Algorithms for Molecular Biology **Fall Semester, 2001**

Lecture 1: October 25, 2001

Lecturer: Ron Shamir *Scribe: Gadi Kimmel and Ariel Farkash*[1]

1.1 Biological Background

1.1.1 Historical Introduction

Genetics as a set of principles and analytical procedures did not begin until 1866, when an Augustinian monk named Gregor Mendel performed a set of experiments that pointed to the existence of biological elements called *genes* - the basic units responsible for possession and passing on of a single characteristic. Until 1944, it was generally assumed that chromosomal proteins carry genetic information, and that DNA plays a secondary role. This view was shattered by Avery and McCarty who demonstrated that the molecule deoxy-ribonucleic acid (DNA) is the major carrier of genetic material in living organisms, i.e. is responsible for inheritance. In 1953 James Watson and Francis Crick deduced the three dimensional structure of DNA and immediately inferred its method of replication. In February 2001, due to a joint venture of the Human Genome Project and a commercial company Celera, the first draft of the human genome was published.

1.1.2 DNA

Composition

The basic elements of DNA had been isolated and determined by partly breaking up purified DNA. These studies demonstrated that DNA is composed of four basic molecules called *nucleotides*, which are identical except that each contains a different nitrogen base. Each nucleotide contains phosphate, sugar (of the deoxy-ribose type) and one of the four bases: *Adenine*, *Guanine*, *Cytosine*, and *Thymine* (usually denoted A,G,C,T) (See Figure 1.1).

Structure

The structure of DNA is described as a *double helix*, which looks rather like two interlocked bedsprings. Each helix is a chain of nucleotides held together by phospho-diester bonds. The two helices are held together by hydrogen bonds. Each base pairs consists of one

[1]Based in part on a scribe by Eran Goldberg and Rotem Sorek, October, 2000, and on [5, 7, 2, 9, 4, 10, 6, 1, 3]

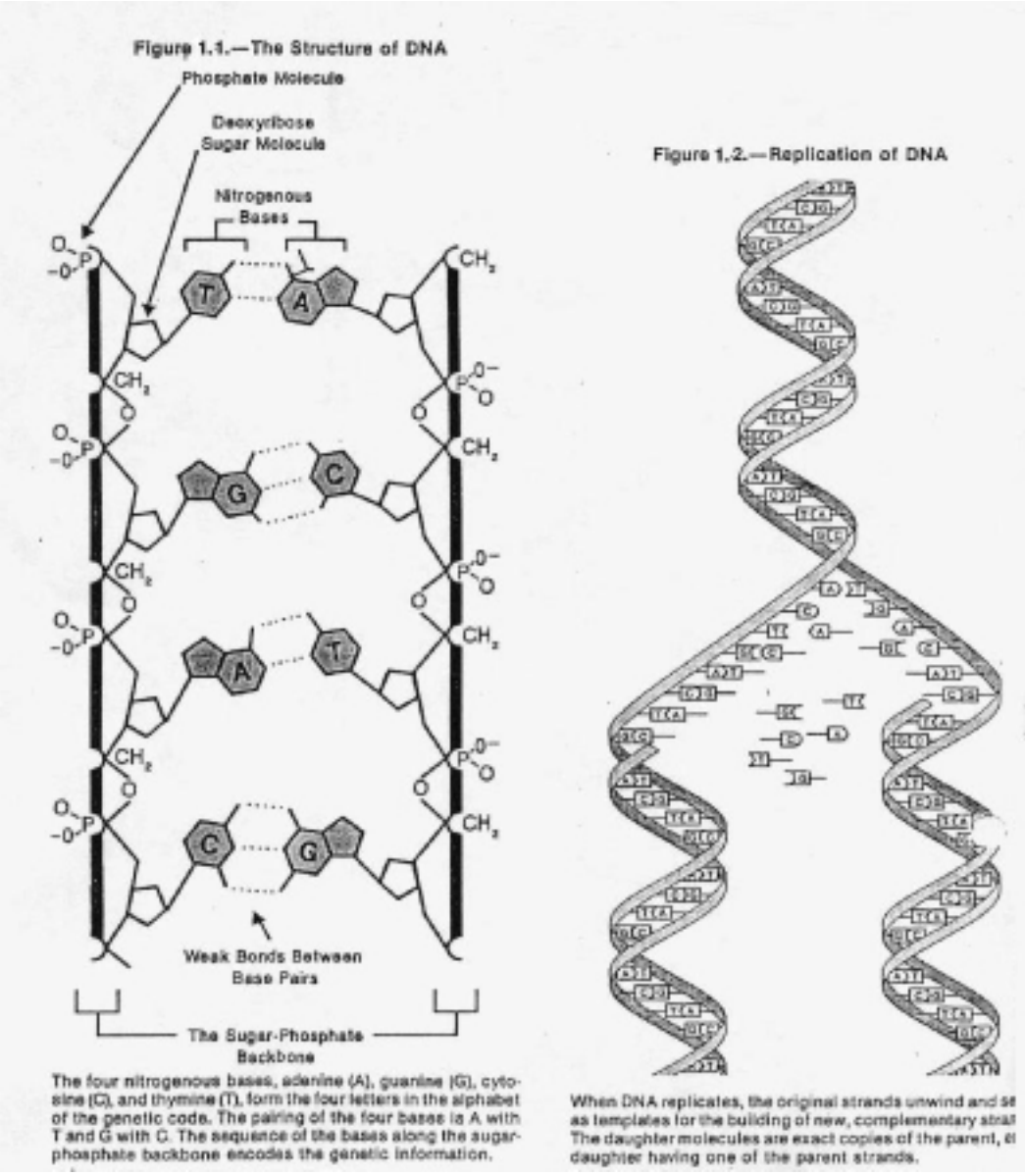


Figure 1.1: Source: [8]. On the left, DNA composition. On the right, DNA double helix structure.

purine base (A or G) and one *pyrimidine* base (C or T), paired according the following rule: $G \equiv C, A = T$ (each '-' symbolizes a hydrogen bond). The DNA molecule is directional, due to the asymmetrical structure of the sugars which constitute the skeleton of the molecule. Each sugar is connected to the strand *upstream* (i.e. preceding it in the chain) in its fifth carbon and to the strand *downstream* (i.e. following it in the chain) in its third carbon. Therefore, in biological jargon, the DNA strand goes from $5'$ (read *five prime*) to $3'$ (read *three prime*). The directions of the two complementary DNA strands are reversed to one another.

Replication

The double helix could be imagined as a zipper that unzips, starting at one end. We can see that if this zipper analogy is valid, the unwinding of the two strands will expose single bases on each strand. Because the pairing requirements imposed by the DNA structure are strict, each exposed base will pair only with its complementary base. Due to this base complementarity, each of the two single strands will act as a template and will begin to re-form a double helix identical to the one from which it was unzipped. The newly added nucleotides are assumed to come from a pool of free nucleotides that must be present in surrounding micro-environment within the cell. 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*, an oligonucleotide (a short 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).

1.1.3 Genes and Chromosomes

Each DNA molecule is packaged in a separate *chromosome*, and the total genetic information stored in the chromosomes of an organism is said to constitute its *genome*. With few exceptions, every cell of a Eukaryotic multi-cellular organism contains a complete set of the genome, while the difference in functionality of cells from different tissues is due the variable expression of the corresponding genes. The human genome contains about 3×10^9 base pairs (abbreviated *bp*), organized as 46 chromosomes - 22 different autosomal chromosome pairs, and two sex chromosomes: either XX or XY. The 24 different chromosomes range from 50×10^6 to 250×10^6 bp. The amount of DNA varies between different organisms. The organism *Amoeba dubia* (a single cell organism), for example, has more than 200 times DNA as human. The living organisms divide into two major groups: *Prokaryotes*, which are single-celled organisms with no cell nucleus, and *Eukaryotes*, which are higher level organisms, and their cells have nuclei. With contemporary knowledge of the biochemical basis of heredity, Mendel's abstract concept of a gene can be redefined as a physical entity. A gene is a region of DNA that controls a discrete hereditary characteristic, usually corresponding

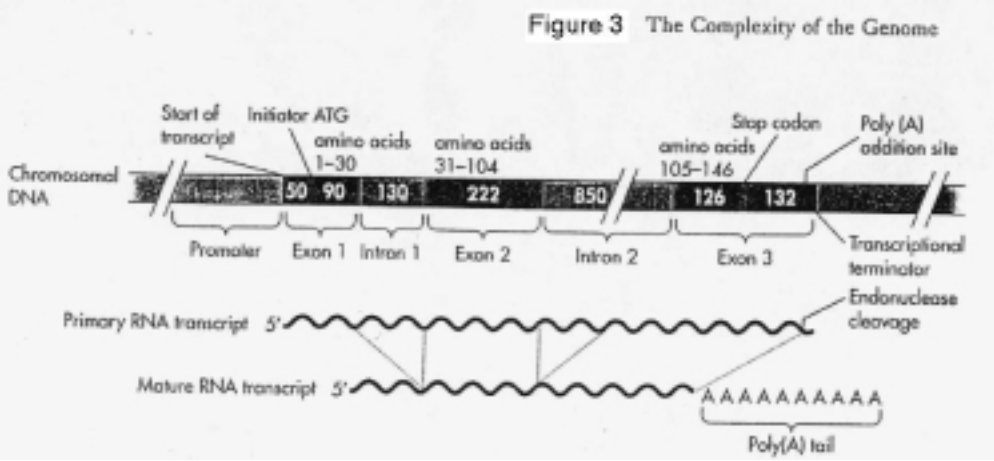


Figure 1.2: Source: [8]. Exons and introns.

to a single mRNA carrying the information for constructing a protein. In 1977 molecular biologists discovered that most Eukariotic genes have their coding sequences, called *exons*, interrupted by non-coding sequences called *introns*, (See Figure 1.2). In humans genes constitute approximately 2-3% of the DNA, leaving 97-98% of non-genic *junk DNA*. The role of the latter is as yet unknown, however experiments involving removal of these parts proved to be lethal. Several theories have been suggested, such as physically fixing the DNA in its compressed position, preserving old genetic data, etc.

1.1.4 The Central Dogma

The expression of the genetic information stored in DNA involves the translation of a linear sequence of nucleotides into a co-linear sequence of amino acids in proteins.
The flow is: DNA → mRNA → Protein (See Figure 1.3).

Transcription

A segment of DNA is first copied into a complementary strand of RNA. This process called *transcription* is catalyzed by the enzyme *RNA polymerase*. Near most of the genes there is a special pattern in the DNA called *promotor*, located upstream of the transcription start site, which informs the RNA polymerase where to begin the transcription. This is achieved with the assistance of transcriptional factors that recognize the promotor sequence and bind to it. Although *ribonucleic acid* (RNA) is a long chain of nucleic acids (as is DNA), it has very different properties. First, RNA is usually single stranded (denoted ssRNA). Second, RNA has a ribose sugar, rather than deoxy-ribose. Third, RNA has the pyrimidine based *Uracil* (abbreviated U) instead of Thymine. Fourth, unlike DNA, which is located primarily

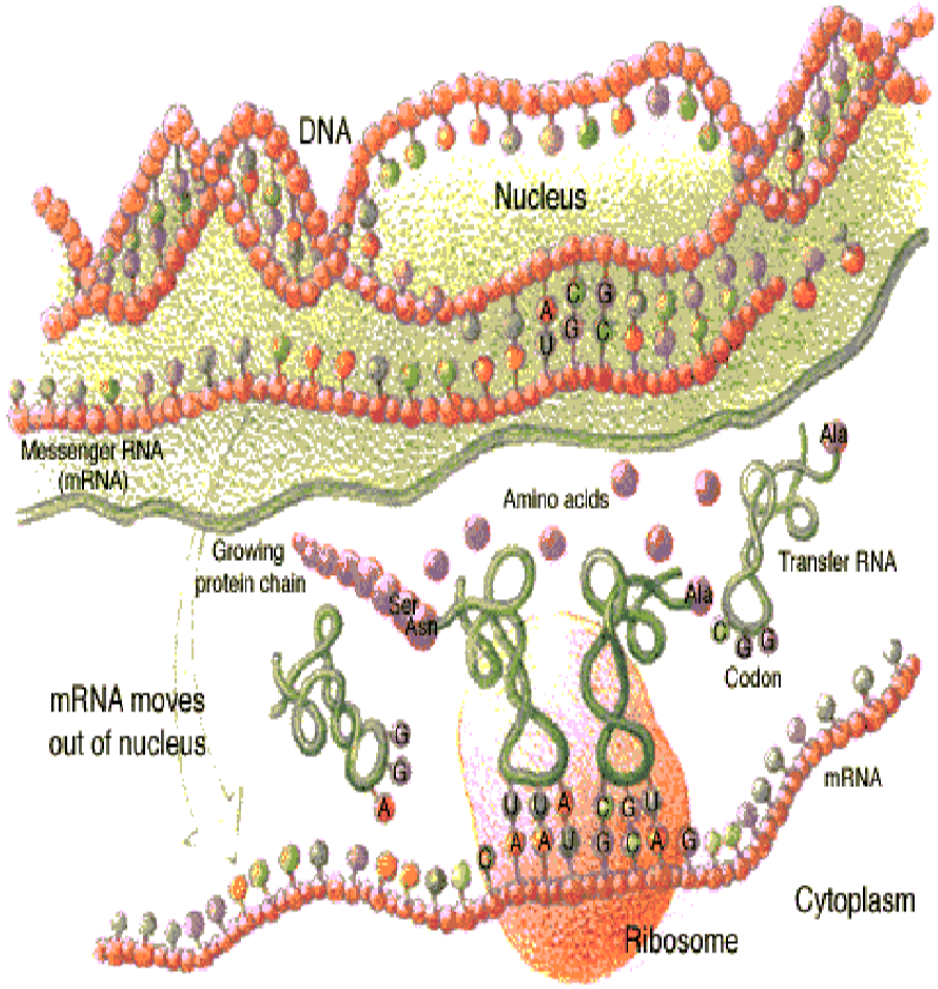


Figure 1.3: Source: [14]. From gene to protein.

in the nucleus, RNA can also be found in the cellular liquid outside the nucleus, which is called the *cytoplasm*.

In Eukariotic organisms, to produce a protein the entire length of the gene, including both its introns and its exons, is first *transcribed* into a very large RNA molecule - the *primary transcript*. At the end of the gene the transcription stops, and a few dozens of Adenine (A) nucleotides are added to the end of the RNA molecule for protection (*poly-A tail*). *5' CAP* lays an important part in the initializing of protein synthesis by the protecting the growing RNA transcript from degradation. Before this RNA molecule leaves the nucleus, a complex of RNA processing enzymes removes all the intron sequence, in a process called *splicing*, thereby producing a much shorter RNA molecule (See Figure 1.4). Typical eukaryotic exons are of average length of 200bp, while the average length of introns is around 10000bp (these lengths can vary greatly between different introns and exons). In many cases, the pattern of the splicing can vary depending on the tissue in which the transcription occurs. For example, an intron that is cut from mRNAs of a certain gene transcribed in the liver, may not be cut from the same mRNA when transcribed in the brain. This variation is called *alternative splicing*, and it contributes to the overall protein diversity in the organism. After this RNA processing step has been completed, the RNA molecule moves to the cytoplasm as a *messanger RNA* molecule (*mRNA*), in order to undergo translation.

The Genetic Code

The rules by which the nucleotide sequence of a gene is translated into the amino acid sequence of the corresponding protein, the so called *genetic code*, were deciphered in the early 1960s. The sequence of nucleotides in the mRNA molecule, that acts as an intermediate was found to be read in serial order in groups of three. Each triplet of nucleotides, called a *codon*, specifies one *amino acid* (the basic unit of a protein, analogous to nucleotides in DNA). Since RNA is a linear polymer of four different nucleotides, there are $4^3 = 64$ possible codon triplets (See Figure 1.5). However, only 20 different amino acids are commonly found in proteins, so that most amino acids are specified by several codons. In addition, 3 codons (of the 64) specify the end of translation, and are called *stop codons*. The codon specifying beginning of translation is AUG, and is also the codon for the amino acid Methionine. The code has been highly conserved during evolution: with a few minor exceptions, it is the same in organisms as diverse as bacteria, plants, and humans.

Translation

In principle, each RNA sequence can be translated in any one of three *reading frames* in each direction, making a total of 6 possible *open reading frames* - *ORFs*, depending on where the process begins. In almost every case, only one of these reading frames will produce a functional protein. However, there are rare cases, especially in viruses, where genes are

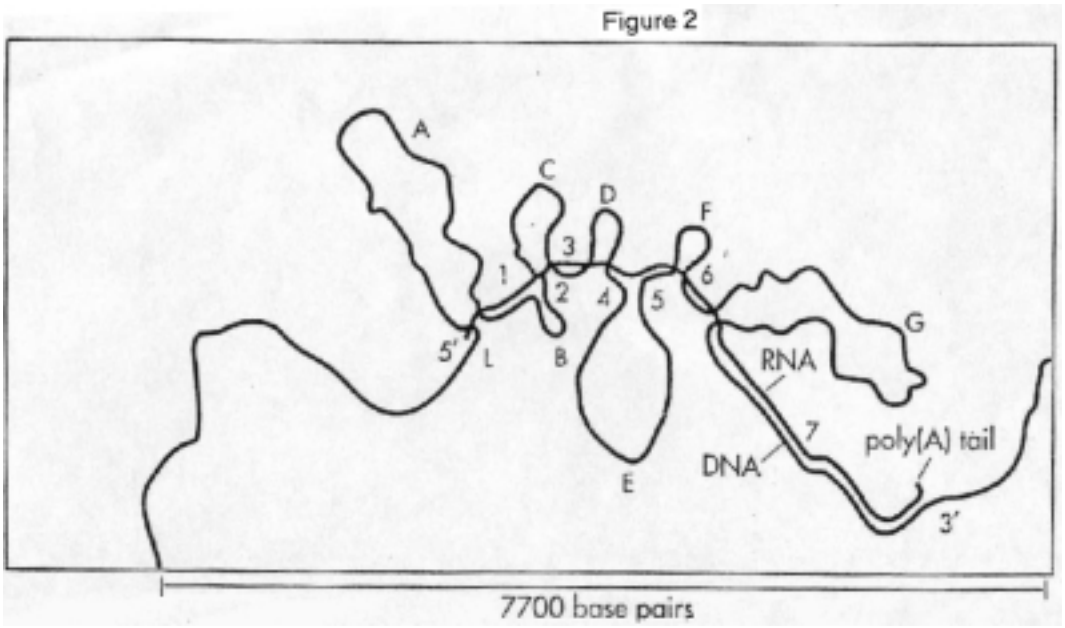
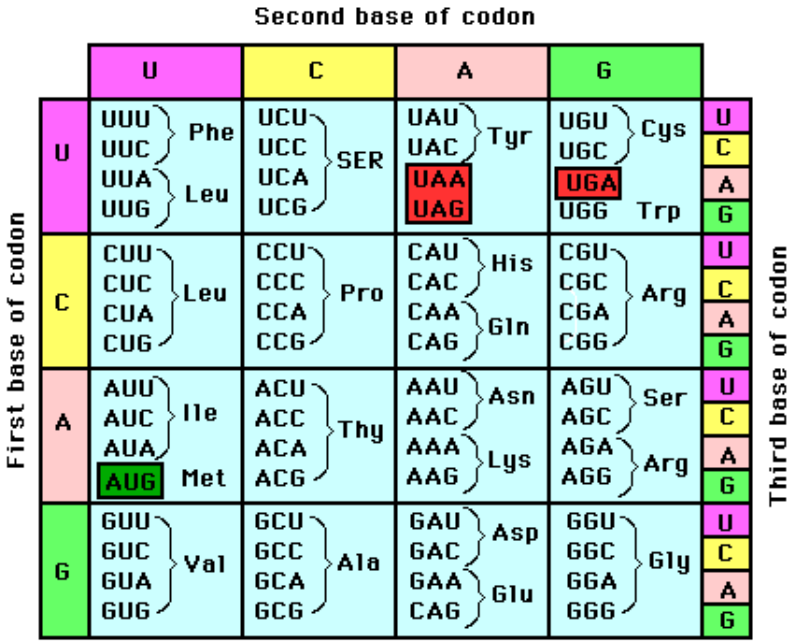


Figure 1.4: Source: [8]. Exons and introns in DNA. In this experiment an mRNA is attached to a single-stranded DNA from which it was transcribed. The regions of the DNA that are attached to the mRNA (1-7) are the exons (present in both the DNA and the mRNA). The regions of the DNA that are not attached to the mRNA (A-G) are the introns (present only in the DNA)



The genetic code, written by convention in the form in which the Codons appear in mRNA. The three terminator codons, UAA, UAG, and UGA, are boxed in red; the AUG initiator codon is shown in green.

Figure 1.5: Source: [12]. The genetic code.

transcribed from overlapping complementary regions of the DNA.

The *translation* of mRNA into protein depends on adaptor molecules that recognize both an amino acid and a triplet of nucleotides. These adaptors consist of a set of small RNA molecules known as *transfer RNA - tRNA*, each about 80 nucleotides in length. The tRNA molecule enforces the universal genetic code logic in the following fashion: On one part the tRNA holds an *anticodon*, a sequence of three RNA bases; on the other side, the tRNA holds the appropriate amino acid. In eukaryotes, the mRNA is formed of *coding* regions flanked by *non-coding* regions. Coding regions (exons or parts of exons) used for the protein creation, while the non coding regions - *3' untranslated region* and *5' UTR* - are mostly regulatory and are not translated. Note, that along the DNA, the coding region may not be contiguous, as it might span several exons. In Prokaryotes, a gene has only one coding region, flanked by the 3' UTR and the 5' UTR.

Due to the mechanic complexity of ordering the tRNA molecules on the mRNA, a mediator is required. The *ribosome* is a complex of more than 50 different proteins associated with several structural rRNA molecules. Each ribosome is a large protein synthesizing machine, on which tRNA molecules position themselves for reading the genetic message encoded in an mRNA molecule. Ribosomes operate with remarkable efficiency: in one second a single bacterial ribosome adds about 20 amino acids to a growing poly-peptide chain. Many ribosomes can simultaneously translate a single mRNA molecule.

1.1.5 Proteins

A protein is linear polymer of amino acids linked together by peptide bonds. The average protein size is around 200 amino acids long, while large proteins can reach over a thousand amino acids. To a large extent, cells are made of proteins, which constitute more than half of their dry weight. Proteins determine the shape and structure of the cell, and also serve as the main instruments of molecular recognition and catalysis. Proteins have a complex structure, which can be thought of as having four hierarchial structural 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* which are single stranded helices of amino acids, and *β-sheets* which are planar patches woven from chain segments that are almost linearly arranged. The *tertiary structure* is formed by packing such structures into one or several 3D *domains*. The final, complete, protein may contain several protein domains arranged in a *quaternary structure* (See Figure 1.6). The whole complex structure (primary to quaternary) is determined by the primary sequence of amino acids and their physico-chemical interaction in the medium. Therefore, its *folding* structure is defined by the genetic material itself, as the three dimensional structure with the minimal free energy. The structure of a protein determines its functionality. Although the amino acid sequence directly determines the proteins structure, 30% amino acid sequence identity will, in most cases, lead to high similarity in structure.

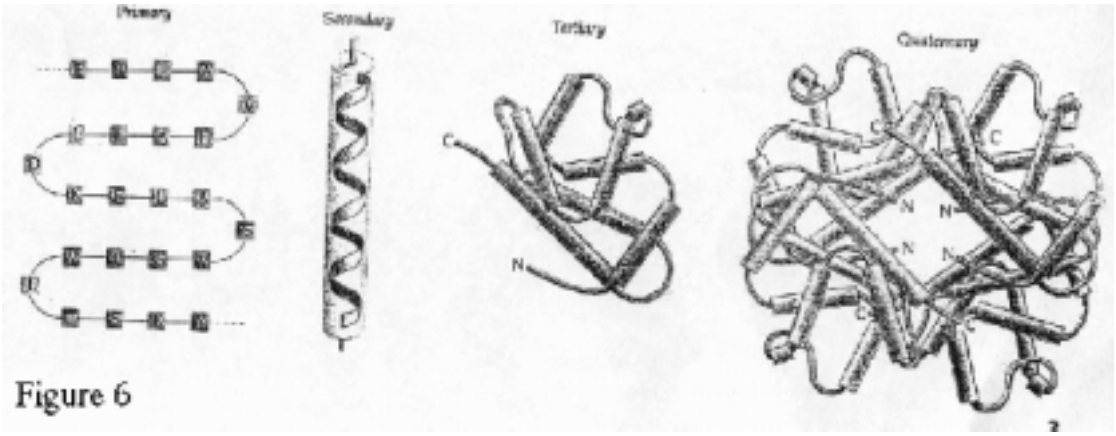


Figure 1.6: Source: [8]. Protein structure.

1.1.6 Mutations

A *mutation* is defined as a heritable change in the nucleotide sequence in the DNA, caused by a faulty replication process. These errors in replication occur often due to exposure to ultra violet radiation or other environmental conditions. There are two different levels at which a mutation may take place. In *gene mutation* an *allele* of a gene changes becoming a different allele. Because such a change occurs within a single gene and maps to one chromosomal *locus*, a gene mutation is sometimes called a point mutation. At the other level of hereditary change, *chromosomal mutation* or *rearrangements* - segments interchange, either on the same chromosome or on different ones (*translocations*). In addition, a chromosome may undergo a more global change, such as *reversal*, *deletion*, *duplication*, etc. For example, *Down's Syndrome* is caused by such a chromosomal mutation.
There are several kinds of point mutations:

- substitution - a change of one nucleotide in the DNA sequence.
- insertion - an addition of one or more nucleotides to the DNA sequence.
- deletion - a removal of one or more nucleotides from the DNA sequence.

The point mutations can also be be divided according to their influence on the resulting protein:

- missense - a mutation that alters the codon so that it encodes for a different amino acid.
- silent - a mutation that does not alter the codon so that it encodes for the same amino acid.

- nonsense - a mutation that alters the codon, as to produce a stop codon.

It is crucial to mention that even though a mutation may change the amino acid sequence of a protein, it does not necessarily affect the protein's functionality. This phenomenon can be explained by the fact that chemical similarity between different amino acids may result in little or no impact on the final folding structure of the protein, therefore preserving its functionality. Furthermore, there are regions in the protein that have very little influence on the structural functionality of the molecule.

Mutations are important for several reasons. They are responsible for inherited disorders and other diseases. For example, Sickle-cell anemia, is a disease caused by a substitution mutation of thymine for adenine, resulting in the codon for Valine instead of Glutamic Acid in the sixth amino acid of the hemoglobin protein. This simple missense mutation causes a lethal disorder in which oxygen is poorly supplied by the hemoglobin in red blood cells to the tissues, causing terminal tissue damage. On a brighter aspect, mutations are the source of phenotypic variation on which natural selection acts, creating new species and adapting existing ones to changing environmental conditions. The gene variation between organisms allows us to investigate species evolution using molecular evidence as our artifacts. Furthermore, it advances medical research in the on-going search for new and better drugs.

1.2 Biological Background - Computational Issues

1.2.1 The Gene Finding Problem

Problem 1.1 Given a DNA sequence, predict the location of genes (open reading frames), exons and introns.

A simple solution would be to seek stop codons in regions along the sequence. Clearly, if several stop codons appear close to each other in a region it would have been terminated, thereby we can safely assume that it is not a coding region. Detecting a relatively long sequence deprived of stop codons could indicate a coding region. The problem complicates in eukaryotic DNA due to the existence of interleaving exons and introns. Further complications arise from the fact that certain DNA sequences can be interpreted in 6 different ways due to their corresponding open reading frames, as mentioned earlier. In most cases, in eukaryotic organisms, a DNA region will encode only one gene, which is not necessarily true in prokaryotes.

1.2.2 The Sequence Alignment Problem

Problem 1.2 Given two DNA or protein sequences, find the best match between them.

In order to do so we define a set of possible operations and their corresponding penalties. For example, a biological phenomenon such as insertion would be mathematically translated into an open gap action which would carry a penalty. In this fashion, we can characterize other features such as deletions, mismatch, frame-shifts etc., each carrying its own specific penalty according to their biological frequencies and gravity. The resulting best match is the one with the minimum sum of such penalties. In the more general Multiple Sequence Alignment Problem, there are more than two sequences.

1.2.3 The Genome Rearrangement Problem

Problem 1.3 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 rare as compared to point mutations. For example, substitutions occur in some organisms about 10 times in each generation, while a non fatal rearrangement event occurs once every 5 to 10 million years. The lower rate of rearrangements allows us to detect a directional evolutionary process, since the chance of reversal is minute. Therefore, by discovering which rearrangement events have occurred, and the order of their occurrence, we might be able to build an evolutionary hypothesis.

1.2.4 The Protein Folding Problem

Since the functionality of the protein is determined by its 3D structure, it is very important to predict the structure of a protein, thus gaining better understanding of its role in the cell.

Problem 1.4 Given a sequence of amino acids, predict the 3D structure of the protein.

The problem of predicting a protein's structure de-novo, i.e. based on its amino acid sequence and their chemical properties, is yet to be solved. Nevertheless, several approaches have been developed to approximate the structure of a protein:

- *Homology modeling* - uses a protein database to search for similar sequences of proteins. If a protein with around 30% sequence identity is found, it is quite safe to assume that the two proteins have similar structures.

- *Threading* - classifies known structures into families with similar foldings. Given a sequence of amino acids, we can select the family to which the given sequence is most likely to belong to.

1.3 Biotechnological Methods

Before the 1970s the goal of isolating a single gene from a large chromosome seamed unattainable. Unlike a protein, a gene does not exist as a discrete entity in cells, but rather as a small region of a much larger DNA molecule. Although the DNA molecules in the cell can be randomly broken into small pieces by mechanical force, a fragment containing a single gene in a mammalian genome would still be only one among a hundred thousand or more DNA fragments, indistinguishable in their average size. In order to simplify this process one can use 'biological machinery' as our ally, this being the major goal in biotechnology. Using biotechnological techniques allows us to produce large quantities of substances necessary for medical procedures, as well as isolation of specific substances for diagnostic purposes.

1.3.1 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 in order to protect the cell from infection. A restriction enzyme breaks the phospho-diester bonds of both strands of a DNA, in a process called *digestion*. The cleavage point of the enzyme is characterized by a target sequence (usually a palindrome). Currently, there are over 150 known nucleotide configurations that serve as target cleavage sites of known restriction enzymes (See Figure 1.7).

1.3.2 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, each containing DNA fragments of the same size (See Figure 1.8).

1.3.3 Sequencing

Sequencing is the operation of determining the nucleotide sequence of a given molecule. DNA can be sequenced by generating fragments through the controlled interruption of enzymatic replication, a method developed by Fredrick Sanger and co-workers. This is now the method of choice because of its simplicity. *DNA polymerase* is used to copy a particular sequence of a single stranded DNA. The synthesis is primed by a complementary fragment, which may be obtained from a restriction enzyme digest, or synthesized chemically. In addition to the four nucleotides (radioactively labelled), the incubation mixture contains a 2',3' dideoxy analog of one of them. The incorporation of this analog blocks further growth of the

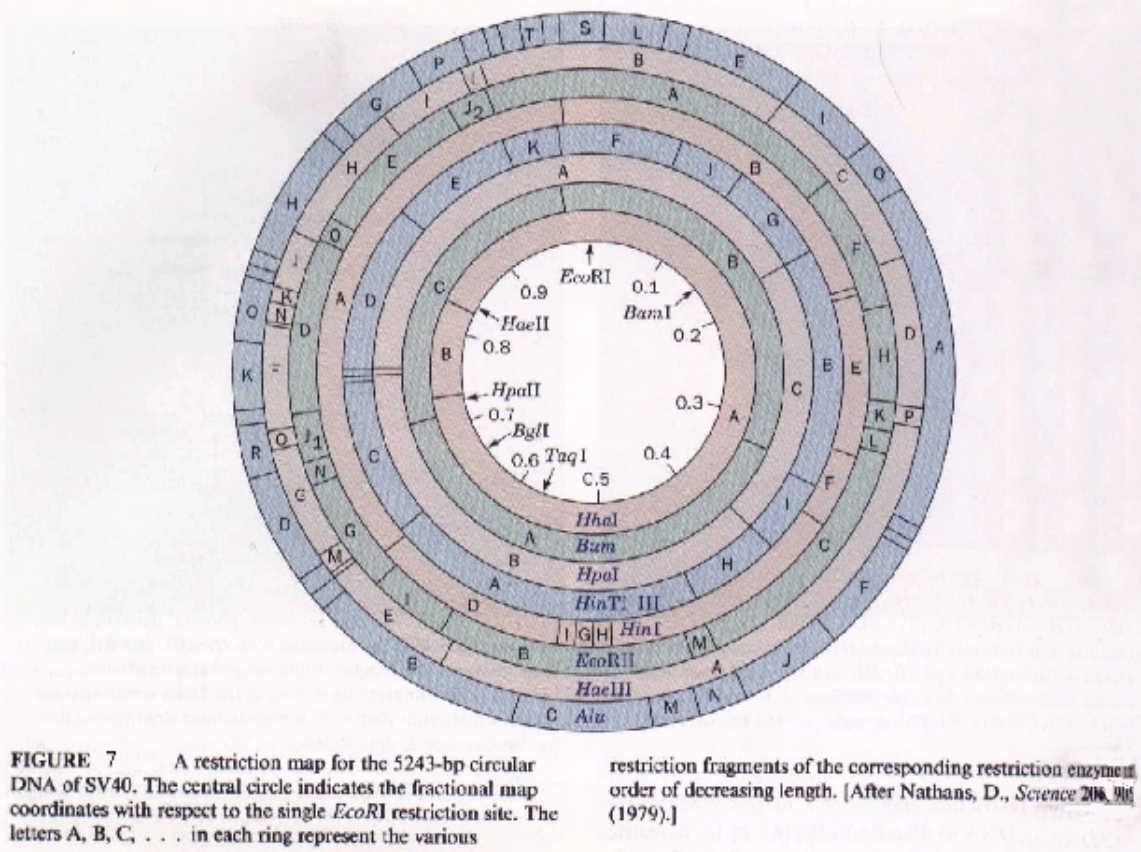


FIGURE 7 A restriction map for the 5243-bp circular DNA of SV40. The central circle indicates the fractional map coordinates with respect to the single *EcoR*I restriction site. The letters A, B, C, . . . in each ring represent the various restriction fragments of the corresponding restriction enzymes in order of decreasing length. [After Nathans, D., *Science* 206, 903 (1979).]

Figure 1.7: Source: [8]. Restriction map: The letters A,B,C,... in each ring represent the various restriction fragments of the corresponding restriction enzymes in order of decreasing length.

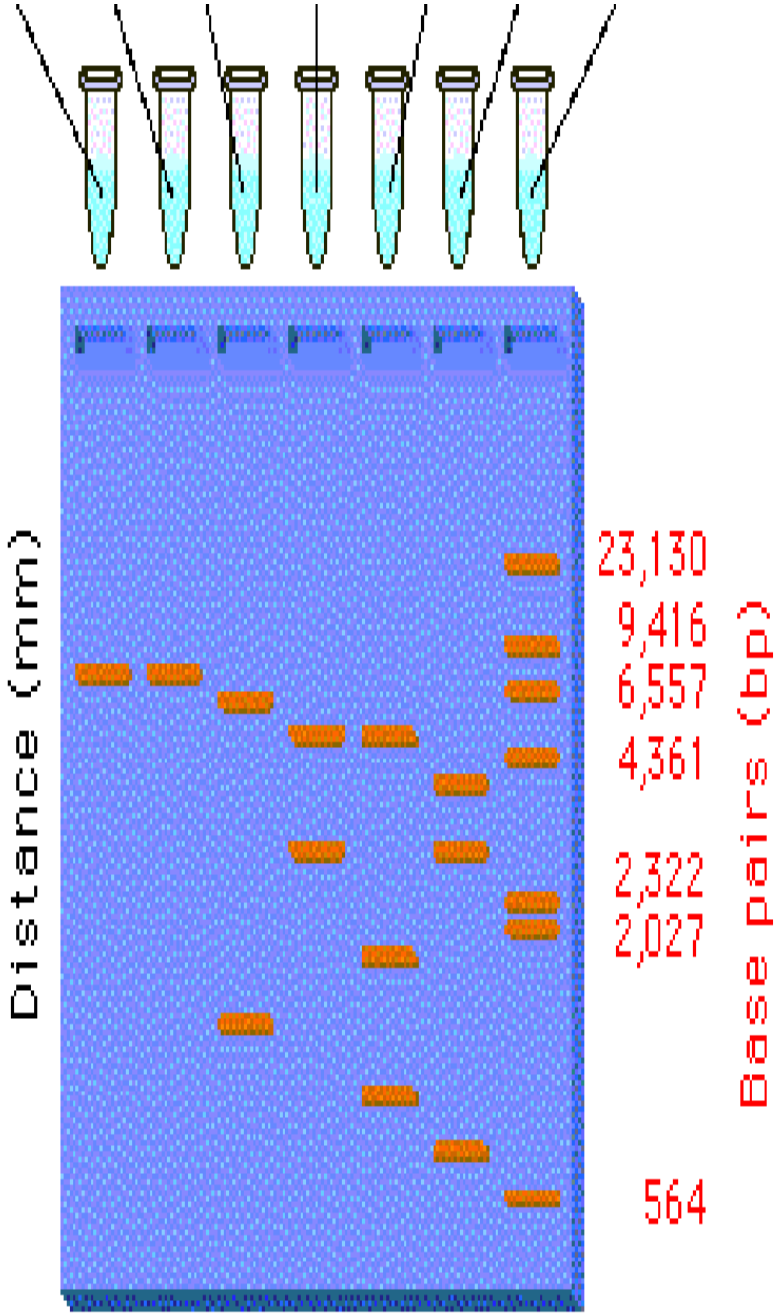


Figure 1.8: Source: [11]. Gel electrophoresis.

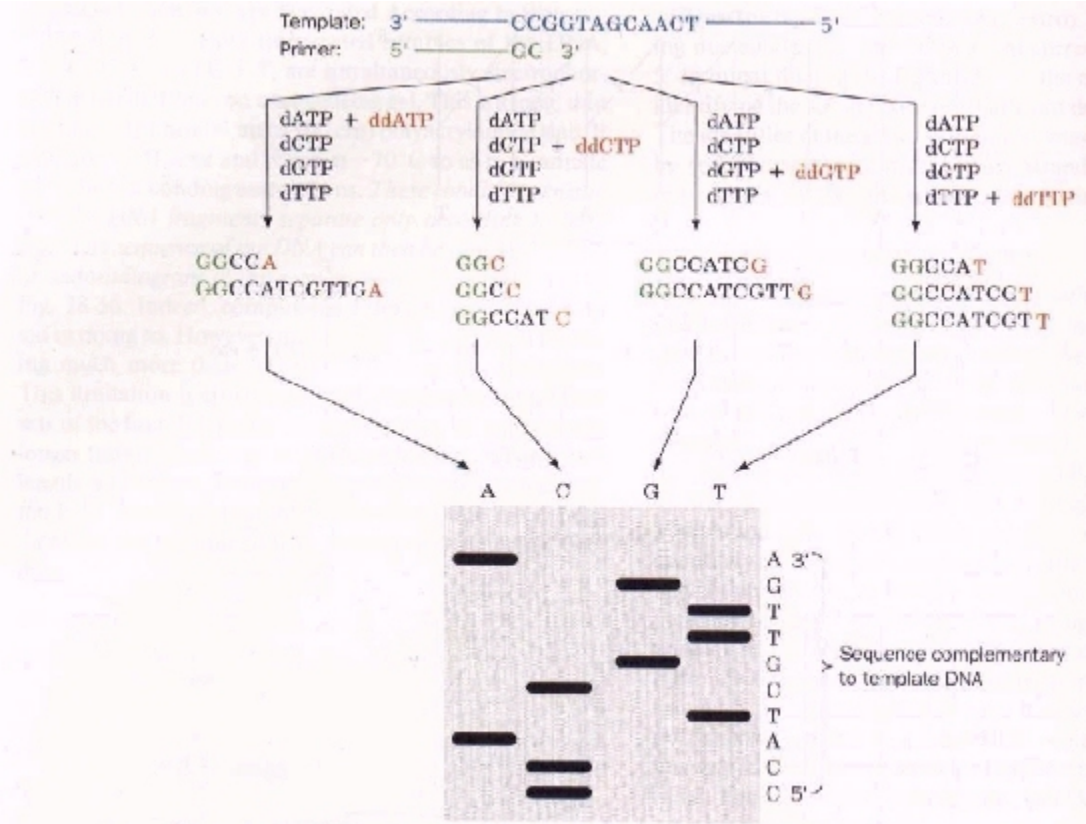


FIGURE 9 . A flow diagram of the chain-terminator (dideoxy) method of DNA sequencing. The symbol ddATP represents dideoxyadenosine triphosphate, *etc.*

Figure 1.9: Source: [8]. DNA sequencing: electrophoresis of four sets of chain terminated fragments. Each di-deoxy analog in a different lane.

new chain because it lacks the 3' terminus needed to form the next phospho-diester bond. Hence, fragments of various lengths are produced in which the di-deoxy analog is at the 3' end. Four such sets of chain terminated fragments (one for each di-deoxy analog) are then electrophoresed, and the base sequence of the new DNA is read from the autoradiogram of the four lanes (See Figure 1.9). Using this method, sequences of 500-800 nucleotides can be determined within reasonable accuracy. The advanced sequencing machines nowadays can sequence simultaneously 96 different sequences of 500-800 nucleotides in a few hours.

1.3.4 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 devel-

opment of molecular *cloning* techniques. A 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-70s'. 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. An example of vectors are plasmids (circular DNAs occurring in some bacteria). Reproduction of DNA segments in appropriate hosts results in the production of large amounts of the inserted DNA segment.

The cloned DNA segment 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, and every two fragments have single strand ends that are complimentary to 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, as 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 system (See Figure 1.10). The cells containing the recombinant DNA are then isolated from non-infected cells using an antibiotic substance to which the original vector is resistant. 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).

1.3.5 Polymerase Chain Reaction - PCR

The availability of purified DNA polymerases and chemically synthesized DNA oligonucleotides, has made it possible to clone specific DNA sequences rapidly without the need for a living cell. The technique called *polymerase chain reaction (PCR)*, allows the DNA from a selected region of a genome to be amplified a billion fold, provided that at least part of its nucleotide sequence is already known. First, the known part of the sequence is used to design two synthetic DNA oligonucleotides, one complementary to each strand of the DNA double-helix and lying on opposite sides of the region to be amplified. These oligonucleotides serve as primers for *in-vitro DNA synthesis*, which is catalyzed by DNA polymerase, and they determine the ends of the final DNA fragment that is obtained.

Each cycle of the reaction requires a brief heat treatment to separate the two strands of the genomic DNA. The success of the technique depends on the use of a special DNA polymerase isolated from a thermophilic bacterium that is stable at much higher temperatures than normal, so that it is not denatured by the repeated heat treatments. A subsequent cooling of the DNA in the presence of large excess of two primer DNA oligonucleotides allows these oligonucleotides to hybridize to complementary sequences in the genomic DNA. The annealed mixture is than incubated with DNA polymerase and an abundance of the four nucleotides (A,C,T,G), so that the regions of DNA downstream from each of the two

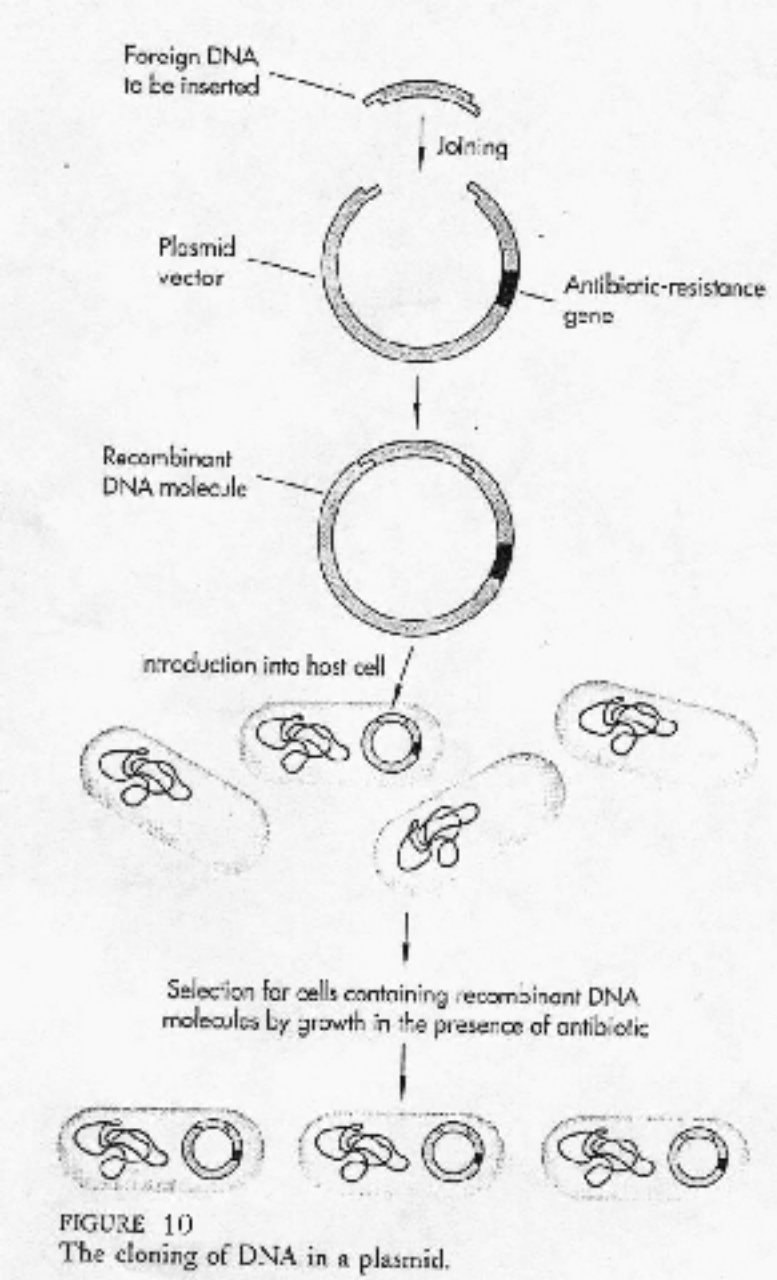


Figure 1.10: Source: [8]. Cloning procedure.

primers are selectively synthesized. When the procedure is repeated, the newly synthesized fragments serve as templates themselves, and within a few cycles the predominant product is a species of DNA fragment whose length corresponds to the distance between the original primers. In practice 20-30 cycles of reaction are required for effective DNA amplification. Each cycle doubles the amount of DNA synthesized in the previous cycle. A single cycle requires only about 5 minutes, and an automated procedure permits "cell free molecular cloning" of a DNA fragment in a few hours, compared with the several days required for some of the cloning procedures. Furthermore, the PCR procedure is usually more reliable than any other cloning procedures.

1.4 Biotechnological Methods - Computational Issues

1.4.1 Restriction Enzyme Digestion Problems

The amount of exposure of the DNA to restriction enzymes determines the portion of possible sites that were actually cleaved. Therefore, by applying different exposure times to the same DNA sequence, we can measure all possible lengths of DNA fragments that one can obtain using a particular enzyme. Using this information we can attempt to deduce the locations of the cleavage sites in the original molecule.

Problem 1.5 (The Double Digest Problem) Let $A = \{A_1, A_2, ..., A_n | A_1 < A_2 < ... < A_n\}$, $B = \{B_1, ..., B_n | B_1 < ... < B_m\}$, $C = A \bigcup B$ s.t. $C_1 < C_2 < ... < C_{n+m}$. Given the three sets of distances $\{|A_i - A_{i-1}|\}_{2 \leq i \leq n}$, $\{|B_i - B_{i-1}|\}_{2 \leq i \leq m}$ and $\{|C_i - C_{i-1}|\}_{2 \leq i \leq n+m}$, reconstruct the original series A_1, ..., A_n, B_1, ..., B_m.

This is an NP-hard problem, but there are some heuristics to solve it.

Problem 1.6 (The Partial Digest Problem) Let $X = \{X_1, X_2, ..., X_n | X_1 < X_2 < ... < X_n\}$. Given a set of distances $\{|X_i - X_j|\}_{1 \leq i < j \leq n}$, reconstruct the original series X_1, ..., X_n.

The complexity of this problem is unknown, although a pseudo-polynomial algorithm does exist. This problem is also known as the highway reconstruction problem.

1.4.2 The Sequence Assembly problem

In order to sequence large fragments of DNA, one can break it up to many small fragments and sequence them as mentioned earlier. The problem that arises from this technique is the assembly of a long DNA chain from the short local sequences. This problem is known as the sequence assembly problem.

Problem 1.7 Given a set of sequences, find a minimal length string containing all members of the set as substrings.

This problem is known to be NP-Complete. However, there are greedy algorithms which perform fairly well in practice. This problem is further complicated due to the existence of repetitive sequences in the genome.

1.5 The Human Genome Project

The ultimate goal of the human genome project is to produce a single continuous sequence for each of the 24 human chromosomes and to delineate the positions of all genes. The working draft sequence described by the international human genome sequencing consortium was constructed by melding together sequence segments derived from over 20,000 large clones.

Human Genome Project Timetable Overview:

- 1985 - The project was first initiated by Charles DeLisi associate director for health and environment research at the depart of energy (DoE) in the United States.
- 1988 - National Institute of Health (*NIH*) establishes the office of human genome research.
- 1990 - human genome project launched with the intention to be completed within 15 years time and a 3 billion dollar budget.
- 1996 - In a meeting in Bermuda international partners in the genome project agreed to formalize the conditions of data access including release of sequence data into public databases. This came to be known as the "Bermuda Principles".
- 1998 - Craig Ventner forms a company with intent to sequence the human genome within three years. The company, later named *Celera*, introduced a new ambitious 'whole genome shotgun' approach.
- 1999 - The public project responds to Ventner's challenge and change their time destination for completing the first draft.
- December 1999 - The first complete human chromosome sequence (number 22) published.
- June 2000 - Leaders of the public project and Celera meet in the white house to announce completion of a working draft of the human genome sequence.

- February 2001 - The first draft of the human genome was published in Nature and Science magazines.

The human genome, the first vertebrate genome sequence to be determined, seems likely to be quite representative of what we will find in other vertebrate genomes. It is around 30 times larger than the recently sequenced worm *Caenorhabditis elegans* and fruit fly *Drosophila melanogaster* genomes (available at public domains) both around 10^8 bp, and 250 times larger then that of yeast *Sacchromyces cerevisiae*. Despite its size, it seems likely to have only two or three times as many genes as the fly or worm genomes, with the coding regions of genes accounting for only 3% of the DNA. Repeat sequences form a large proportion of the remaining DNA, around 46%. These repeats may or may not have a function but they are certainly characteristic of large vertebrate genomes. The rest of the sequence contains promoters, transcriptional regulatory sequences and other features.

As of today, more than 98.5% of the human genome is sequenced and around 47% is in a *finished* state, i.e. assembled into long pieces and reviewed (See Figures 1.11, 1.12). The total number of genes in human is estimated to be between 25,000 and 40,000.

The human genome project is but the latest increment in a remarkable scientific program whose origins date back a hundred years to the rediscovery of Mendel's laws and whose end is nowhere in sight. In a sense it provides a capstone for efforts in the past century to discover genetic information and a foundation for efforts in the coming century to understand it. The scientific work would have profound long term consequences for medicine, leading to the elucidation of the underlying molecular mechanisms of disease and thereby facilitating the design in many cases of rational diagnostics and therapeutics targeted at those mechanisms.

> "We shall not cease from exploration. And the end of all our exploring will be to arrive where we started, and know the place for the first time". —T.S. Eliot

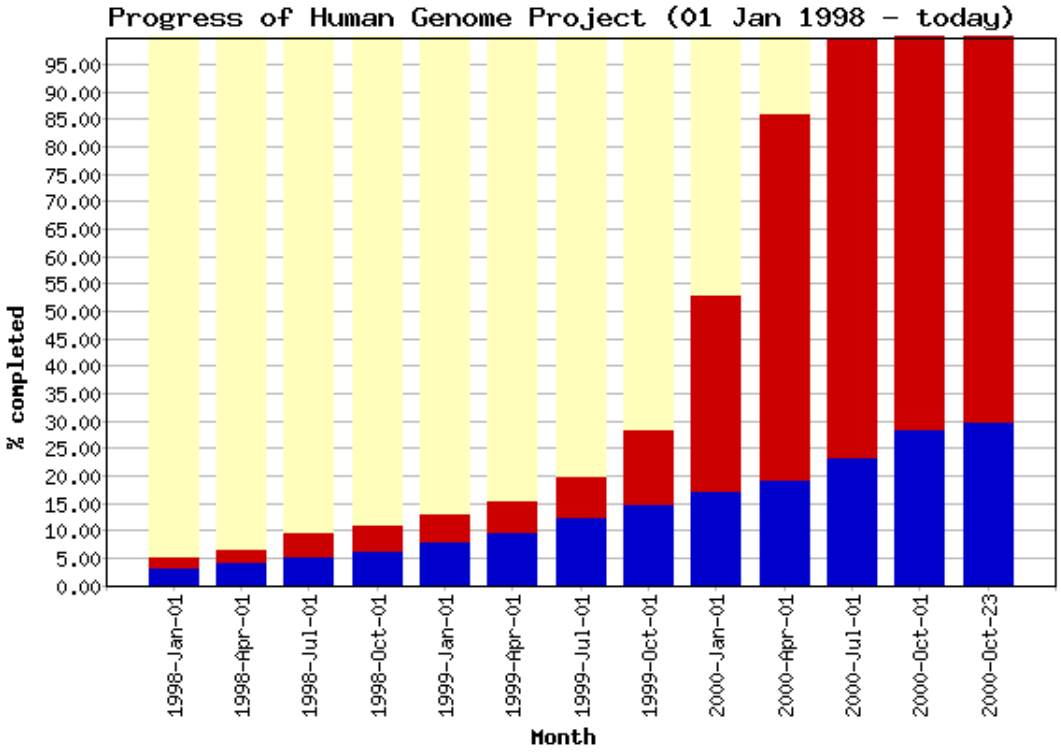


Figure 1.11: Source: [13]. The Public Human Genome Project - Progress until Oct. 2000: blue - finished, red - draft, yellow - yet to be sequenced.

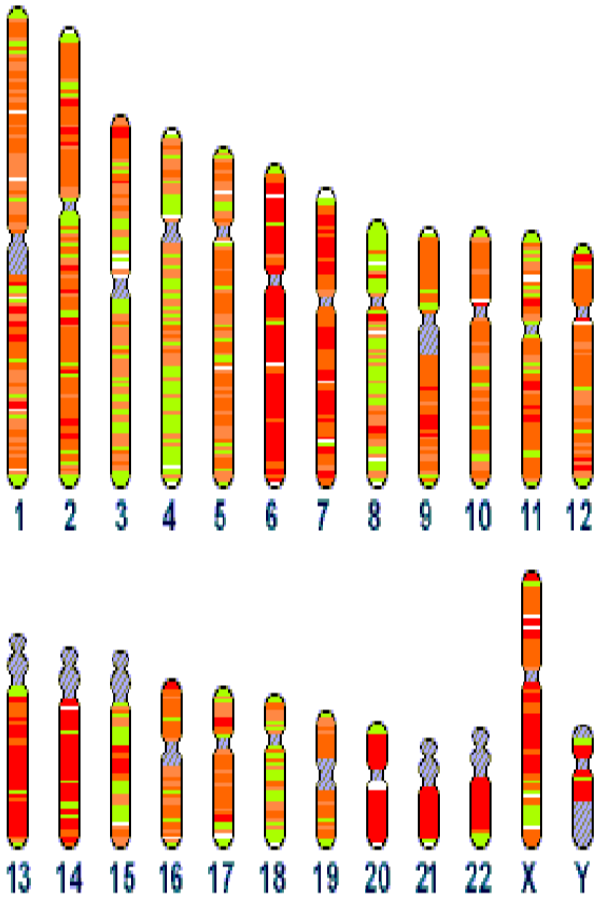


Figure 1.12: Source: [13]. The Public Human Genome Project - status as of Oct. 2001, scaled from green (draft) to red (finished).

Bibliography

[1] J. Aach, M.L. Bulyc, G.M. Church, J. Comander, A. Derti, and J. Shendure. Computational comparison of two draft sequences of the human genome. *Nature*, 409, 2001.

[2] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. Watson. *Molecular Biology of the Cell*. Garland Publishing, Inc, 1994.

[3] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409, 2001.

[4] C. Dennis, R. Gallagher, and P. Camphell. Everyone's genome. *Nature*, 409, 2001.

[5] A.J.F. Griffiths, J.H. Miller, D.T. Suzuki, R.C. Lewontin, and W.M. Gelbart. *An Intorduction to Genetic Analysis*. W. H. Freeman, New York, 1996.

[6] W-H. Li, Z. Gu, H. Wang, and A. Nekrotenko. Evolutionary analyses of the human genome. *Nature*, 409, 2001.

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

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

[9] J.D. Wilson, E. Braunwald, K.J. Isselbacher, R.G. Petersdorf, J.B. Martin, A.S. Fauci, and R.K. Root. *Priciples of Internal Medicine*. McGraw-Hill, 2000.

[10] T.G. Wolfsberg and J. McEntyre. Guide to the draft human genome. *Nature*, 409, 2001.

[11] `http://dlab.reed.edu/projects/vgm/vgm/VGMProjectFolder/VGM/`.

[12] `http://ntri.tamuk.edu/cell/ribosomes.html/`.

[13] `http://www.ncbi.nlm.nih.gov/genome/seq/`.

[14] `http://www.ornl.gov/hgmis/publicat/tko/index.htm/`.