Algorithms for Molecular Biology **Fall Semester, 1998**

Lecture 1: December 6, 1998

Lecturer: Ron Shamir *Scribe: Sigal Korczyn and Nimrod Hoofien* [1]

This background lecture will begin by addressing elementary biological background. Topics include the DNA, genes, chromosomes, and the transfer of genetic information from the DNA to proteins. We shall also review several biotechnological techniques in . A discussion of the type of questions addressed in the field of Bioinformatics will be presented with a description of some of the techniques used in solving them. This is supposed to mainly give a flavor of questions asked, and the chioce of problems is rather arbitrary.

1.1 Biological Background

1.1.1 Genetic information

Laws of inheritance

The basic laws of inheritance were discovered in 1866 by Gregor Mendel. He defined the concept of a *gene* - the basic unit responsible for oppression and passing on of a single characteristic. A period of over 75 years has passed from the time the laws of inheritance were discovered until the biological role of *DNA (Deoxy-Ribonucleic acid)* was elucidated. It is now known that the DNA is the major carrier of genetic material in living organisms.

DNA

The determination of the structure of DNA by Watson and Crick in 1953 is often said to mark the birth of modern molecular biology. The DNA is a large molecule composed of four basic units called *nucleotides*. Each nucleotide contains phosphate, sugar and one of the four bases: *Adenine, Guanine, Cytosine* and *Thymine* (usually denoted as A,G,C and T). The structure of DNA is described as a *double helix*. Each helix is a polymer of nucleotides chained together by phosphodiester bonds. The two helices are held together by hydrogen bonds. These bonds are formed by pairs of bases, with each base pair consisting of one *purine* base (A or G) and one *pyrimidine* base (C or T), paired according to the following rule: G pairs with C, and A pairs with T. The total length of the human DNA is about 3×10^9 base pairs (abbreviated *bp*).

[1] Based in part on a scribe by Roded Sharan, October 29, 1995, and on [3, 2, 1]. Figures taken from [3].

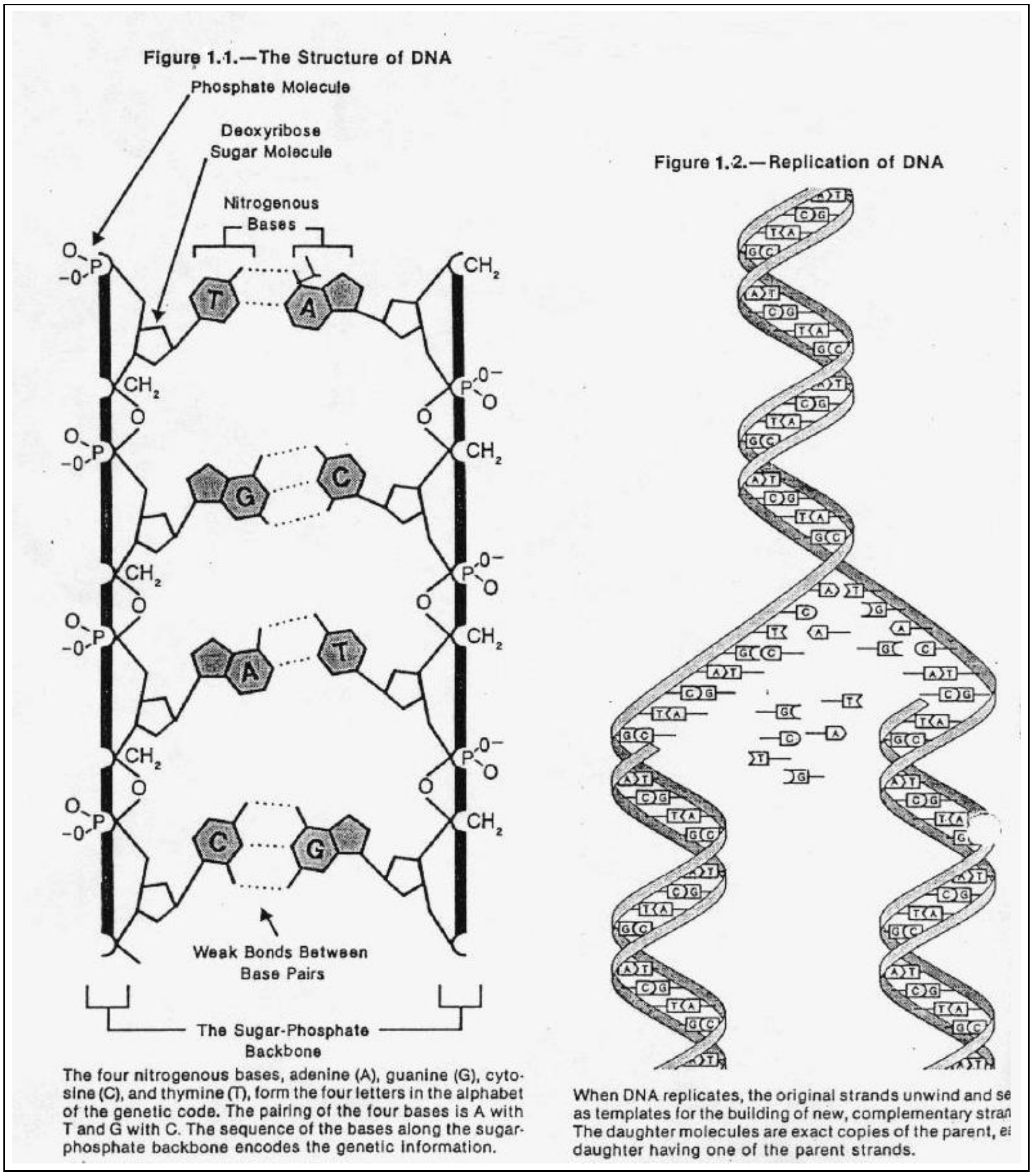
Phosphate Molecule
Deoxyribose Sugar Molecule
Nitrogenous Bases
T
A
G
C
A
T
C
G
O
-O-P
CH$_2$
O
P
O-
Weak Bonds Between Base Pairs
The Sugar-Phosphate Backbone

Figure 1.1.—The Structure of DNA

The four nitrogenous bases, adenine (A), guanine (G), cytosine (C), and thymine (T), form the four letters in the alphabet of the genetic code. The pairing of the four bases is A with T and G with C. The sequence of the bases along the sugar-phosphate backbone encodes the genetic information.

Figure 1.2.—Replication of DNA

When DNA replicates, the original strands unwind and se as templates for the building of new, complementary stran The daughter molecules are exact copies of the parent, ea daughter having one of the parent strands.

Figure 1.1: DNA

Genes and Chromosomes

Eukariotic cell nuclei contain *chromosomes* - the contiguous structures in which DNA is stored. Every somatic cell usually includes two copies of each chromosome (excluding the sex X,Y chromosomes). The number of chromosomes varies among species. Humans have 22 pairs of chromosomes, plus the sex chromosomes.
A *gene* is a specific sequence of nucleotide bases along a chromosome carrying information for constructing a protein. Genes are parts of the cromosomes. In humans genes constitute approximately $5-10\%$ of the DNA, leaving $90-95\%$ of noncoding *"junk DNA"*. The role of the latter is as yet unknown, but it is speculated that it plays a very important role. Certain theories have been suggested, such as physically fixing the DNA in its compressed position, preserving old genetic data, etc.

RNA

Cells have a second type of nucleic acid - *RNA (Ribonucleic Acid)* which can also carry genetic information. Unlike DNA, which is located primarily in the nucleus, RNA can also be found in the cell's cytoplasm. Like DNA, RNA is also built from purine and pyrimidine nucleotides (*Uracil* taking the place of Thymine), but forms a single helix (unlike the DNA's double helix).

The *messenger RNA (mRNA)* carries genetic information from the DNA to the *ribosomes* - the intra-cellular constructs where it is translated into a protein. mRNA is synthesized in the nucleus based on a single DNA strand, using the *RNA polymerase* enzyme. mRNA is *transcribed* from a DNA strand only in locations called *open reading frames*. When such transcription occurs, the two DNA strands are split apart, and one of them serves as a mold for the generated mRNA molecule, which is complementary to this strand, and therefore, a replica of the other one. Consecutive triplets of mRNA bases, called *codons*, each determine a certain amino acid (see below). In eukaryotes, the mRNA is formed of *coding* and *non-coding* regions. Coding regions are the regions used to carry real genetic information. Non coding regions do not carry such information (see below).

The coding regions are called *exons*, since they are able to leave the nucleus and reach the Ribosome. The non-coding regions are called *introns* and never leave the nucleus.

After being synthesized from a DNA strand, the exons of the RNA molecule are merged together, exiting the nucleus, heading for a ribosome.

The *transfer RNA (tRNA)* is a small RNA molecule serving as an adapter between mRNA and amino acids. The molecule is composed of two parts. On one part the tRNA holds an *anticodon*. The anticodon is a sequence of three RNA bases. On the other side, the tRNA holds an amino acid. The many-to-one mapping from anticodons to amino acids defined by tRNA molecules is the *universal genetic code* (see below).

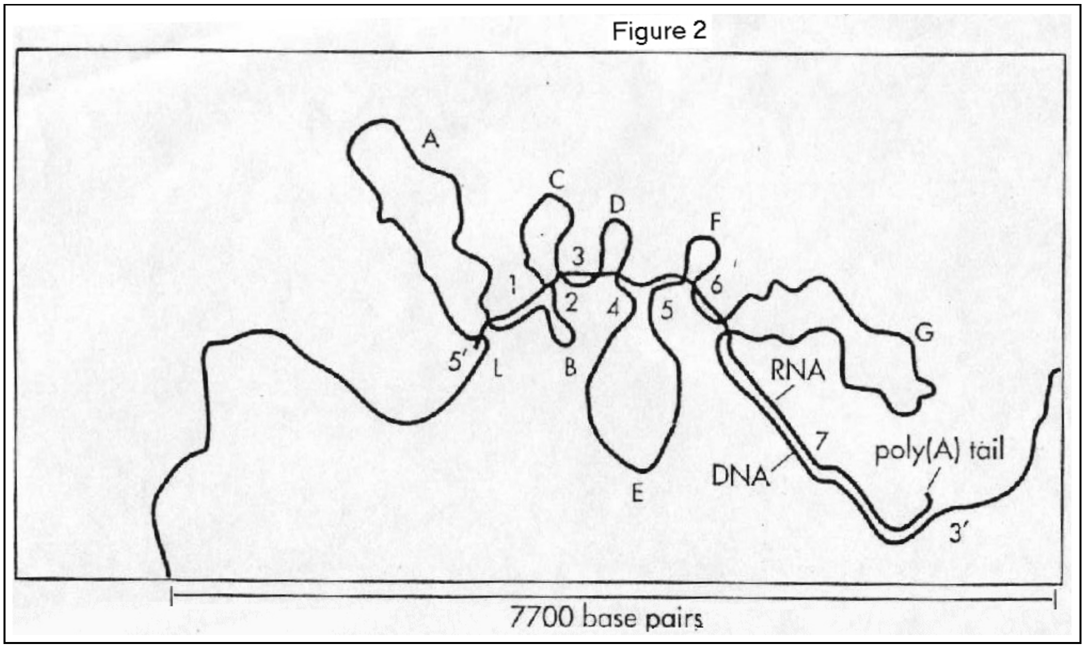

Figure 1.2: coding and non-coding sections of DNA

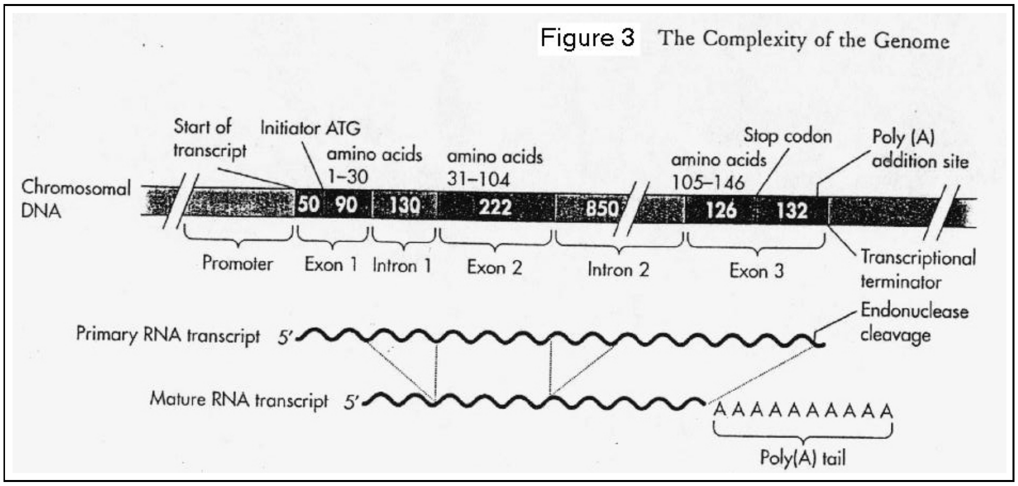

Figure 1.3: exons and introns

DNA replication

In addition to being translated into proteins, the genetic information embodied in DNA can also be replicated into more DNA. When DNA is replicated, its two strands are split and each strand serves as a template for a new forming strand: Each base in the template dictates the complementary base in the new strand. The replication reaction is catalyzed by the enzyme *DNA polymerase*. This enzyme can extend a chain, but cannot start a new one. Therefore, DNA synthesis must first be initiated with a *primer*, a short oligonucleotide (nucleotide chain). The oligonucleotide generates a segment of duplex DNA that is then turned into a new strand by the replication process (see figure 1.1).

Mutations

Mutations are local changes in the DNA content, caused by inexact replication. Mutations occur often due to radiation and other environmental conditions. *Substitution* occurs when one base is replaced by another. *Insertion* and *deletion* are the addition and removal of one or more bases, respectively. Substitution, as well as insertion or deletion of a single base is called point mutation. A *rearrangement* is a change in the order of complete segments along a chromosome.

Mutations are important to us for several reasons. They are responsible for inherited disorders and other diseases such as cancer, that involve alterations in the genes. At the same time, mutations are the source of phenotypic variation on which natural selection acts, creating speciesand changing them. For example, the human and the mice genomes are very similar. The major difference between them is the internal order of DNA segments. A complex problem that arises is:

The Genome Rearrangement Problem

Problem 1 *Given two permutations of a set of genomic segments, find the minimal set of operations to transform one permutation into the other.*

Rearrangement events are much less frequent than point mutations. For example, substitutions occur in some organisms about 10 times in each generation. A non fatal rearrangement event occurs once every 5 to 10 million years. By discovering what rearrangement events have occurred, and what was their order of occurrence, there is a chance to get a better understanding of the evolutionary process.

Protein Coding and Ribosomes

The process of *translating* DNA into an active protein has two phases: The first is *transcribing* a DNA open reading frame into an mRNA molecule. The mRNA is synthesized from one strand of double stranded DNA helix. This transcription is catalyzed by the *RNA polymerase* enzyme, using the DNA strand as a template for creating the RNA strand. The mRNA leaves the nucleus, leaving the introns behind, heading for a ribosome. The process of synthesizing a protein from an RNA molecule is accomplished by tRNA. These molecules carry the amino acids to the ribosome, where they can be attached to a growing chain of amino acids, a.k.a. a *polypeptide*. The Ribosome moves along the mRNA, so that successive codons are brought into position for their respective amino acids.

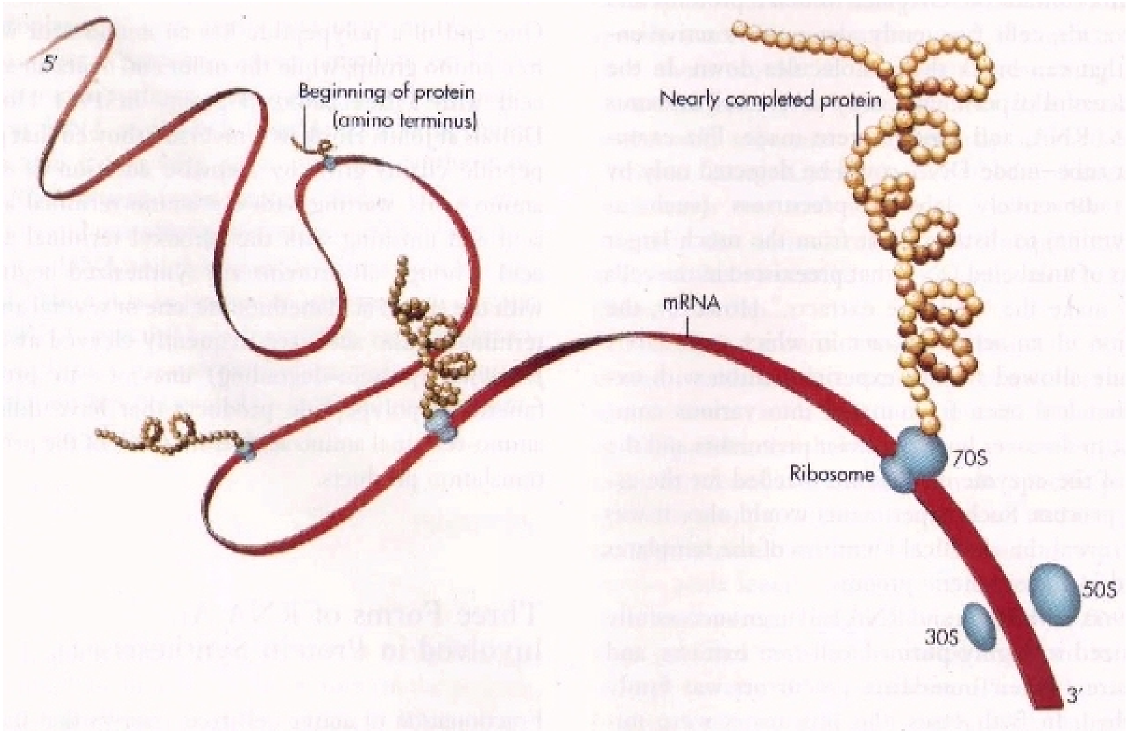

FIGURE 4
Messenger RNA carries genetic information from the DNA to the ribosomes, where it is translated into protein. The polypeptide chains are elongated as ribosomes move along the mRNA molecules, with the 5' ends of the mRNA being translated first.

Figure 1.4: protein coding

The Genetic Code

The universal genetic code is the logical mapping that specifies how genetic information stored in DNA and mRNA determines protein sequence. It was discovered by Watson and Crick and since then was found to be common to all living organisms, with only minor and rare differences. Each triplet of bases is called a *codon*, and codes for a specific amino acid.

There are also special codons, called *stop codons* , which are used to signal the termination of the protein synthesis process. Since there are 64 possible codons (minus three stop codons), and only 20 amino acids, different codons may be used to code for the same amino acid.

The Gene Finding Problem

Problem 2 *Given a DNA sequence, predict the location of genes (open reading frames), exons and introns.*

A simple solution might be seeking stop codons in the section. Clearly, if several stop codons exists close to each other in a section, the section cannot be a gene, since it would have been terminated. When a relatively long sequence does not contain stop codons, it becomes more probable that it contains a gene. The problem becomes more complex in eukaryotic DNA due to the existence of interleaved exons and introns. In that case, a stop codon does not indicate that the sequence is not in a gene, but merely that the sequence is not in an exon. Further complications arise from the fact that a certain DNA sequence can be interpreted in 6 different ways: 3 different offsets for each of the possible 'starting points' (the reading frame of the codons) and two for the reading directions. It is safe to assume that in most cases, apart from prokaryotic species, a DNA section will encode only one gene.

Proteins

Proteins are organic molecules which are resposible to most checmical reactions performed in the cell, and thus, are essential for all cell functions. A protein is a polypeptide - a macromolecule composed of building blocks called amino acids attached end to end in a linear string. There are 20 amino acids, with an average protein containing about 200 amino acids. Proteins have a complex structure, which can be thought of as having four logical levels. The amino acid sequence of a protein's chain is called its *primary structure.* Different regions of the sequence form local regular *secondary structures*, such as *α-helices* and *β-sheets.* The *tertiary structure* is formed by packing such structures into one or several *domains.* The final, complete, protein may contain several protein chains arranged in a *quaternary structure.*

The whole complex structure (primary to quaternary) is determined solely by the primary sequence of amino acids (and therefore, is defined by the genetic material itself).

The Protein Folding Problem

Problem 3 *Given a sequence of amino acids, predict the 3D structure of the protein.*

Since the functionality of the protein is determined by its 3D structure, it would be very helpful to be able to predict the protein's structure, thus helping to understand its role and

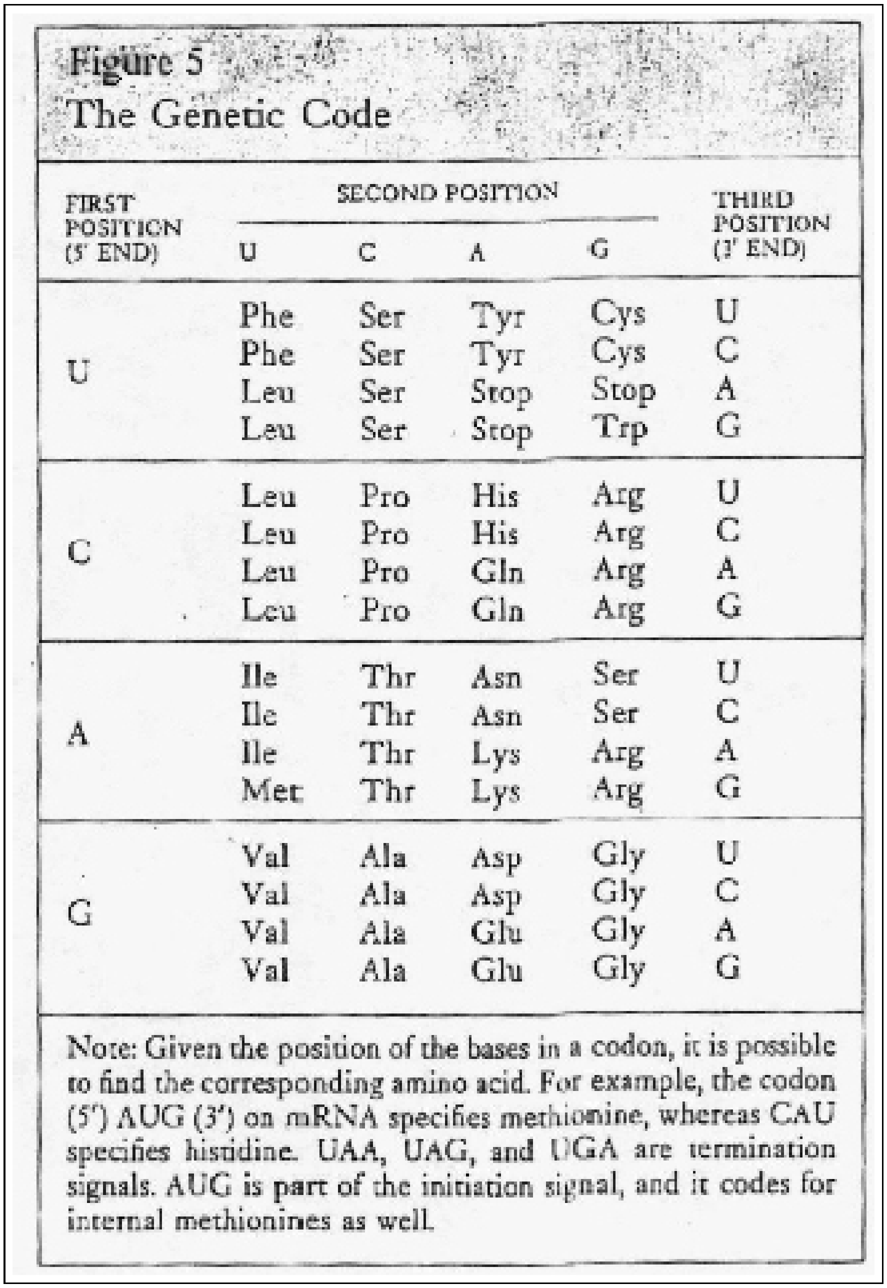

Figure 5

The Genetic Code

FIRST POSITION (5' END)	SECOND POSITION				THIRD POSITION (3' END)
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Note: Given the position of the bases in a codon, it is possible to find the corresponding amino acid. For example, the codon (5') AUG (3') on mRNA specifies methionine, whereas CAU specifies histidine. UAA, UAG, and UGA are termination signals. AUG is part of the initiation signal, and it codes for internal methionines as well.

Figure 1.5: the genetic code

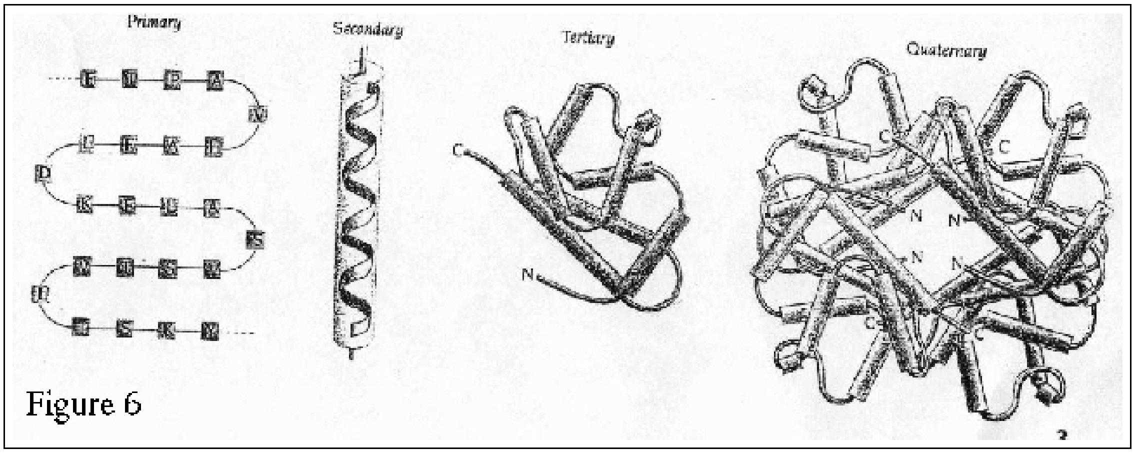

Figure 1.6: protein structure

responsibilities in the cell. There are several approaches towards this problem. The first, called *homology*, uses a protein database to search for similar sequences of proteins. If such a protein is found (a protein with 25% of the amino-acids in its sequence being identical to the original is usually similar enough), it is quite safe to assume that the two proteins will have the same structure. The second approach, *threading*, classifies known structures into families with similar foldings. Given a sequence of amino acids, we select the family to which the given sequence is most likely to fold. A third approach, *de novo* (from Scratch), tries to find the structure which minimizes the energy of the protein.

1.2 Biological Technology

The main motivation for biotechnology is to use 'biological machines' to fulfill human needs. Using biotechnological techniques allows us to produce large quantities of substances necessary for medical and other purposes.

Restriction Enzymes

One of the basic tools used in biotechnology is *restriction enzymes.* In natural circumstances, one of the main roles of these enzymes is to break foreign DNA entering the cell. A restriction enzyme breaks the phosphodiester bonds of a DNA upon appearance of a certain cleavage sequence. Each such enzyme is characterized by a different cleavage sequence. Today there are more then 150 known different cleavage sites, namely, different nucleotide configurations that known enzymes can digest. Application of restriction enzymes to a sequence is called *digestion.*

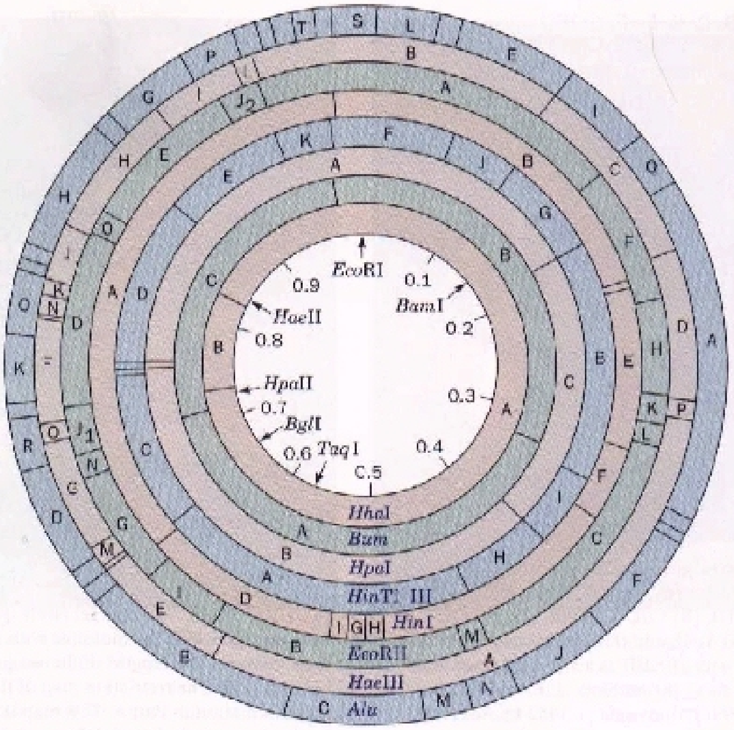

FIGURE 7 A restriction map for the 5243-bp circular DNA of SV40. The central circle indicates the fractional map coordinates with respect to the single *EcoRI* restriction site. The letters A, B, C, . . . in each ring represent the various restriction fragments of the corresponding restriction enzyme order of decreasing length. [After Nathans, D., *Science* 206, 90 (1979).]

Figure 1.7: restriction map

Gel Electrophoresis

Gel electrophoresis is a technique used to separate a mixture of digested DNA fragments. An electrical field is used to move the negatively charged DNA molecules through porous agarose gel. Fragments of the same size and shape move at the same speed, and because smaller molecules travel faster then larger molecules, the mixture is separated into bands.

The amount of exposure the DNA receives to restriction enzymes determines the portion of possible sites that were actually separated. Therefore, by applying different exposures to the same DNA sequence, we can measure all possible lengths of DNA fragments, that one can obtain using a particular enzyme. From this information we can attempt to find out where the sites are located in the original molecule. This problem is known as:

The Complete Digest Problem

Problem 4 *Given a set of distances* $\{|X_i - X_j|\}$ $1 \leq i < j \leq n$, *reconstruct the original series* $X_1, \ldots, X_n$.

The complexity of this problem is yet unknown.

Sequencing

Sequencing is the operation of determining the nucleotide sequence of a given molecule. There are There are several approaches to sequencing, but generally, the most successful is based on gel electrophoresis. As mentioned earlier, the *DNA polymerase* enzyme catalyzes the replication reaction of DNA. DNA polymerase extends the chain by adding nucleotides to its end. Current biotechnology enables synthesis of nucleotides which cause the strand to terminate. For instance, A* denotes an *Adenine* molecule which does not allow other molecules to extend the strand after itself. By catalyzing DNA replication in an environment containing mixtures of normal Adenine and sythesized Adenine* instead of only Adenine, it is possible to create DNA strands of different lengths. By applying gel electrophoresis to these molecules, it is possible to determine the lengths of all the strings and from it to conclude the location of all Adenines in the tested DNA strand. In a similar fashion it is possible to locate other nucleotides and eventually to fully sequence a whole segment of DNA. Using this method, sequences of 500-800 nucleotides can be mapped.

The problem that arises from this sequencing technique is the creation of a long DNA chain from the local sequences. This problem is known as:

The Sequence Assembly problem

Problem 5 *Given a set of sub-strings, find the minimal string containing all of the members of the set.*

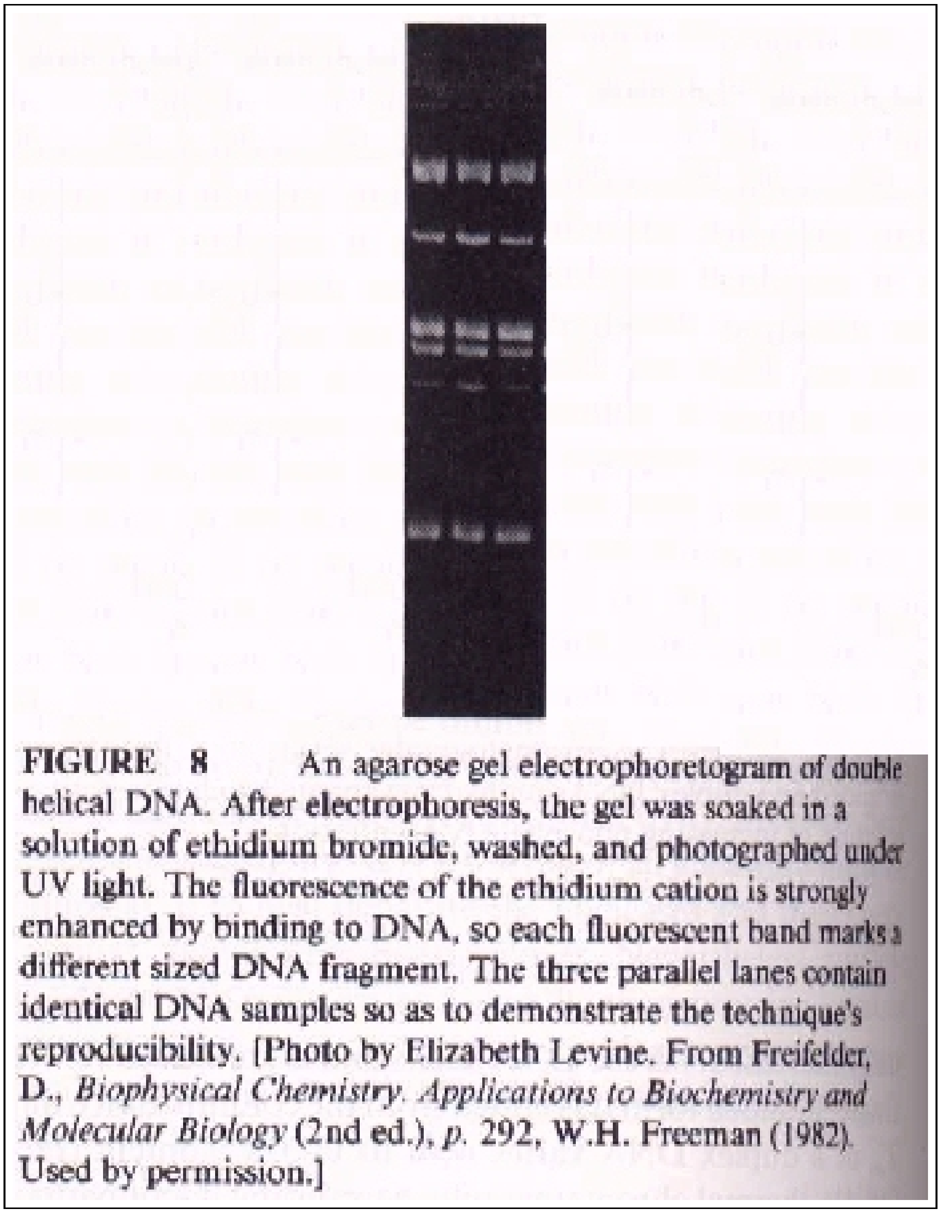

FIGURE 8 An agarose gel electrophoretogram of double helical DNA. After electrophoresis, the gel was soaked in a solution of ethidium bromide, washed, and photographed under UV light. The fluorescence of the ethidium cation is strongly enhanced by binding to DNA, so each fluorescent band marks a different sized DNA fragment. The three parallel lanes contain identical DNA samples so as to demonstrate the technique's reproducibility. [Photo by Elizabeth Levine. From Freifelder, D., *Biophysical Chemistry. Applications to Biochemistry and Molecular Biology* (2nd ed.), *p.* 292, W.H. Freeman (1982). Used by permission.]

Figure 1.8: gel electrophoresis

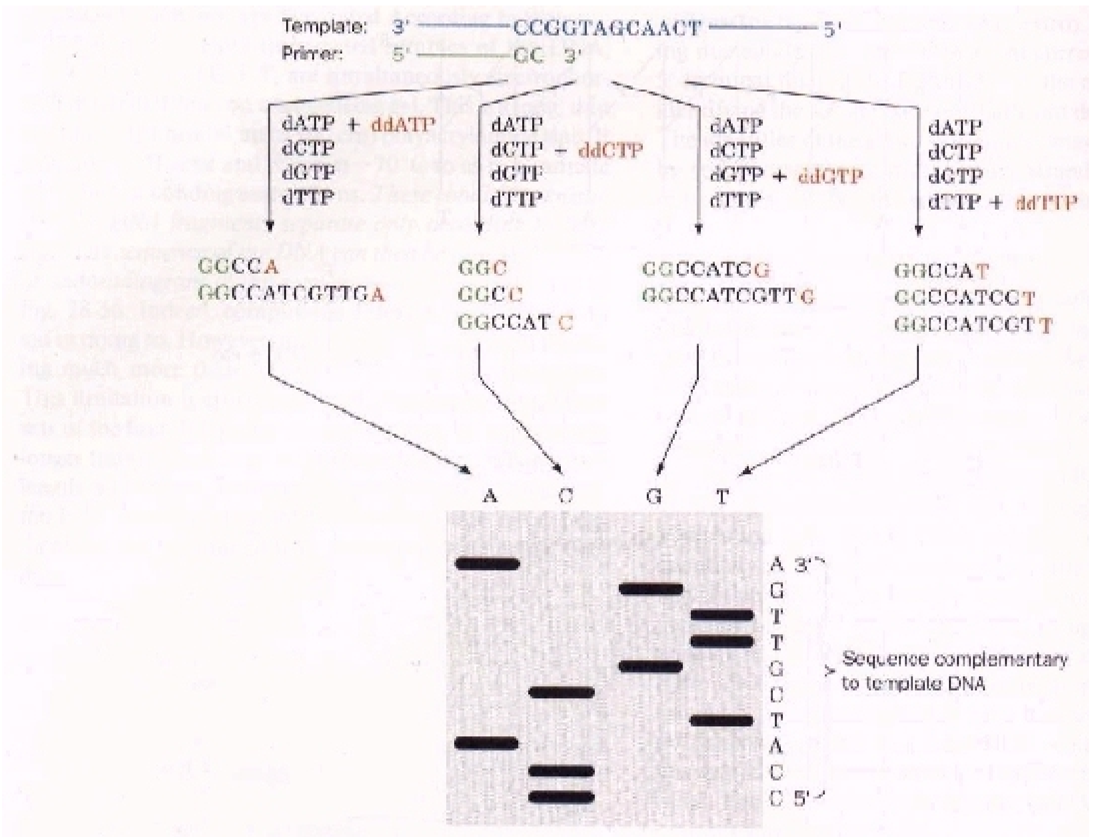

FIGURE 9 . A flow diagram of the chain-terminator (dideoxy) method of DNA sequencing. The symbol ddATP represents dideoxyadenosine triphosphate, *etc.*

Figure 1.9: DNA sequencing

Although this problem may seem simple, it is known to be NP-Complete. However, there are greedy algorithms which perform fairly well in practice.

Cloning

A major problem in biochemical research is obtaining sufficient quantities of the substance of interest. These difficulties have been largely eliminated in recent years through the development of molecular *cloning* techniques. The clone is a collection of identical organisms that are all replicas of a single ancestor.
Methods of creating clones of desired properties, usually called *genetic engineering* and *recombinant DNA technology*, deserve much of the credit for the dramatic rise of biotechnology since the mid-70'. The main idea of molecular cloning is to insert a DNA segment of interest into an autonomously replicating DNA molecule, called a*cloning vector*, so that the DNA segment is replicated with the vector. Such vectors could be, for instance, plasmids (circular DNAs occuring in some bacteria). Reproduction of DNA segments in appropriate hosts, results in the production of large amount of the inserted DNA segment.
A DNA to be cloned is usually a fragment of a genome of interest, obtained by application of restriction enzymes. Most restriction enzymes cleave duplex DNA at specific palindromic sites, generally two fragments that have single strand ends that are complimentary with each other (known as 'sticky ends'). Therefore, a restriction fragment can be inserted into a cut made in a cloning vector by the same restriction enzyme, because the segment ends stick (chemically bond) to the loose ends of the vector. Such a recombinant DNA molecule is inserted into a fast reproducing host cell, and is duplicated in the process of the host's reproduction. The cells containing the recombinant DNA are then isolated from non-infected cells using an antibiotic substance which the original vector is resistant to . The cloning technique provides both high quantities of DNA fragments, as well as a mean to preserve them for long periods of time (by keeping the host cells alive).

1.3 The Human Genome Project

The human genome project is a worldwide endeavor aimed to sequence the entire 3×10^9bp-long human DNA. The project, initiated by the United States in the late 80' was launched in 1990 and was intended to be completed in 15 years with a 3 billion dollar budget. Due to the massive support the project got all over the world, it's completion is expected considerably sooner then expected. Since so far the effort was focused mainly at constructing biotechnical tools and methods, only 5% of the human genome has been sequenced so far. The largest fully sequenced genome to date is the 10^8-long genome of the nematode worm *caenorhabditis elegans.*

The human genome project was entitled 'biology's space conquest', although its effects

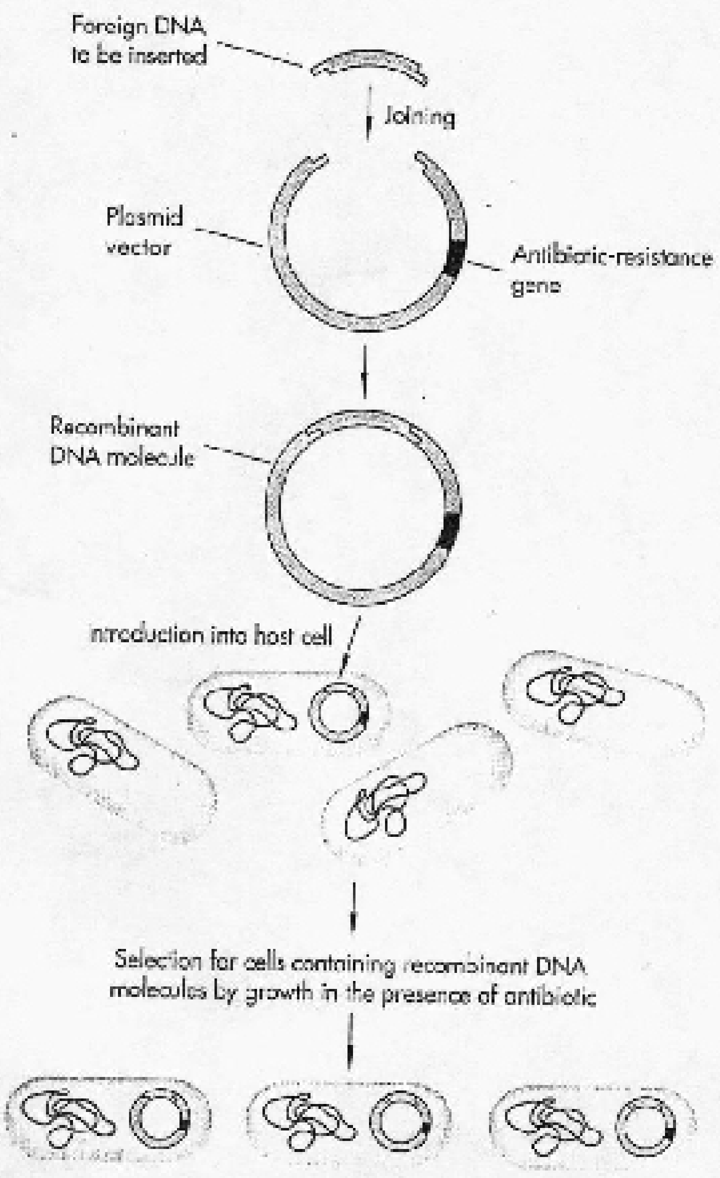
Foreign DNA to be inserted
Joining
Plasmid vector
Antibiotic-resistance gene
Recombinant DNA molecule
introduction into host cell
Selection for cells containing recombinant DNA molecules by growth in the presence of antibiotic

FIGURE 10
The cloning of DNA in a plasmid.

Figure 1.10: cloning

and benefits to biological and medical science will probably be much more significant and far reaching.

Bibliography

[1] M. Chalfie. Genome sequencing: The work revealed. *Nature*, 396:1–4, 1998.

[2] L. Stryer. *Biochemistry*. W.H. Freeman, New York, 4th edition, 1995.

[3] J.D. Watson, M. Gilman, J. Witkowski, and M. Zoller. *Recombinant DNA*. W.H. Freeman, New York, 2nd edition, 1992.