The Biotechnology Education Company®

EDVOTEK®

339

EDVO-Kit #

Sequencing the Human Genome

Experiment Objective:

In this experiment, DNA sequences obtained from automated sequencers will be submitted to Data bank searches using the World Wide Web to identify genes and gene products. The impact of Genomics will be discussed in the context of our society in the 21st century.

Major Section Headings

Experiment Components

This experiment contains a total of ten automated gel run sequence printouts. Students can use any Human Genome sequence database to perform the activities in this lab. For purposes of simplification we have chosen to illustrate the database offered by the NCBI.

Requirements

- Internet access (WWW)

BACKGROUND INFORMATION

Bioinformatics

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. Although DNA sequencing has existed since the early 1970's, it has not been until the 1990's that the whole process has been automated. In particular, automated DNA sequencers rapidly and efficiently analyze the reactions in a one-lane sequencing process that uses four-dye fluorescent labeling methods and a real-time scanning detector. These machines automatically separate the labeled DNA molecules of varying sizes by gel electrophoresis and also "call" the bases and record the data. In contrast to running and reading the DNA sequencing gels manually, these automated sequencers can provide much more information (up to several thousands of base pairs) per gel run.

The entire process of collecting and analyzing sequencing data is automated. Robots perform the sequencing reactions, which are then loaded onto automated sequencers. After the automated sequencing run is complete, the sequence information is transferred to computers, which analyze the data. This highly efficient automated DNA sequencing process has produced many large-scale DNA sequencing efforts creating a new field of biology called genomics. Genomics involves using DNA sequence information to understand the biological complexity of an organism. The Human Genome Project (HGP) will furnish a complete human genetic blueprint by the year 2002. The goal of the HGP is to determine the complete nucleotide sequence of human DNA and thus localizing the estimated 80,000-100,000 genes within the human genome. Advances in DNA sequencing and bioinformatics will soon make it possible to use information from the Human Genome Project as a clinical diagnostic tool.

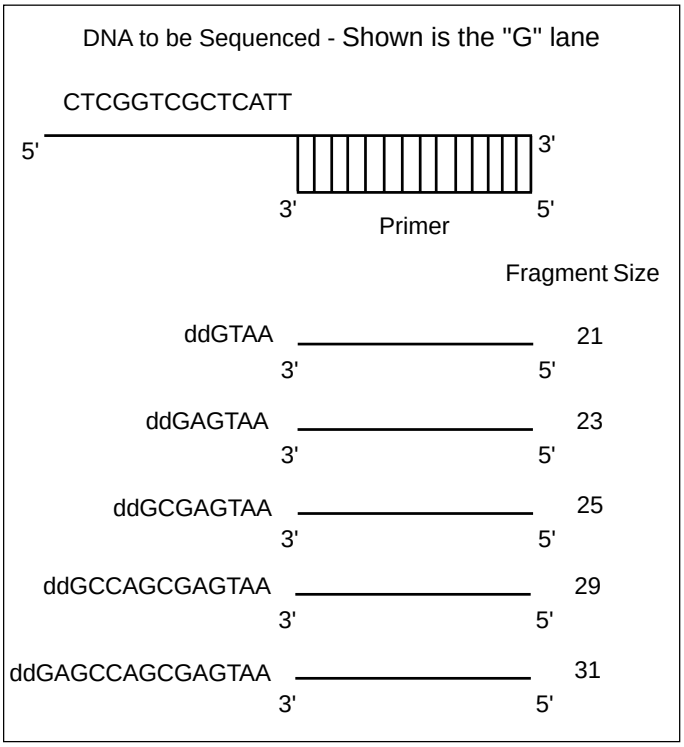


Figure 1

In addition to the human genome, some of the first genomes to be sequenced are those of microbes. Information about genes in microbes represents new leads for developing new therapeutic agents. It should be noted that several smaller genomes such as that for *Saccharomyces cerevisae* and *Helicobacter pylori* have already been completed. Additional efforts are ongoing for sequencing the genomes of other organisms that are used exten-

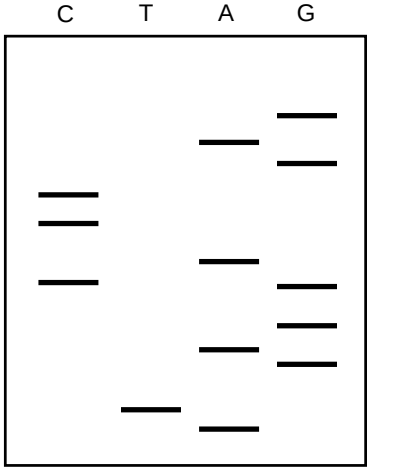
BACKGROUND INFORMATION

Background Information, continued

sively in research laboratories as model systems (e.g. mice) or for commercial reasons (e.g. corn).

The genetic revolution will continue to yield new discoveries. While scientist continue to identify genes that cause disease or phenotypic differences (tall verses short), there is a growing danger to see humans merely as a sum of their genes. Understanding the ethical, legal, and social implications of genetic knowledge, and the development of policy options for public consideration are therefore yet another major component of the human genome research effort. For example, one particular area of debate is that of psychiatric disorders whereby researchers are trying to characterize traits such as schizophrenia, intelligence and criminal behavior purely in terms of genes. This simplistic view may create situations in which genetic information has the potential to cause inconvenience or harm. Additionally, ethical debate about prenatal screening of diseases in human embryos is also controversial. Thus in depth discussion is needed to balancing improvements to human health with the ethical implications of the genetic revolution.

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.



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

About DNA Sequencing

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

BACKGROUND INFORMATION

Background Information, continued

primer has been hybridized, the four deoxyribonucleotide triphosphates (dATP, dGTP, dCTP, dTTP), often an isotopically labeled deoxynucleotide triphosphate, such as ^{32}P or ^{35}S 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. 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, ^{32}P-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 ^{32}P-labeled fragments ranging in size successively from 19 to 31 nucleotides for the hypothetical sequence in Figure 2 .

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 con-

Background Information, continued

tact with a sheet of x-ray film. Since the DNA fragments are labeled with ^{32}P, their position can be detected by a dark exposure band on the sheet of x-ray film. In addition to ^{32}P or ^{35}S-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 ^{32}P 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.

The purpose of this exercise is to introduce students to Genomics and 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 non-redundant 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. For each of the four sequences, students should identify a potential human disease and discuss related bioethical issues.