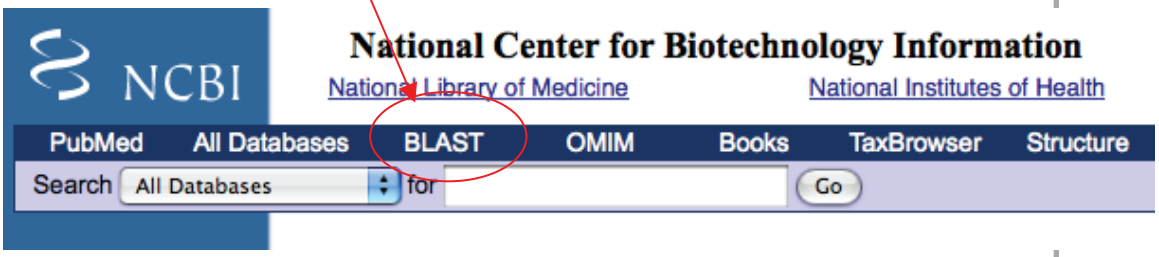
Background Information

At this time, various research groups from around the world are engaged in determining the complete Human Genome sequence. Advances in DNA sequencing and bioinformatics will soon make it possible to use information from the Human Genome Project as a clinical diagnostic tool. It should be noted that several smaller genomes such as that for *Saccharomyces cerevisae* and *Helicobacter pylori* have already been completed.

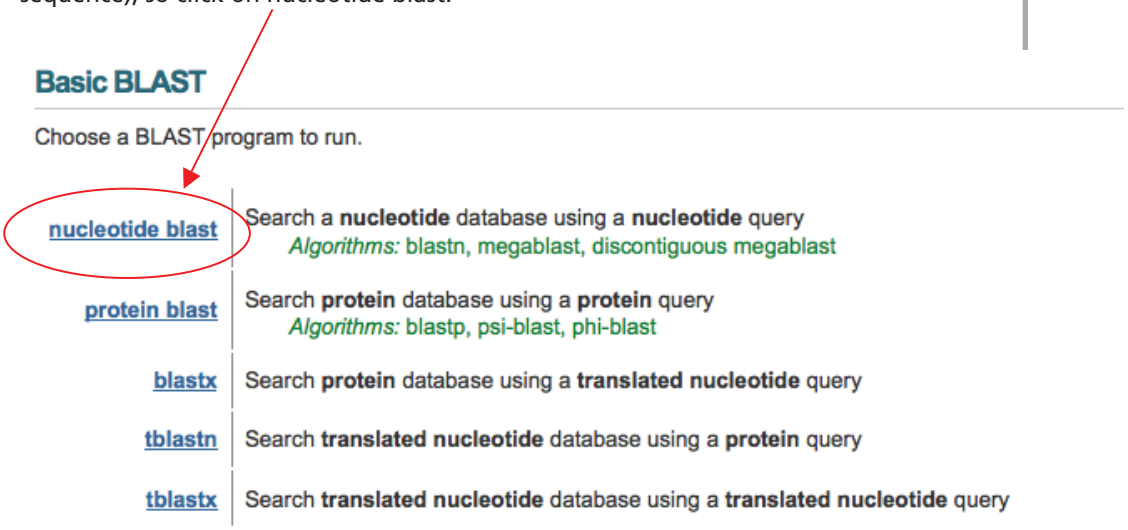
The purpose of this exercise is to introduce students to bioinformatics. In order to gain experience in database searching, students will utilize the free service offered by the National Center for Biotechnology (NCBI) which can be accessed on the WWW. At present there are several Databases of GenBank including the GenBank and EMBL nucleotide sequences, the nonredundant GenBank CDS (protein sequences) translations, and the EST (expressed sequence tags) database. Students can use any of these databases as well as others available on the WWW to perform the activities in this lab. For purposes of simplification we have chosen to illustrate the database offered by the NCBI. These exercises will involve using BLASTN, whereby a nucleotide sequence will be compared to other sequences in the nucleotide database. BLASTP will also be used to compare the amino acid sequence of a protein with other protein sequences in the databank.

Experiment Procedures

1. Type:

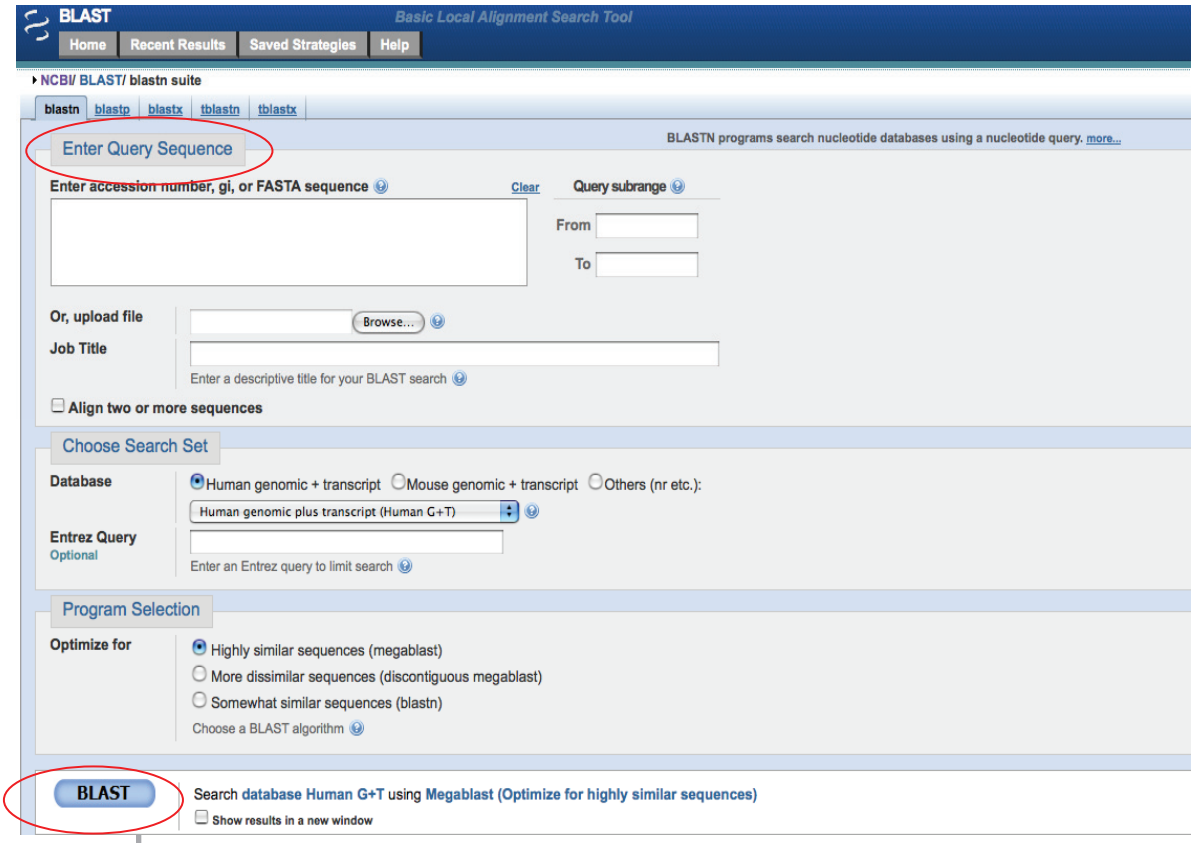


2. The following new menu should now appear. This page presents us with choices for which database is being searched, and which variety of BLAST to use for the searching. For our purpose, which is a nucleotide sequence (not a protein amino acid sequence), so click on nucleotide blast.

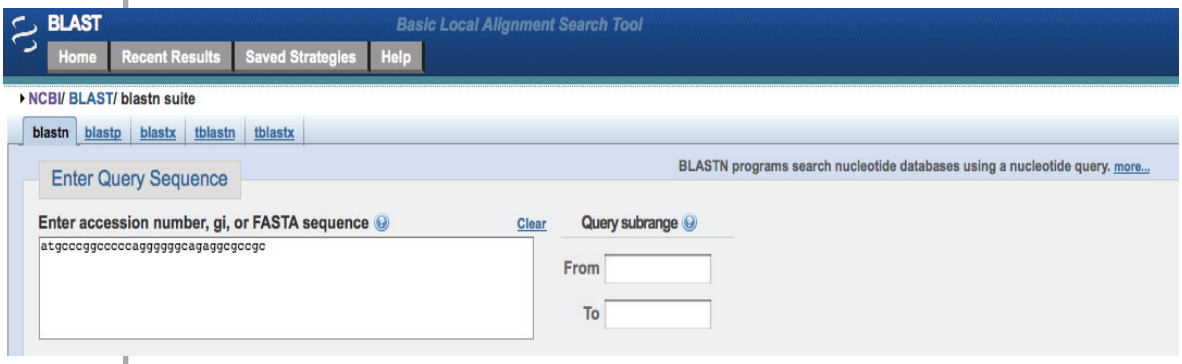


3. Next, under Nucleotide BLAST, click on Standard Nucleotide- Nucleotide BLAST [blastn]. The other options are more complicated than needed for our specific application. The new screen should appear as below. There are three selections on this screen: Enter Query Sequence, Choose Search Set, and Program Selection.

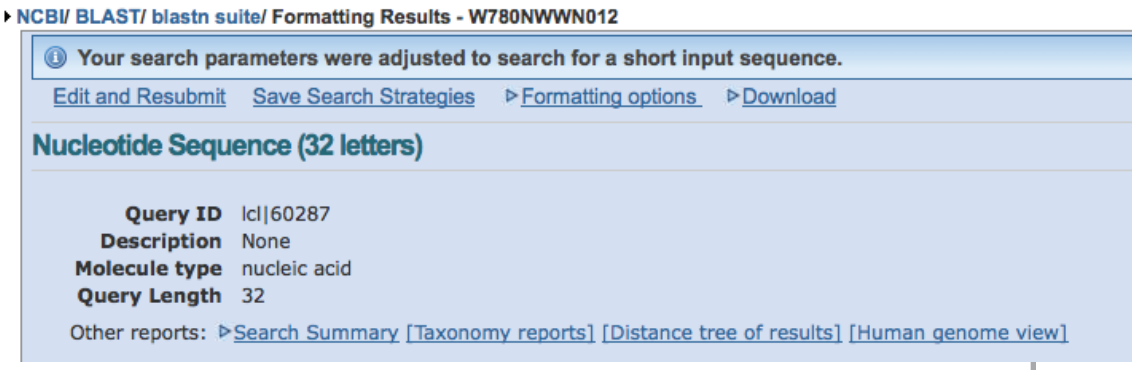
Experiment Procedure



4. To get started entering the nucleotide sequence, start typing the sequence (the sequence can be in either capitals or non-capital letters) shown in screen above under Enter Query Sequence. Be careful to type exactly what is shown in the box! Use the default selections and then click on the BLAST query box.



5. Once the BLAST query box has been clicked, you will receive a reply regarding the status of your search results. Once your entry is in the BLAST Queue, you will be assigned an ID# that can be used to check your results at a later time ("Retrieve results for an existing Request ID" under the BLAST menu).



6. Sometimes the search is busy. If your results are not ready at this point, try again a bit later or log on at another time and use the same ID number. Below is an example of what you might expect to find from your example search:

```
>ref|NM_007061.3| GM Homo sapiens CDC42 effector protein (Rho GTPase binding) 1 (CDC42EP1),
transcript variant 2, mRNA
Length=2131

 GENE ID: 11135 CDC42EP1 | CDC42 effector protein (Rho GTPase binding) 1
[Homo sapiens] (Over 10 PubMed links)


 Score = 63.9 bits (32),  Expect = 5e-09
 Identities = 32/32 (100%), Gaps = 0/32 (0%)
 Strand=Plus/Plus

Query  1    ATGCCCGGCCCCCAGGGGGGCAGAGGCGCCGC  32
            ||||||||||||||||||||||||||||||||
Sbjct  374  ATGCCCGGCCCCCAGGGGGGCAGAGGCGCCGC  405


>ref|NM_152243.1| GM Homo sapiens CDC42 effector protein (Rho GTPase binding) 1 (CDC42EP1),
transcript variant 1, mRNA
Length=2135


 Score = 63.9 bits (32),  Expect = 5e-09
 Identities = 32/32 (100%), Gaps = 0/32 (0%)
 Strand=Plus/Plus

Query  1    ATGCCCGGCCCCCAGGGGGGCAGAGGCGCCGC  32
            ||||||||||||||||||||||||||||||||
Sbjct  375  ATGCCCGGCCCCCAGGGGGGCAGAGGCGCCGC  406
```

One can see by inspecting the BLASTN search report that our query sequence shows the best match to human CDC42 effector protein. Specifically, CDC42 showed the highest score. Inspection of the two sequences alignment shows that our query of 32 nucleotides was identical to the nucleotide segment within CDC42. As a general rule, identical nucleotide sequence spanning greater than 21 bp between two samples, usually indicates that the sequences are related or identical. One clear exception to this rule, is a long stretch of A's or T's, which may correspond to the Poly (A+) tail found at the 3'-end of all eukaryotic mRNAs.

Experiment Procedures

EXERCISE 1

Familiarize yourself with the autorads by reading the DNA sequence for sample #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.
- 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 directly the sequence into the query box or first write the sequence down on a piece of paper.
- 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.
- Notice that for the most part that the spacing and intensity of most of the bands is fairly constant. Ignore lightly colored bands choosing 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.

Results to be obtained for Sample 1:

a. What are the names of these genes?
b. What species do these genes belong to?

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.

Now, begin the exercise by reading the DNA sequence for sample number 2 approximately 6 cm from the bottom of the strip. It should read as follows:

5'...GGACGACGGTATGGAATAGAGAGGAAGTTCCTC...3'

Experiment Procedures

Submit it to NCBI using the BLASTN program.

- 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).
- If an exact band at a position is ambiguous, you can enter an N which denotes it could be either A, C, G or T.

After receiving the BLAST results, scroll down to look at the entries that have nucleotide matches with your query sequence. Again remember that when two sets of nucleotide sequences show identity of a 21 basic pair continuous strand, it usually indicates that the sequences are related or identical.

Results to be obtained for Sample 2:

a. What is the name of this gene?
b. Compared to the gene bank entry, what strand have you read?

EXERCISE 3

- DNA entries can be further accessed by clicking on the Genbank accession number.
- The information shown describes the DNA sequence and/or gene, the contributing scientist's name and information such as the protein and the amino acid sequence for which it codes.
- Read the DNA sequence from sample number 3. Start at the bottom of the strip and record the DNA sequence.

Results to be obtained for Sample 3:

a. What is the name of this gene?
b. Approximately how many amino acids does this gene have?

Experiment Procedures

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 protein-protein interactions of the signaling molecules. This process, called signal transduction, is of central importance in many biological processes, such as cell division, cytoskeleton formation, etc.

- Read the DNA sequence obtained from sample number 4. Start at the bottom of the strip and record the DNA sequence.
- Next, move approximately a third of the way up the strip and read a portion of this section of the DNA sequence.
- Submit each sequence section individually to the BLASTN program.

Results to be obtained for Sample 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?

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