Lecture 1: March 14, 2002

Lecturer: Ron Shamir *Scribe: Dana Torok and Adar Shtainhart*[1]

1.1 Basic Biology

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., responsible for inheritance. In 1953 James Watson and Francis Crick deduced the three dimensional double helix structure of DNA and immediately inferred its method of replication (see [2], pages 859-866). In February 2001, due to a joint venture of the Human Genome Project and a commercial company Celera (see [8]), the first draft of the human genome was published.

1.1.2 DNA (Deoxy-Ribonucleic acid)

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* (denoted A, G, C, T) (see Figures 1.1 and 1.2). The length of human DNA is about 3×10^9 base pairs (abbreviated *bp*).

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 *purine* base (A or G) and one*pyrimidine* base (C or T), paired according the following rule:

[1]Based in part on a scribe by Gadi Kimmel and Ariel Farkash, October 2001, and on a scribe by Amos Tanay and Eyal Zach, January 2002.

$G \equiv C, A = T$ (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 (see Figure 1.1).

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 reform a double helix identical to the one from which it was unzipped (see [1], pages 629-639). The newly added nucleotides are assumed to come from a pool of free nucleotides that must be present in the surrounding micro-environment within the cell. The replication reaction is catalyzed by the enzyme *DNA polymerase*. This enzyme can extend a chain, but can not start a new one. Therefore, DNA synthesis must first be initiated with a *primer*, a short nucleotide sequence (oligonucleotide). The oligonucleotide generates a segment of duplex DNA that is then turned into a new strand by the replication process (see Figure 1.3).

1.1.3 Chromosomes

The cell's DNA is stored in the nucleus. The space inside the nucleus is limited and has to contain billions of nucleotides (see Figure 1.5). Therefore, the DNA has to be highly organized. There are several levels to the DNA packaging: At the finest level, the nucleotides are organized in the form of linear strands of double helices. Zooming out, the DNA strand is wrapped around his tones, a form of DNA binding proteins. Each unit of DNA wrapped around an *octamer* of histones molecule is called a*nucleosome*. The nucleosomes are linked together by the long strand of DNA. To further condense the DNA material, nucleosomes are grouped together to form chromatin fibers. The chromatin fibers then fold together into large looped domains. During the mitotic cycle, the looped domains are organized into distinct structures called the chromosomes (see Figures 1.4 and 1.6). Chromosomes are also used as a way of referring to the genetic basis of an organism as either diploid or haploid. Many eukaryotic cells have two sets of the chromosomes and are called *diploid*. Other cells that only contain one set of the chromosomes are called *haploid*. The chromosome also plays an important role in cell death-related aging phenomena. At the tips of chromosomes are segments called *telomeres*. As the cell replicates, the telomeres are shortened. Once

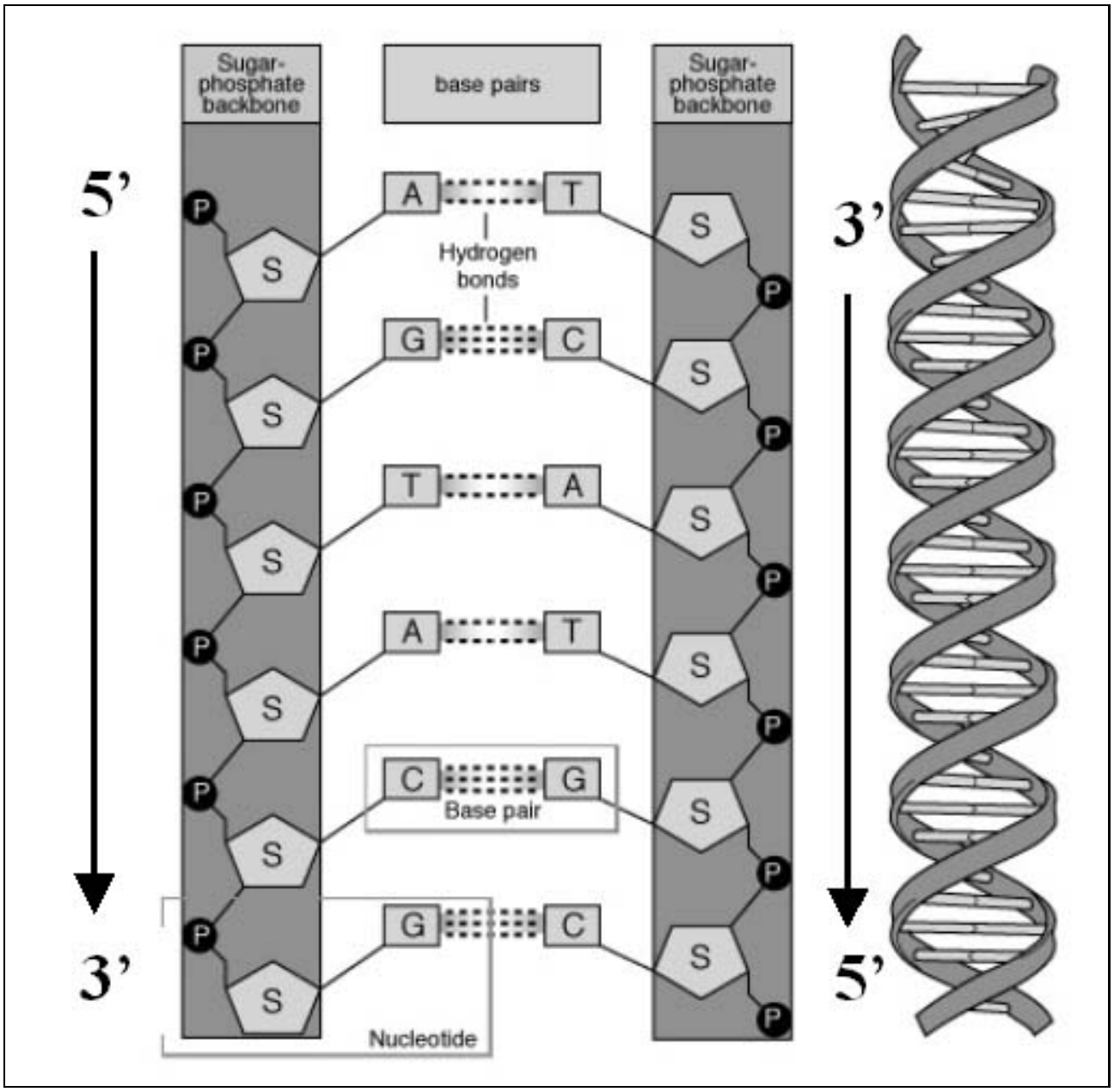


Figure 1.1: Source: [9]. DNA structure.

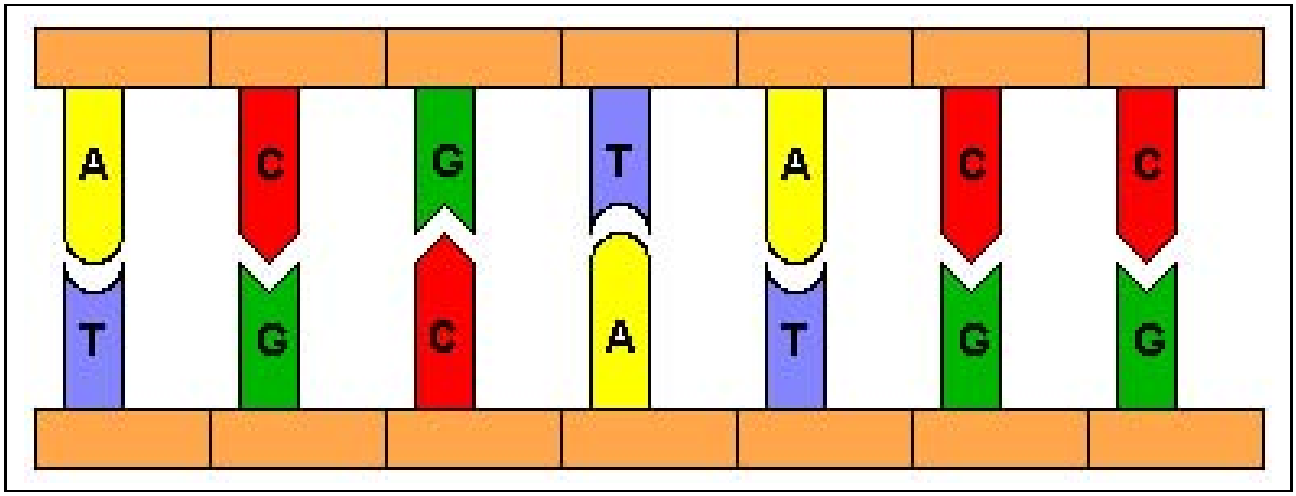


Figure 1.2: Source: [12]. Base pairs in DNA bond together to form a ladder-like structure. Because bonding occurs at angles between the bases, the whole structure twists into a helix.

the telomeres have been reduced to a certain level, the cell can no longer replicate itself and initiates *apoptosis*, the cellular death process. Today, much research efforts have been devoted to elucidating the specific mechanisms by which telomeres cause cell death (see [1], pages 397-401).

1.1.4 Genes

A gene is a region of DNA that controls a discrete hereditary characteristic, usually corresponding to a single mRNA carrying the information for constructing a protein (see [1], pages 98-99). It contains one or more regulatory sequences that either increase or decrease the rate of its transcription (see Figure 1.7). In 1977 molecular biologists discovered that most Eukaryotic genes have their coding sequences, called *exons*, interrupted by non-coding sequences called *introns* (see Figure 1.11). 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.

1.1.5 The Central Dogma

The expression of the genetic information stored in DNA involves the translation of a linear sequence of nucleotides into aco-linear sequence of amino acids in proteins.
The flow is: DNA→ RNA → Protein (see Figures 1.8 and 1.9).

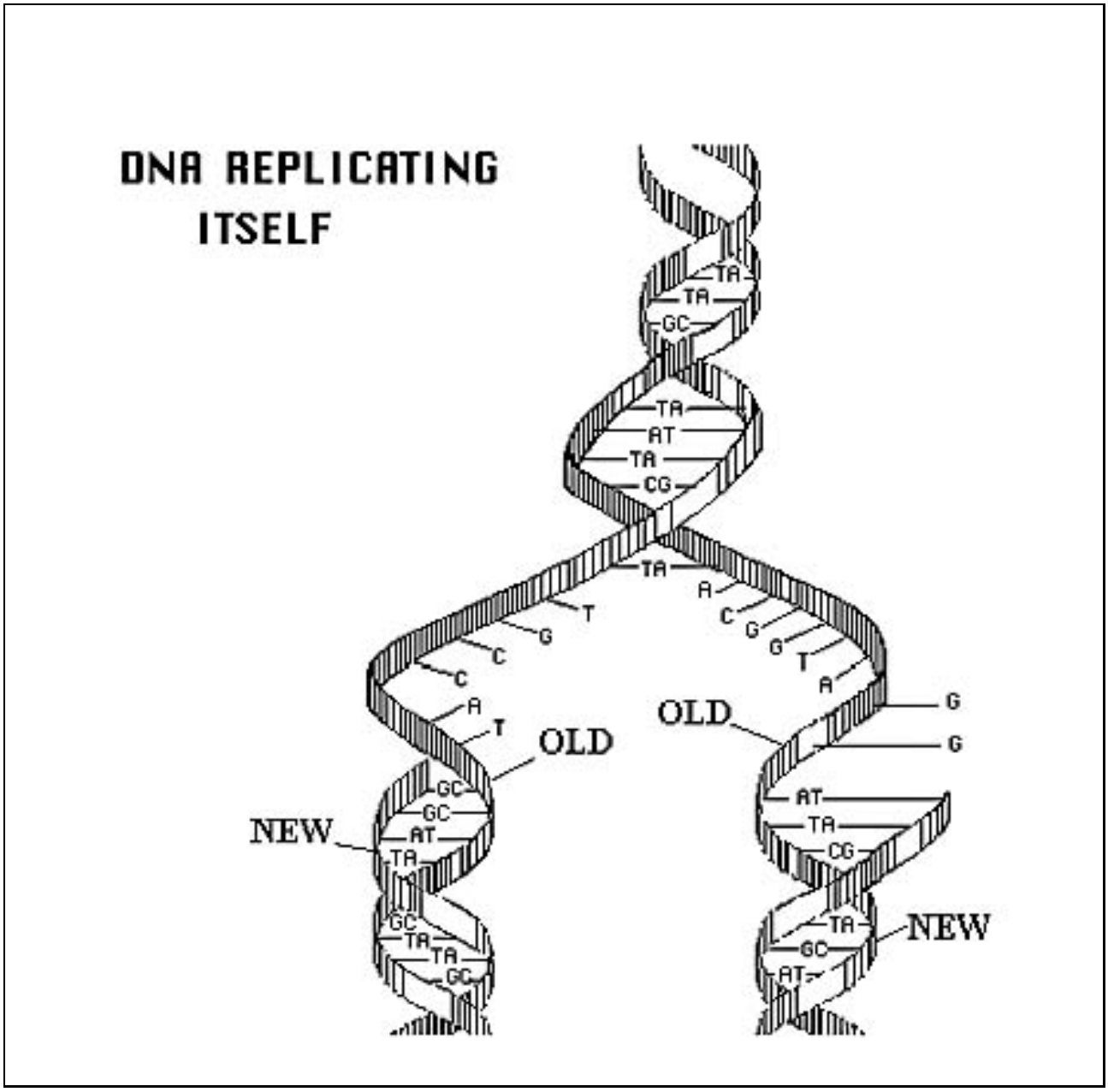


Figure 1.3: Source: [6]. DNA Replication.

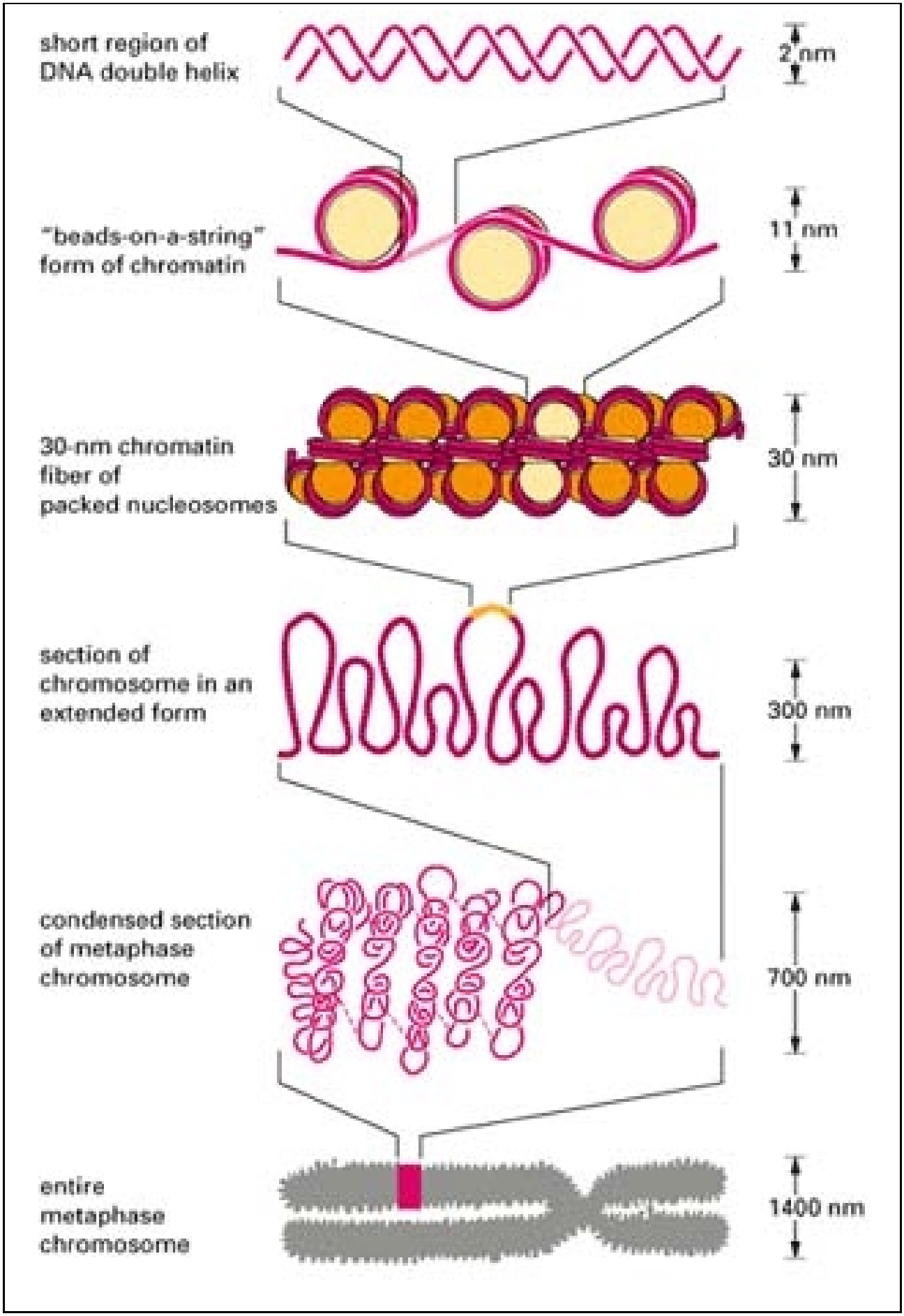


Figure 1.4: Source: [3]. A diagram showing the different levels of DNA packaging. Notice that DNA starts out as a single strand double helix and is condensed until it reaches the

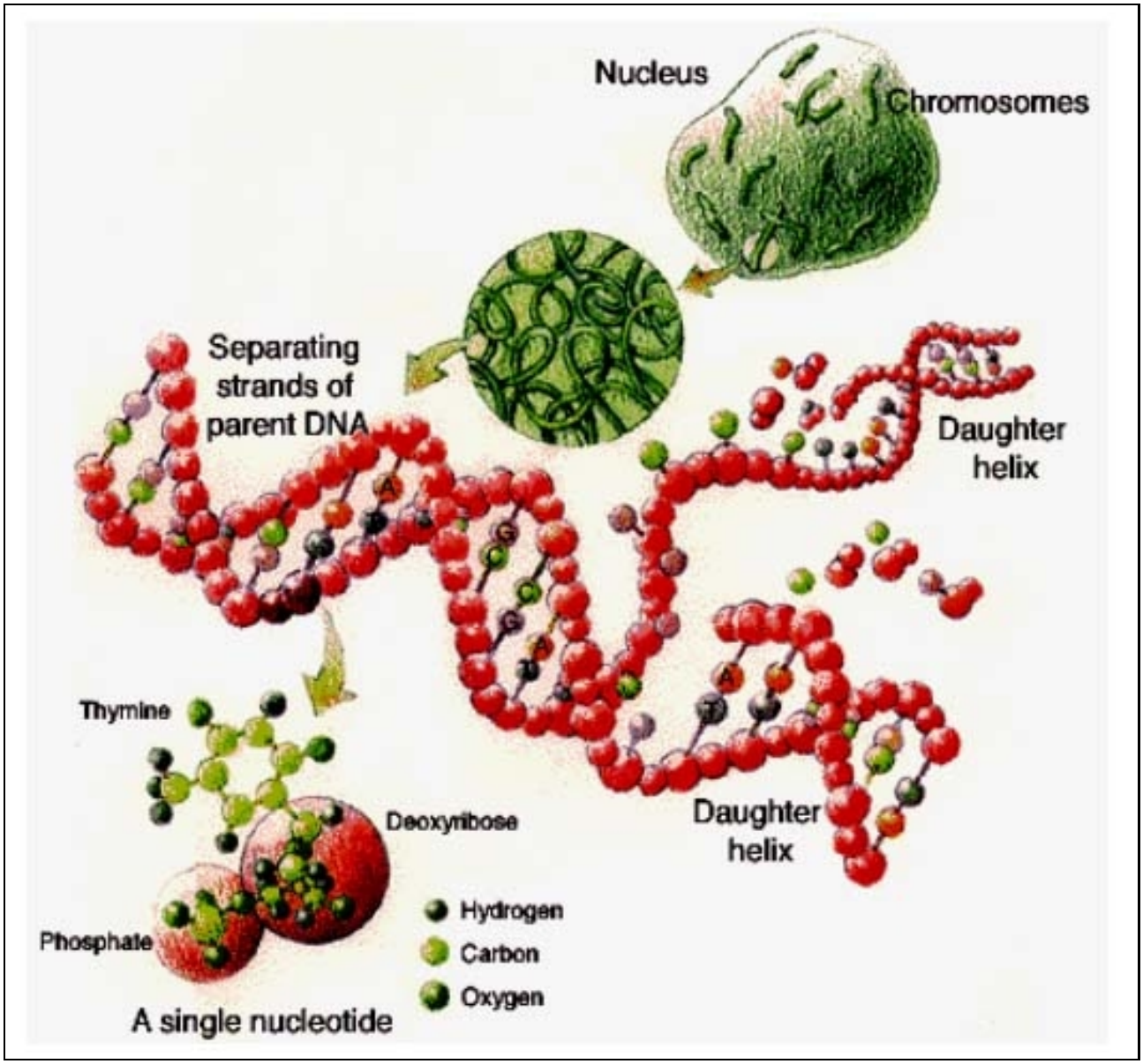


Figure 1.5: Source: [16]. Chromosomes.

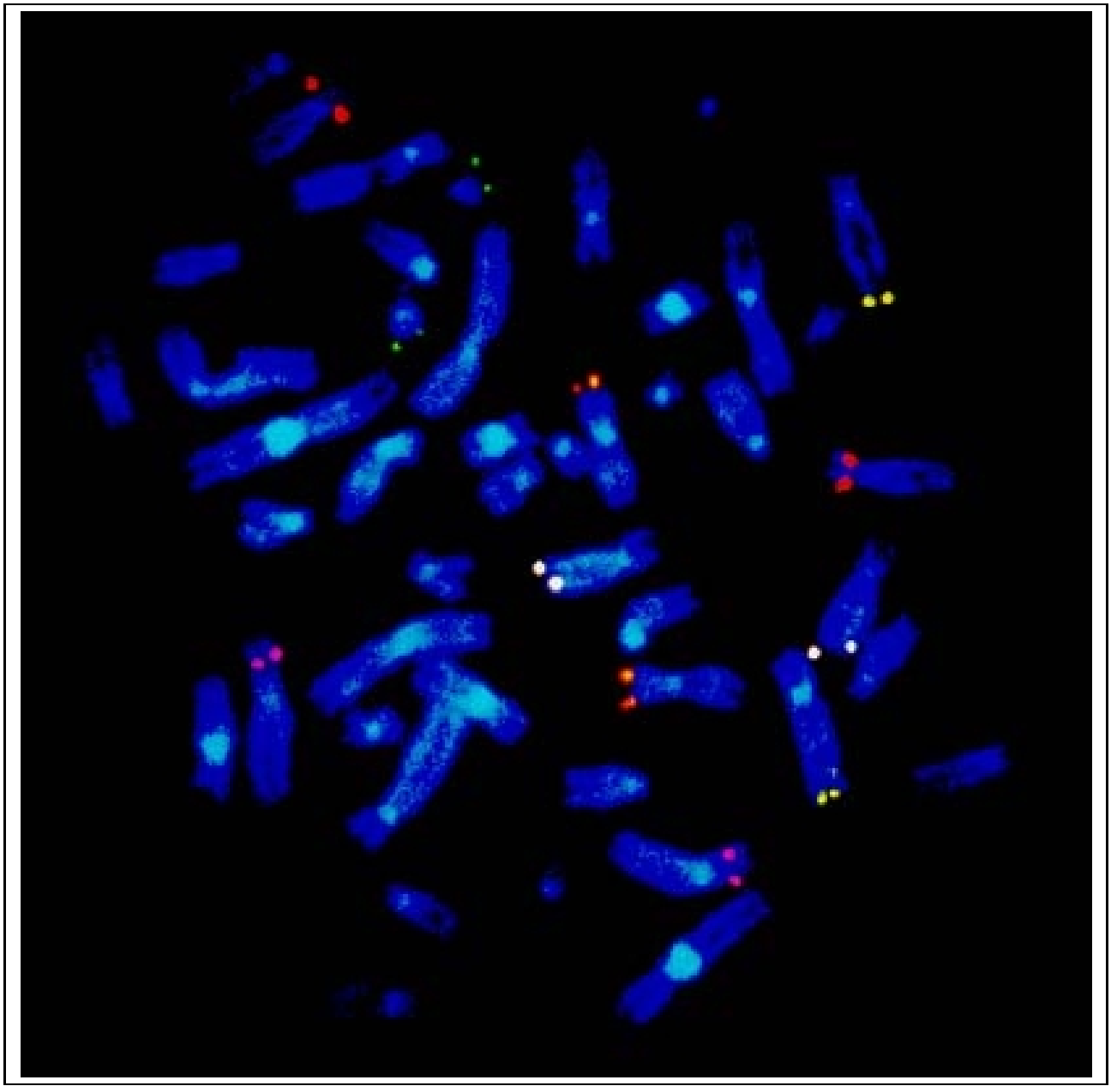


Figure 1.6: Source: [18]. Stained chromosomes. Chromosome 1 has the most genes (2968), and the Y chromosome has the fewest (231).

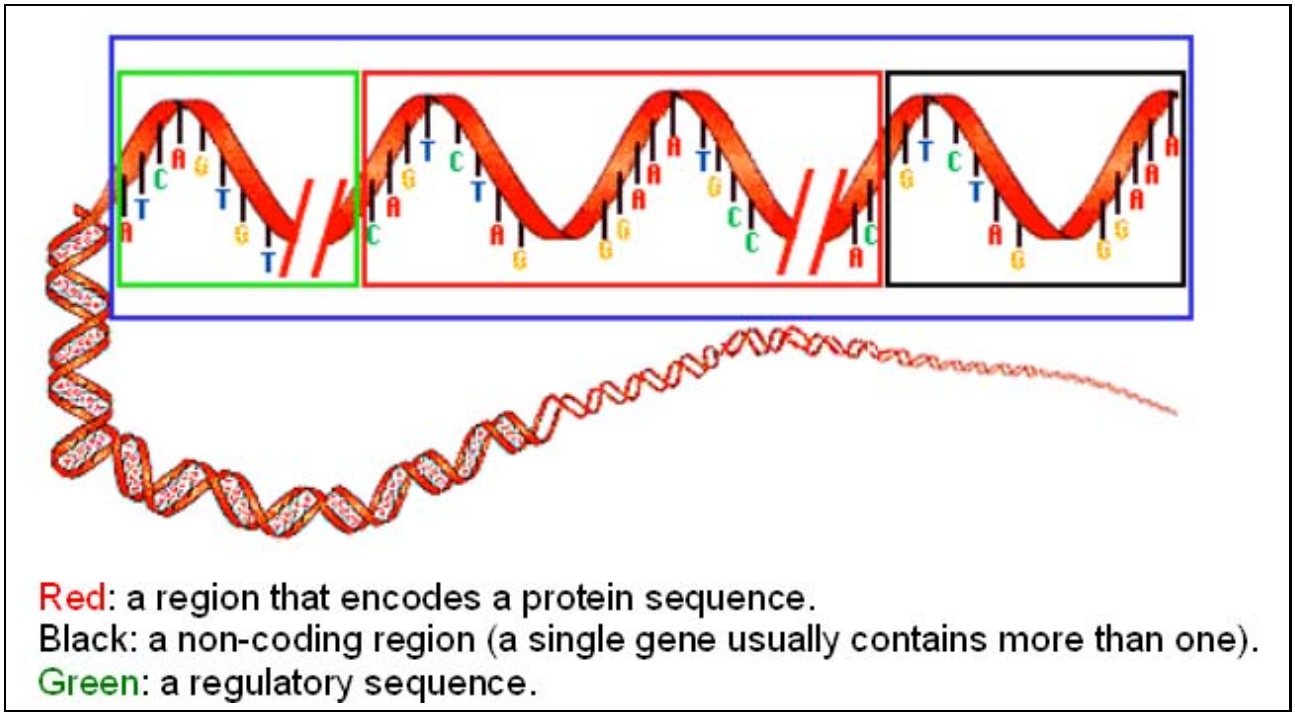


Figure 1.7: Source: [4]. Gene structure.

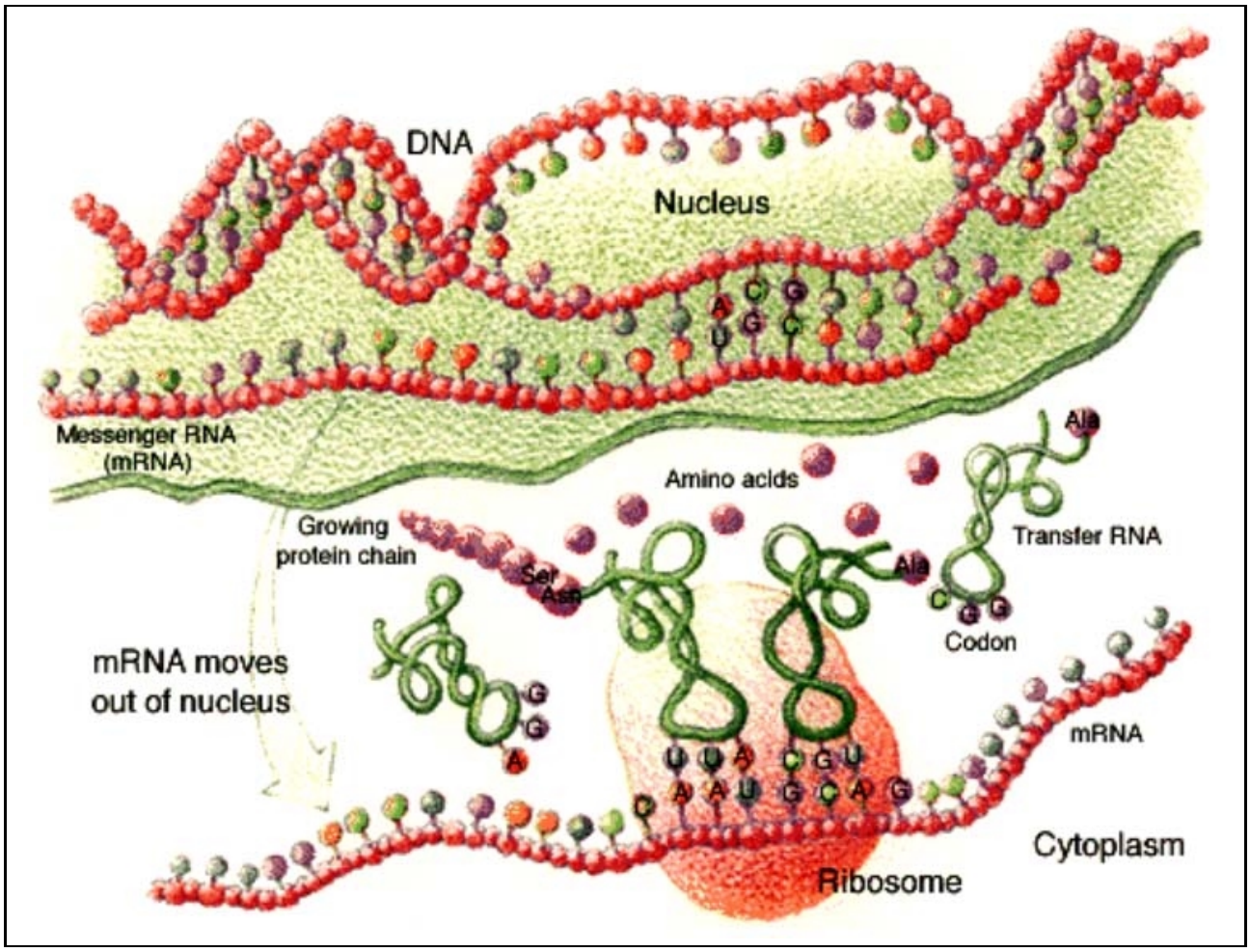


Figure 1.8: Source: [16]. From gene to protein.

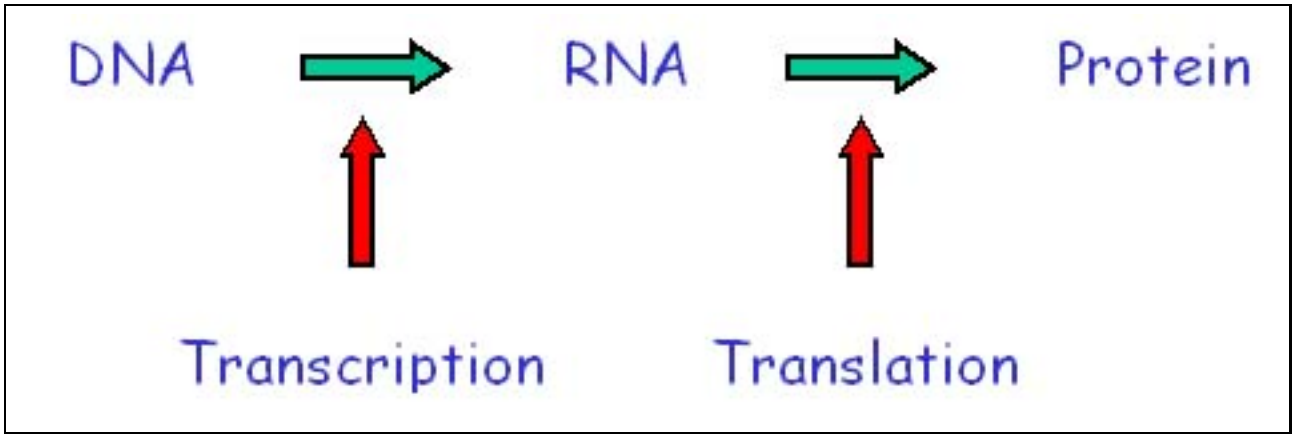


Figure 1.9: From DNA to Protein.

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* (see [1], pages 107, 200-202). Near most of the genes lies a special DNA pattern called *promoter*, 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 promoter 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 in the nucleus, RNA can also be found in the *cytoplasm* outside the nucleus, e.g. messenger RNA (mRNA) - molecules that direct the synthesis of proteins in the cytoplasm.

In Eukaryotic organisms, 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 3' end of the RNA molecule for protection (*poly-A tail*). The 5' end of the RNA is modified by capping with a (+)Gppp - positively charged 7-methylguanosine (see [1], pages 331, 413). This 5' CAP plays an important part in the initializing of protein synthesis by the protecting the growing RNA transcript from degradation.

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 (see [1], page 108). The sequence of nucleotides in the mRNA molecule 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 1.10). 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 the 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.

Splicing

Before the RNA molecule leaves the nucleus, a complex of RNA processing enzymes removes all the intron sequences, in a process called *splicing* (see [1], pages 412-422), thereby producing a much shorter RNA molecule (see Figure 1.12). Typical Eukaryotic exons are of average length of 200bp, while the average length of introns is around 10,000bp (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, called *alternative splicing*, 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 mRNA, in order to undergo translation (see Figure 1.14).

Translation

The *translation* of mRNA into protein (see [1], pages 109-110, 199-213) 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) are used for the protein translation, while the non coding regions *3' untranslated region* (UTR) 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

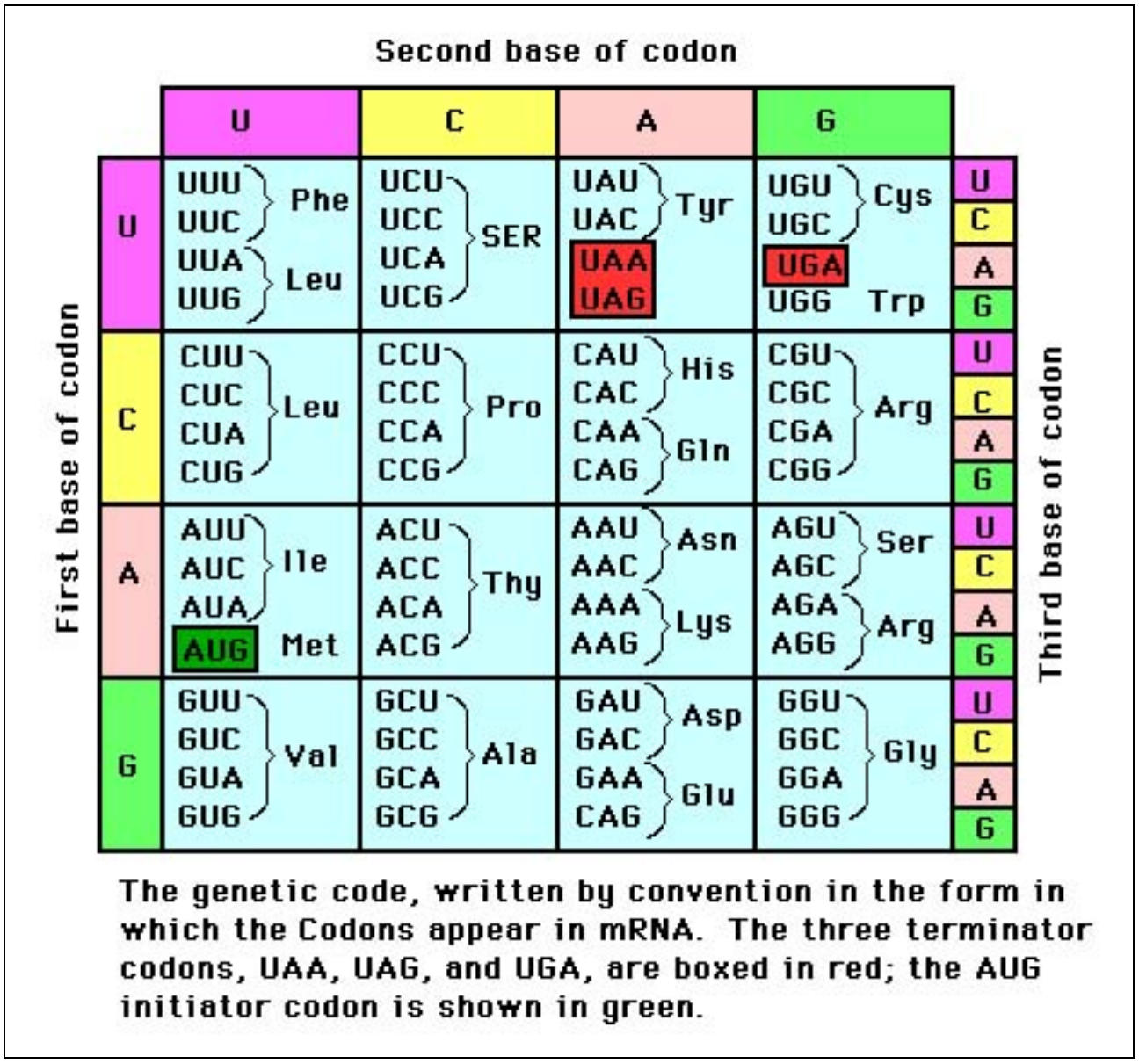


Figure 1.10: Source: [5]. The genetic code table.

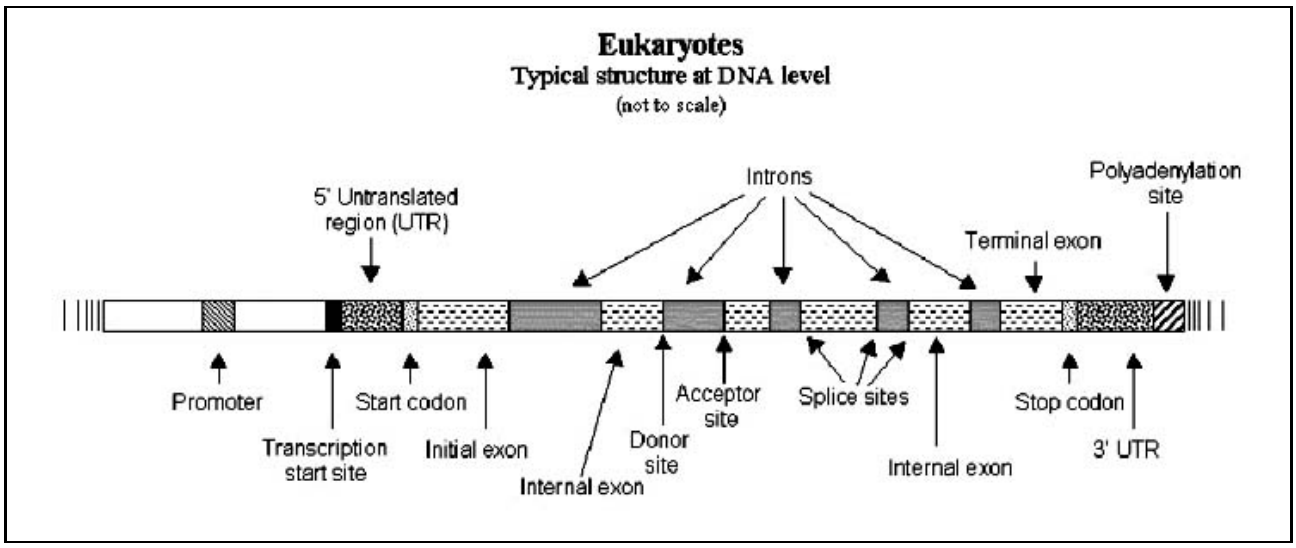


Figure 1.11: Gene structure in Eukaryotes.

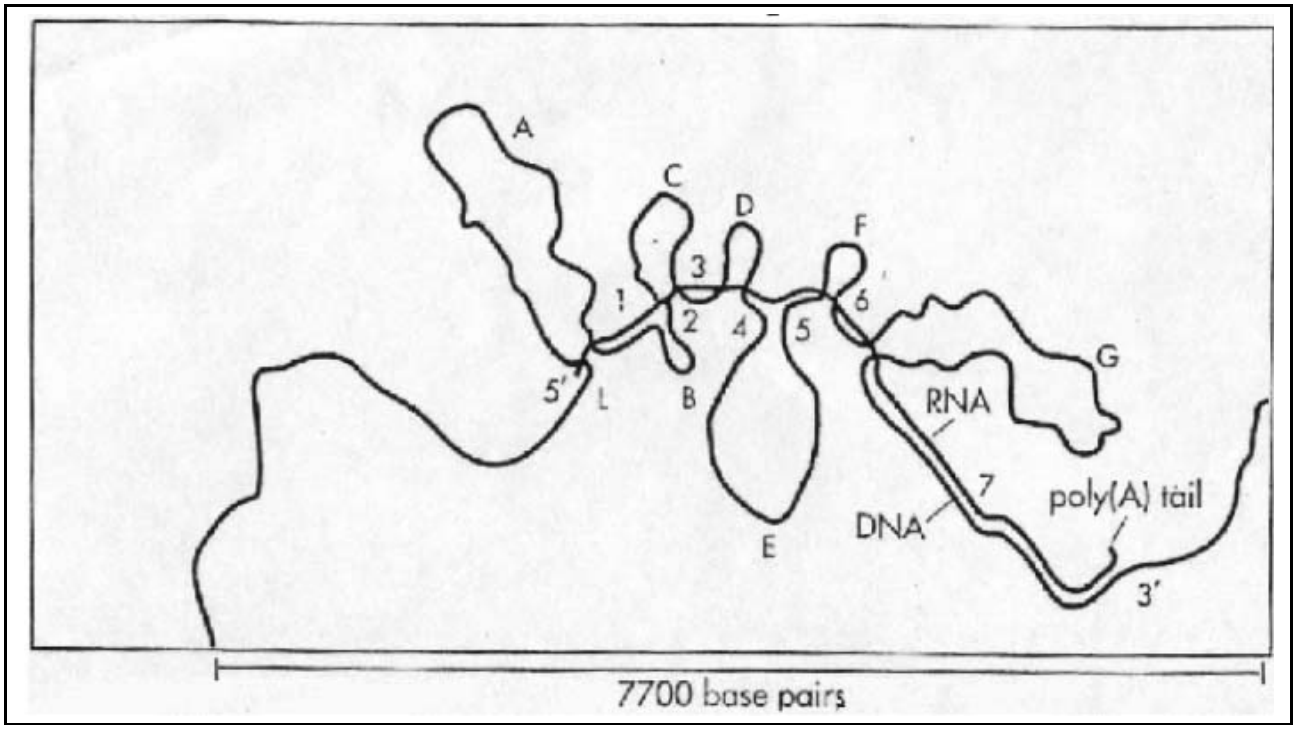


Figure 1.12: Introns are spliced out to form the mature mRNA.

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 (see Figure 1.13). 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. 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 transcribed from overlapping complementary regions of the DNA.

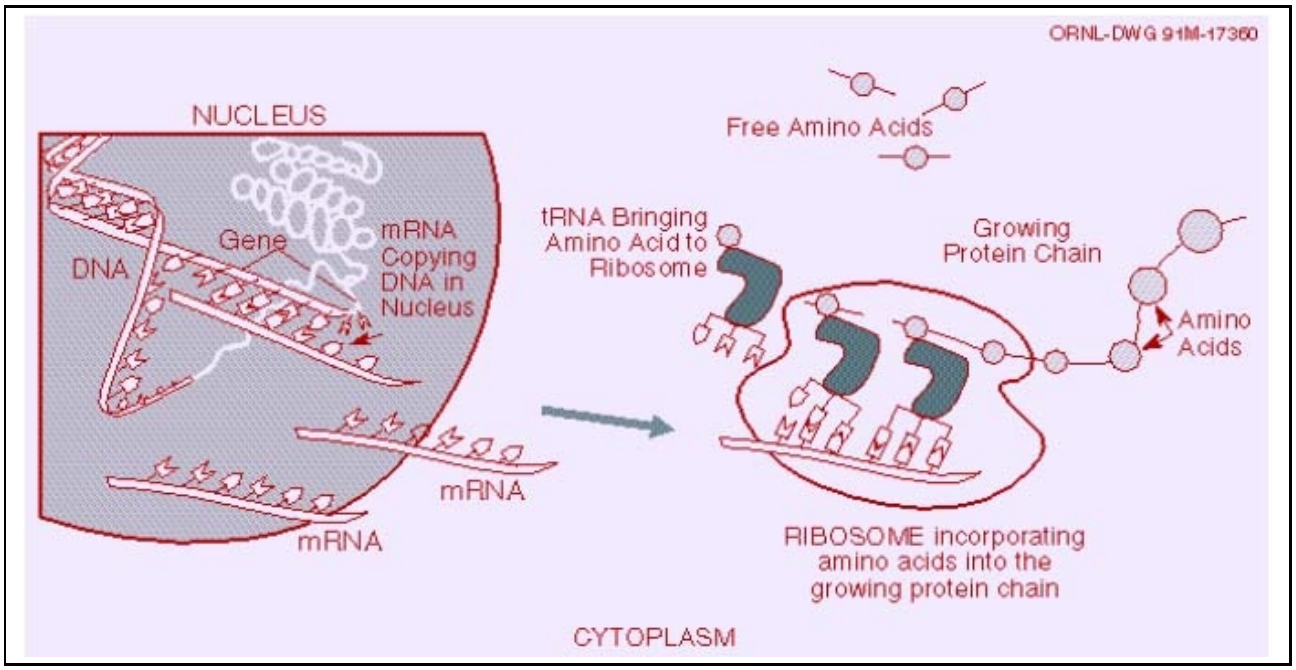


Figure 1.13: Source: [7]. Translation.

For animation of transcription, splicing and translation see [13].

Proteins

A protein is linear polymer of amino acids linked together by peptide bonds (see [1], pages 111-127). 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 hierarchical

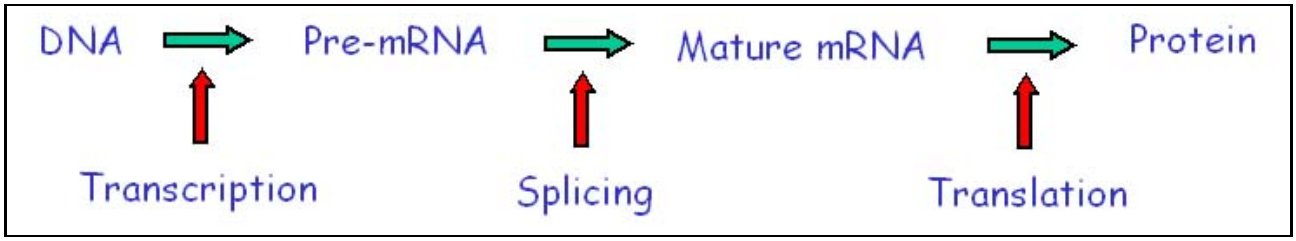


Figure 1.14: A diagram showing the process of gene expression.

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 *alpha-helices* which are single stranded helices of amino acids, and *beta-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*. 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 (see [1], pages 58-59). 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.

1.2 Basic Biotechnology

1.2.1 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 , the incubation mixture contains a 2',3' di-deoxy (radioactively labeled) analog of one of them. The incorporation of this analog, blocks further growth of the 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. 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 500-700 nucleotides in a few hours. For animation of sequencing see [17] and [10].

1.2.2 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 emphin-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 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.

For animation of the PCR procedure see [19].

1.3 The Human Genome

Following are some statistics on the human genome:

- 23 pairs of chromosomes comprise the human genome.

- The human genome contains 3,164.7 million nucleotide bases.
- The average gene consists of 3,000 bases, but sizes vary greatly, with the largest known human gene being dystrophin at 2.4 million bases.
- The total number of genes is estimated at 30,000 to 40,000, much lower than previous estimates of 80,000 to 140,000 that had been based on extrapolations from gene-rich areas as opposed to a composite of gene-rich and gene-poor areas.
- The total number of protein variants is estimated as 1,000,000.
- The order of almost all (99.9%) nucleotide bases is exactly the same in all people.

1.3.1 The Human Genome Project

The human genome project was launched in 1990 and was planned to be completed by 2005. There are over 50 participating laboratories located mainly in USA, Europe and Japan. The ultimate goal of the project is to produce a single continuous sequence for each of the 24 human chromosomes and to delineate the positions of all genes. The Human Genome Project is expected to produce a sequence of DNA representing the functional blueprint and evolutionary history of the human species. However, only about 3% of this sequence is thought to specify the portions of our 25,000 to 40,000 genes that encode proteins. Thus an important part of basic and applied genomics is to identify and localize these genes in a process known as transcript mapping. When genes are expressed, their sequences are first converted into messenger RNA transcripts, which can be isolated in the form of complementary DNAs (cDNAs). A small portion of each cDNA sequence is all that is needed to develop unique gene markers, known as sequence tagged sites or STSs, which can be detected in chromosomal DNA by assays based on PCR. To construct a transcript map, cDNA sequences from a master catalog of human genes were distributed to mapping laboratories in North America, Europe, and Japan. These cDNAs were converted to STSs and their physical locations on chromosomes determined on one of two radiation hybrid (RH) panels or a yeast artificial chromosome (YAC) library containing human genomic DNA. This mapping data was integrated relative to the human genetic map and then cross-referenced to cytogenetic band maps of the chromosomes.

The 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 - The human genome project is 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 the intent to sequence the human genome within three years. The company, later named *Celera* (see [8]), introduced a new ambitious 'whole genome shotgun' approach.

- 1999 - The public project responds to Ventner's challenge and changes their time destination for completing the first draft.

- December 1999 - The first complete human chromosome sequence (number 22) is 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 (see [14] and [20]).

For a more detailed timetable of the Human Genome Project see [21].

Progress

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 Figure 1.15). The total number of genes in human is estimated to be between 25,000 and 40,000. The Human Genome Project needs 1-2 more years for "real" completion.

Contribution

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 now here 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.

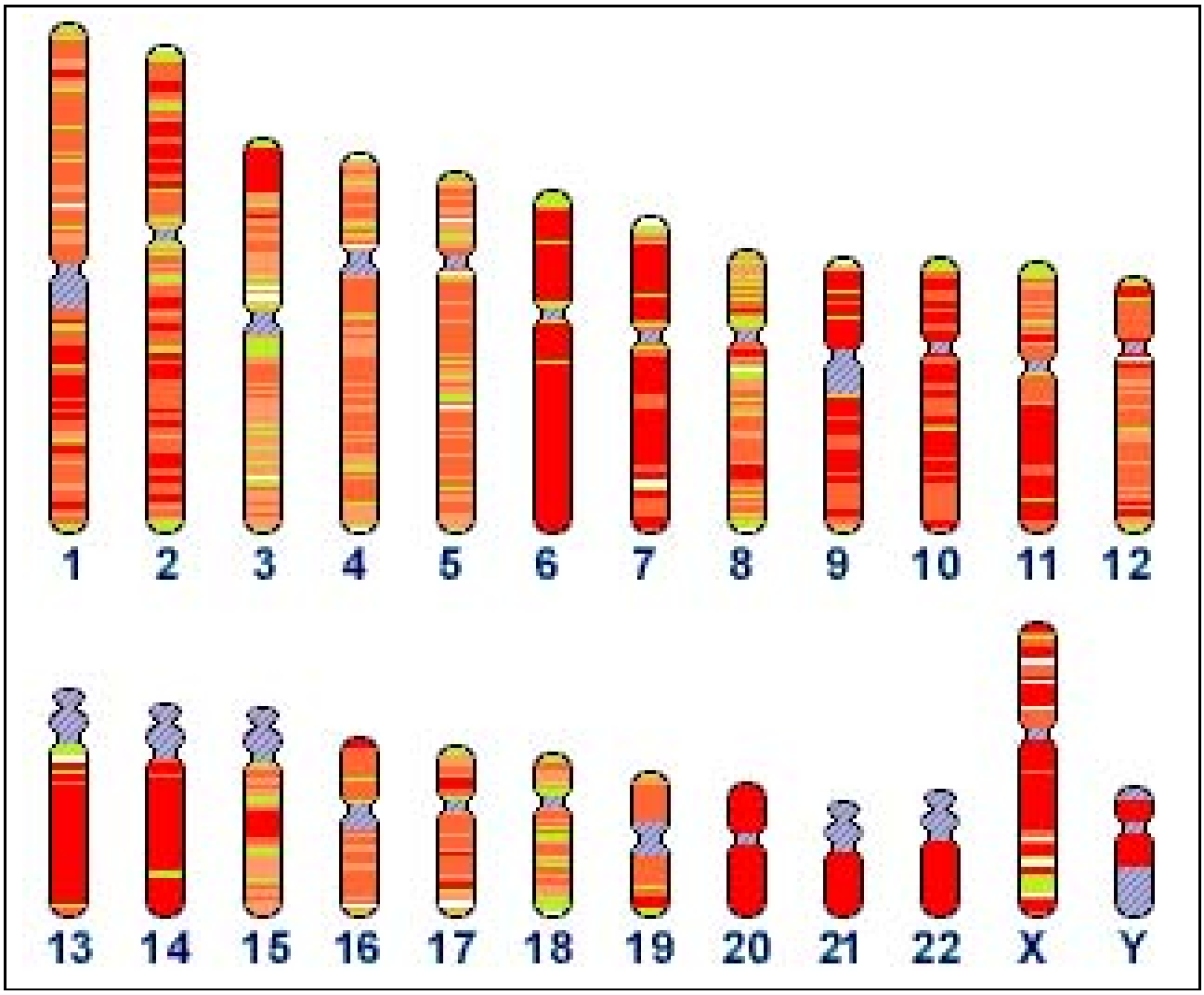


Figure 1.15: Source: [15]. The Public Human genome sequencing progress as of 31/12/2001, scaled from green = draft to red = finished, 63% finished, 34.8% draft, 97.8% total.

Current and Potential Applications of Genome Research

Molecular Medicine:

- Improve diagnosis of disease.
- Detect genetic predispositions to disease.
- Create drugs based on molecular information.
- Use gene therapy and control systems as drugs.
- Design "custom drugs" based on individual genetic profiles.

Microbial Genomics:

- Rapidly detect and treat pathogens (disease-causing microbes) in clinical practice.
- Develop new energy sources (biofuels).
- Monitor environments to detect pollutants.
- Protect citizenry from biological and chemical warfare.
- Clean up toxic waste safely and efficiently.

Risk Assessment:

- Evaluate the health risks faced by individuals who may be exposed to radiation (including low levels in industrial areas) and to cancer-causing chemicals and toxins.

Bioarchaeology, Anthropology, Evolution, and Human Migration:

- Study evolution through germline mutations in lineages.
- Study migration of different population groups based on maternal genetic inheritance.
- Study mutations on the Y chromosome to trace lineage and migration of males.
- Compare breakpoints in the evolution of mutations with ages of populations and historical events.

DNA Identification:

- Identify potential suspects whose DNA may match evidence left at crime scenes.
- Exonerate persons wrongly accused of crimes.
- Identify crime, catastrophe, and other victims.
- Establish paternity and other family relationships.
- Identify endangered and protected species as an aid to wildlife officials (could be used for prosecuting poachers).
- Detect bacteria and other organisms that may pollute air, water, soil, and food.
- Match organ donors with recipients in transplant programs.
- Determine pedigree for seed or livestock breeds.
- Authenticate consumables such as caviar and wine.

Agriculture, Livestock Breeding, and Bioprocessing:

- Grow disease-, insect-, and drought-resistant crops.
- Breed healthier, more productive, disease-resistant farm animals.
- Grow more nutritious produce.
- Develop biopesticides.
- Incorporate edible vaccines into food products.
- Develop new environmental cleanup uses for plants like tobacco.

1.3.2 After the HGP, the Next Steps

The words of Winston Churchill, spoken in 1942 after 3 years of war, capture well the HGP era: "Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning". The avalanche of genome data grows daily. The new challenge will be to use this vast reservoir of data to explore how DNA and proteins work with each other and the environment to create complex, dynamic living systems. Systematic studies of function on a grand scale–functional genomics–will be the focus of biological explorations in this century and beyond. These explorations will encompass studies in transcriptosmics, proteomics, structural genomics, new experimental methodologies, and comparative genomics.

1.4 DNA Chips and Microarrays

Even though there are no doubts about the importance of the Human Genome Project, the knowledge of the full DNA sequence of an organism represents only the first necessary step towards the understanding of the connection between the DNA sequence and the phenotypical characteristics of a living organism.

1.4.1 Functional Genomics

Functional Genomics is a study of the functionality of specific genes, their relations to diseases, their associated proteins and their participation in biological processes. It is widely believed that thousands of genes and their products (i.e., RNA and proteins) in a given living organism function in a complicated and orchestrated way that creates the mystery of life. However, traditional methods in molecular biology generally work on a "one gene in one experiment" basis, which means that the throughput is very limited and the "whole picture" of gene function is hard to obtain. *Reductionist* approach to functional genomics is hypothesis driven - we proceed by suggesting a hypothesis and designing an experiment to check its correctness. However, the complexity of living organisms makes the challenge of fully understand complex biology unachievable using these methods, and a new paradigm, holistic and high throughput is emerging instead. Technologies for simultaneously analyzing the expression levels of large numbers of genes provides the opportunity to study the activity of whole genomes, rather than the activities of single, or a few, genes. In the long-term, large-scale gene expression analysis will enable the behavior of co-regulated gene networks to be studied. The technology can be used to look for groups of genes involved in a particular biological process or in a specific disease by identifying genes whose expression levels change under certain circumstances. The RNA transcription profiles of wild type (a normal organism) and mutant or transgenic organism can be compared using gene expression technologies, thus providing an overall analysis of the impact of a particular genetic change on gene expression.

1.4.2 The DNA Chip

Terminologies that have been used in the literature to describe this technology include, but not limited to: biochip, DNA chip, DNA microarray, and gene array. An array is an orderly arrangement of samples. Those samples can be either DNA or DNA products. Each spot in the array contains many copies of the sample. The array provides a medium for matching known and unknown DNA samples based on base-pairing (hybridization) rules and automating the process of identifying the unknowns. The sample spot sizes in microarray are typically less than 200 microns in diameter and these arrays usually contain thousands of spots. As a result microarrays require specialized robotics and imaging equipment. An

experiment with a single DNA chip can provide researchers information on thousands of genes simultaneously - a dramatic increase in throughput.

Oligonucleotide Arrays

Before going over the process of research with DNA chips it is best to clarify two basic and sometimes confusing nomenclatures. A *probe* is the tethered nucleic acid with known sequence which we use in order to discover information about the *target* which is the free nucleic acid sample whose identity/abundance is being detected. The basic idea, developed (and patented) by a company named Affymetrix, is to generate probes that would capture each coding region as specifically as possible. The length of the oligos (a sequence of nucleotides) used depends on the application, but the yare usually no longer than 25 bases. Since the oligos are short, the density of these chips is very high, for instance, a chip that of 1cm by 1cm can easily contain 100,000 oligo types. There are two variants of the oligo nucleotide arrays technology, in terms of the property of arrayed DNA sequence with known identity:

- Format I: The target (500-5,000 bases long) is immobilized to a solid surface such as glass using robot spotting and exposed to a set of probes either separately or in a mixture. This format is traditionally called DNA microarray.

- Format II: An array of oligonucleotide (20-80 mer oligos) or peptide nucleic acid (PNA) probes is synthesized either in situ (on-chip) or by conventional synthesis followed by on-chip immobilization. The array is exposed to labeled sample DNA, hybridized, and the identity/abundance of complementary sequences are determined. This format is historically called DNA chip.

Manufacturing Oligonucleotide Arrays

Oligonucleotide arrays are produced in a way that is similar to the way computer chips are. We start with a matrix created over a glass substrate. Each cell in the matrix contains a "chain" with appropriate chemical properties, and ending with an *emterminator*, a chemical gadget that prevents chain extension. This substrate is covered with a mask, covering some of the cells, but not others, and then illuminated. Covered cells are unaffected. In cells that are hit by the light, the bond with the terminator is severed. If we now expose the substrate to a solution containing a nucleotide base, it will form bonds with the non-terminated chains. Thus, some of the cells will now contain this nucleotide. The process can then be repeated with different masks (which covers different cells), and for different nucleotides. This way one can insert a specific nucleotide to each cell of the matrix, and manufacture a specific oligonucleotide. Figure 1.16 demonstrates the production process.

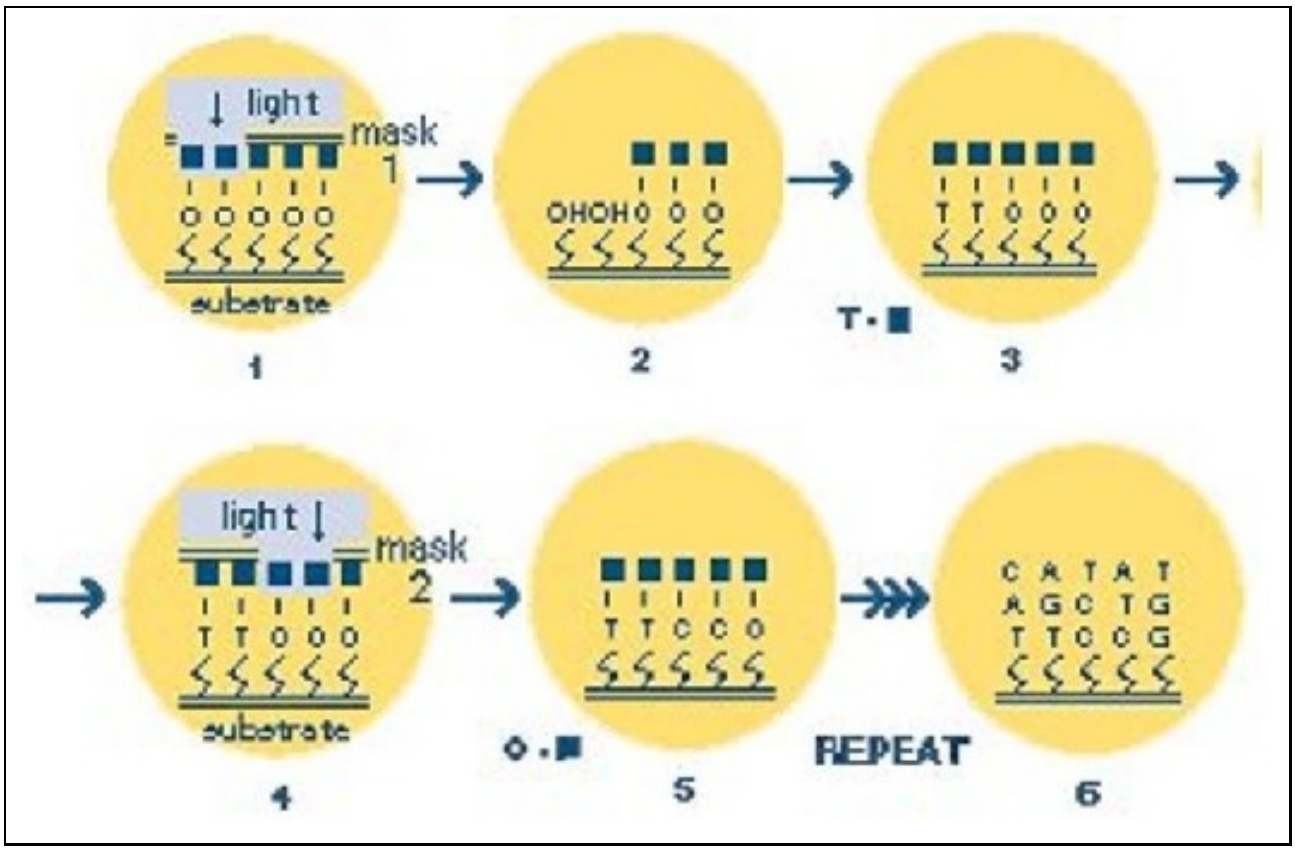


Figure 1.16: Manufacturing DNA chips. 1-2) The light removesthe terminator from the chains not covered by the mask, creating hydrogen bonds instead. 3) Bonds are formed with a nucleotide base. 4-6) The process is repeated with a different base.

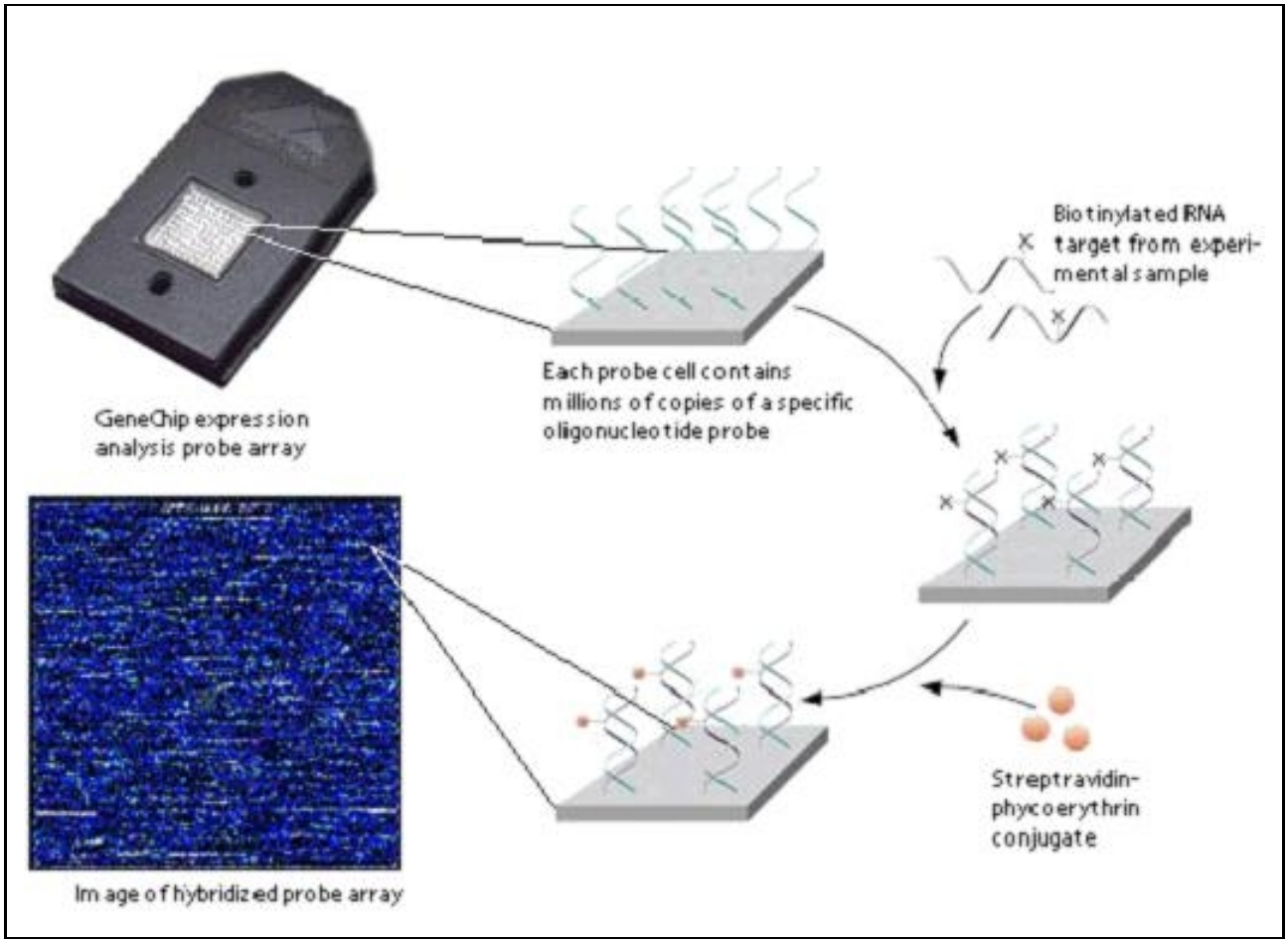


Figure 1.17: A typical experiment with an oligonucleotide chip. Labeled RNA molecules are applied to the probes on the chip, creating a fluorescent spot where hybridization has occurred.

Sequencing by Hybridization

This is one application of DNA chips intended to identify a sequence of a gene/gene mutation. Sequencing by Hybridization (SBH) uses sequencing chips in the Format II method. A chip contains all the possible sequences of a given length (usually 8-10 bases long). For example, if a chip is based on 3-mer oligos segments, the chip will contain 64 different segments or probes (AAA, TTT, ..., AAC, ...). Target samples are marked, either with fluorescent dye or radioactive label and then introduced to the chip. A specific Target "sticks to" (or hybridizes with) the segments, which are part of the target itself. After the hybridization the spots with sequences which the target hybridized with will be marked and be part of the *target spectrum*. The spectrum consists of the sequences of k base pairs, which are part of the target, and according to it the target is deduced. In case of a chip containing all the sequences of 8 base pairs, the target can be up to 150-200 base pairs. For targets longer than that there will be computational problems because of repetitions of sequences inside the target. Even though a marked spot will appear different if it contains a sequence which appears once in the target than if the sequence is contained twice in it, it will be much harder to conclude if a sequence is contained four, five or six times in along target. As a consequence this method is not effective for very long targets and SBH is not competitive for sequencing targets. But we will see that some improvements can be made that may yet prove competitive.

Oligo-Fingerprinting

This type of chip was the first to be used, and is, in a sense, the opposite to Affymetrix approach. The chip consists of a matrix, with each cell of the matrix containing a target DNA. The chip is exposed to a solution containing many **identical oligos**, and hybridization occurs between matching DNA and oligos. Again, if the oligos are tagged, either with fluorescent dye or radioactive label, we can then see at each point of the matrix indications of whether the hybridization occurred (i.e., which of the DNAs hybridized to the oligo we tested). The chip can then be heated, separating the oligos from the DNA, and the experiment can be repeated with a different type of oligo. Finally, we get a matrix, with each row representing a specific target DNA from the matrix, and each column representing an oligo. The advantage of this method is the fact that a single experiment produces information about multiple targets. The main disadvantage, however, is that each experiment contains only a limited number of probes (about 300 oligos of 8 bases for example) in contrast to the SBH method.

cDNA Microarrays

This technology enables a researcher to analyze the expression of thousands of genes in a single experiment and provides quantitative measurements of the differential expression of these genes. In this approach, each spot in the chip contains, instead of short oligos, a

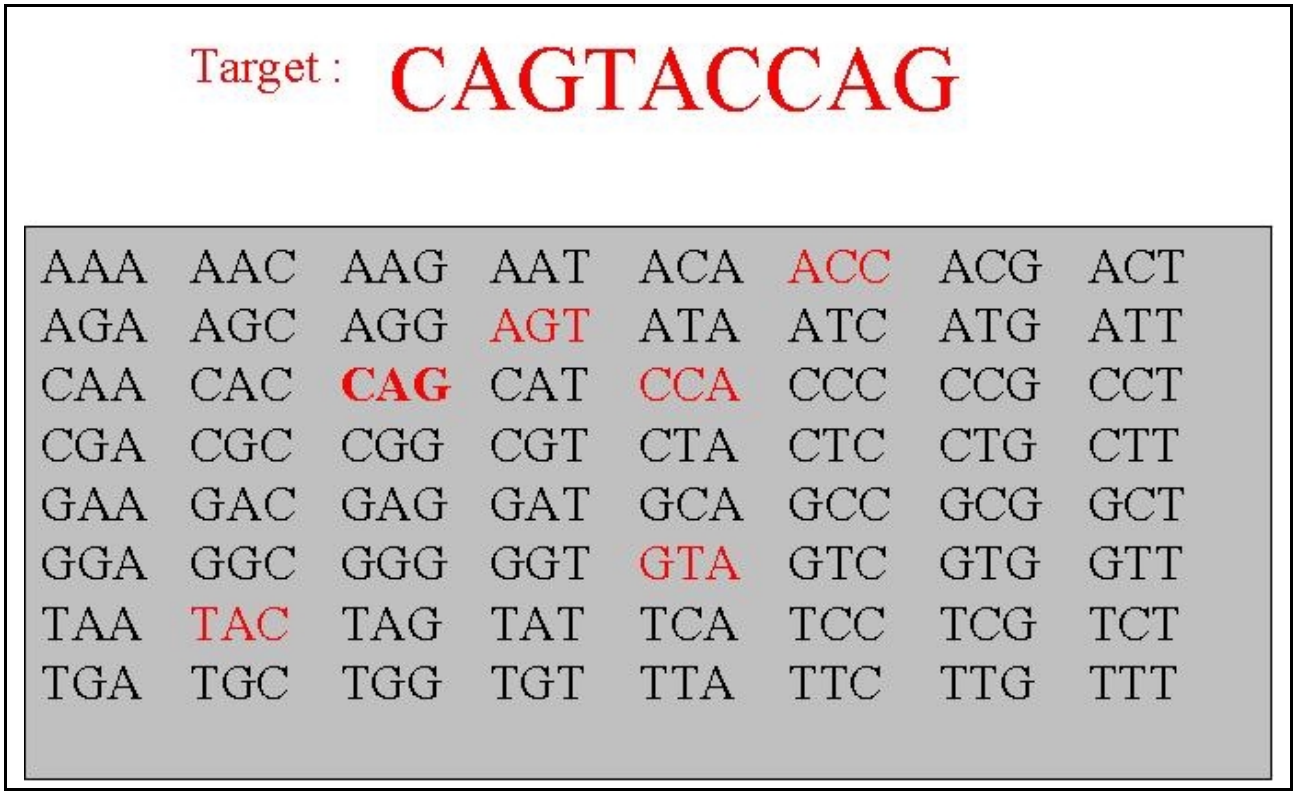
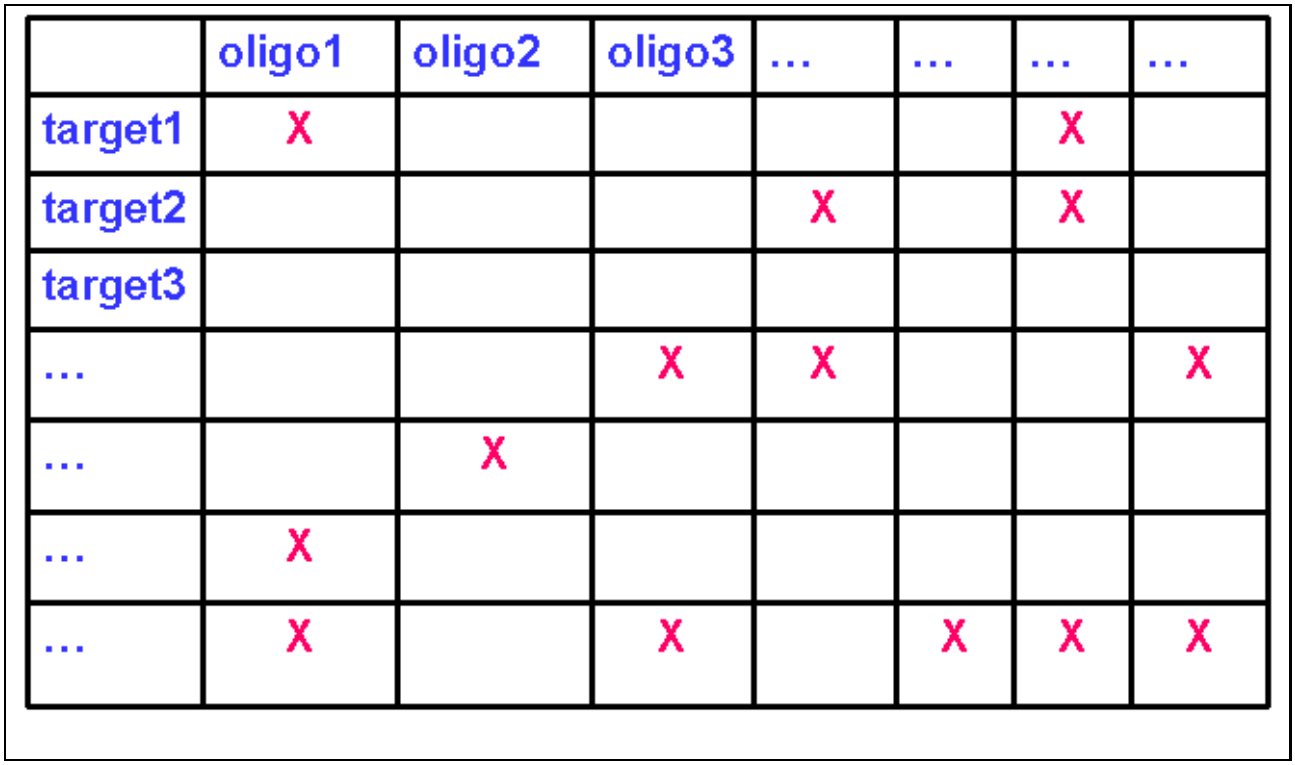


Figure 1.18: The Target Spectrum for 9 pair bases target introduced to a DNA chip with 3-mer oligos. Notice that the CAG segment is contained twice in the target and as a result the segment has a different color on the chip.



	oligo1	oligo2	oligo3	…	…	…	…
target1	X					X	
target2				X		X	
target3							
…			X	X			X
…		X					
…	X						
…	X		X		X	X	X

Figure 1.19: A matrix obtained from a fingerprinting experiment. Each line represents a fingerprint of a specific target - all the oligos that hybridizised with it.

cDNA clone, which represents a gene. The chip as a whole represents thousands of genes. The target is the mRNA extracted from a specific cell. Since almost all the mRNA in the cell is translated into a protein, the total mRNA in a cell represents the genes expressed in that cell. Therefore hybridization of mRNA is an indication of a gene being expressed in the target cell. Since cDNA clones are much longer than oligos (can be thousands of nucleotides long), a successful hybridization with a clone is an almost certain match for the gene. However, due to the different structure of each clone and the fact that unknown amount of cDNA is printed at each probe, we cannot associate directly the hybridization level with transcription level and so cDNA chips experiments are limited to comparisons of a reference extract and a target extract. Comparative genomic hybridization is designed to help clinicians determine the relative amount of a given genetic sequence in a particular patient. This type of chip is designed to look at the level of aberration. This is usually done by using a healthy tissue sample as a reference and comparing it with a sample from the diseased tumor. To perform a cDNA array experiment, we label green the reference extract, representing the normal level of expression in our model system, and label red the target culture of cells which were transformed to some condition of interest. We hybridize the mixture of reference and target extracts and read a green signal in case our condition reduced the expression level and a red signal in case our condition increased the expression level.

Expression Data The outcome of DNA Microarrays is a matrix associating for each gene (row) and condition/profile (column) the expression level. Expression levels can be absolute or relative.

Computational Challenges We wish to identify biological meaningful phenomena from the expression matrix, which is often very large (thousands of genes and hundreds of conditions). The most popular and natural first step in this analysis is clustering of the genes or experiments. Clustering techniques are used to identify subsets of genes that behave similarly under the set of tested conditions. By clustering the data, the biologist is viewing the data in a concise way and can try to interpret it more easily. Using additional sources of information (known genes annotations or conditions details), one can try and associate each cluster with some biological semantics. Other computational challenges are:

- Classification: Given a partition of the conditions into several types, classify an unknown tissue.
- Feature selection: Given a partition of the conditions into several types, find a subset of the genes for each type that distinguishes it from the rest.
- Normalization: How does one best normalizes thousands of signals from same/different conditions/experiments?

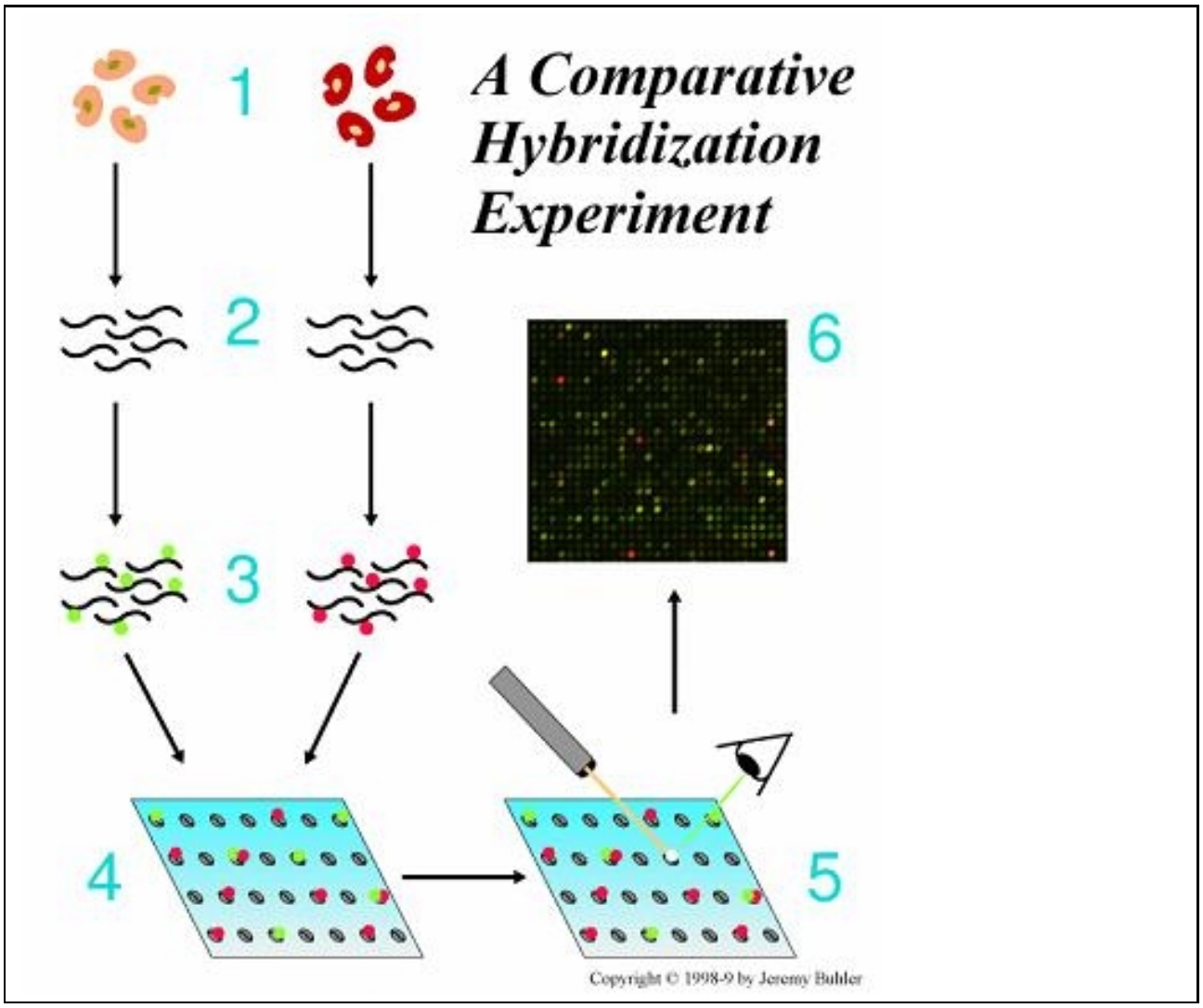


Figure 1.20: cDNA Microarray. 1) Two cells to be compared. On the left is the reference cell and on the right the target cell. 2) The mRNA is extracted from both cells. 3) Reference mRNA is labeled green, and the target mRNA is labeled red. 4) The mRNA is introduced to the Microarray. 5) According to the color of each gene clone the relative expression level is educed. 6) cDNA chip after scanning.

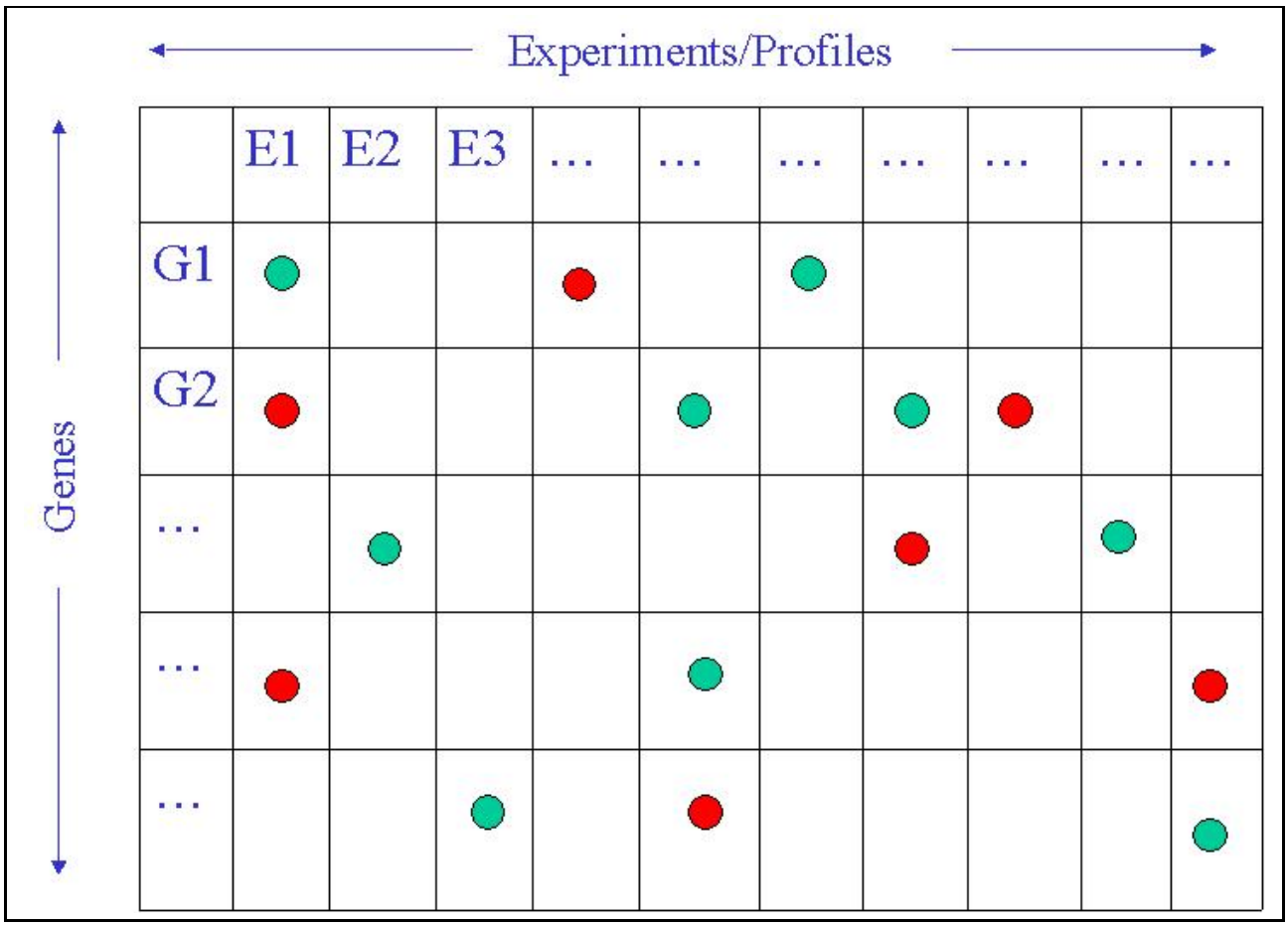


Figure 1.21: Raw Data. For each gene (row) we can see what was its relative expression in each cell, and for each profile or condition (column) we can see what genes were expressed abnormally.

- Experiment design: Choose which (pairs of) conditions will be most informative. The cells must differ in the condition under research but be alike as much as possible in all other aspects (phenotypes) in order to avoid distractions.

- Detect regulatory signals in promoter regions of co-expressed genes.

1.5 Gene Networks

A gene network is a set of molecular components such as genes and proteins and interactions between them that collectively carry out some cellular function. Since the development of the microarray technique in 1995, there has been an enormous increase in gene expression data from several organisms. Based on the view of gene systems as a logical network of nodes that influence each other's expression levels, scientists wish to be able to reconstruct the precise gene interaction network from the expression data obtained with this large scale arraying technique. Computer science shows that inference of a logical regulatory network is possible solely from sets of expression data, and mathematicians are working on the question how much data is necessary for reverse engineering. Meanwhile, experimental biologists are experiencing problems in the field. The number of experiments that are necessary before attempting network reconstruction is a lot more than is generally possible in "wet" laboratories, so data compression algorithms are applied to reduce the number of nodes considered. This is however an extremely coarse representation of the intricate interconnections that exist between single genes. The resulting network of only a handful of nodes is therefore usually only sufficient to describe the experiments performed, while any possible predicting properties are absent.

1.5.1 Examples

Figure 1.22 depicts gene expression and its role in catalyzing certain chemical reaction in the cell. The proB gene is being expressed into the gamma-glutamyl-kinase protein, which catalyzes a reaction involving glutamate and ATP, which produces gamma-glutamyl-phosphate and ADP compounds.

This gene expression is a part of a simple metabolic pathway, involving a chain of generated proteins, which is shown on figure 1.23. One of the final products of the chain, proline, inhibits the initial reaction, which has started the whole process. This "feedback inhibition" pattern is highly typical to genetic networks, and serves to regulate the process execution rate.

The following two figures (1.24 and 1.25) show a more complex gene network, describing Methionine biosynthesis in E-coli. The second figure is a shortcut representation of the pathway, with most nodes omitted, but it can give a better idea on overall topology.

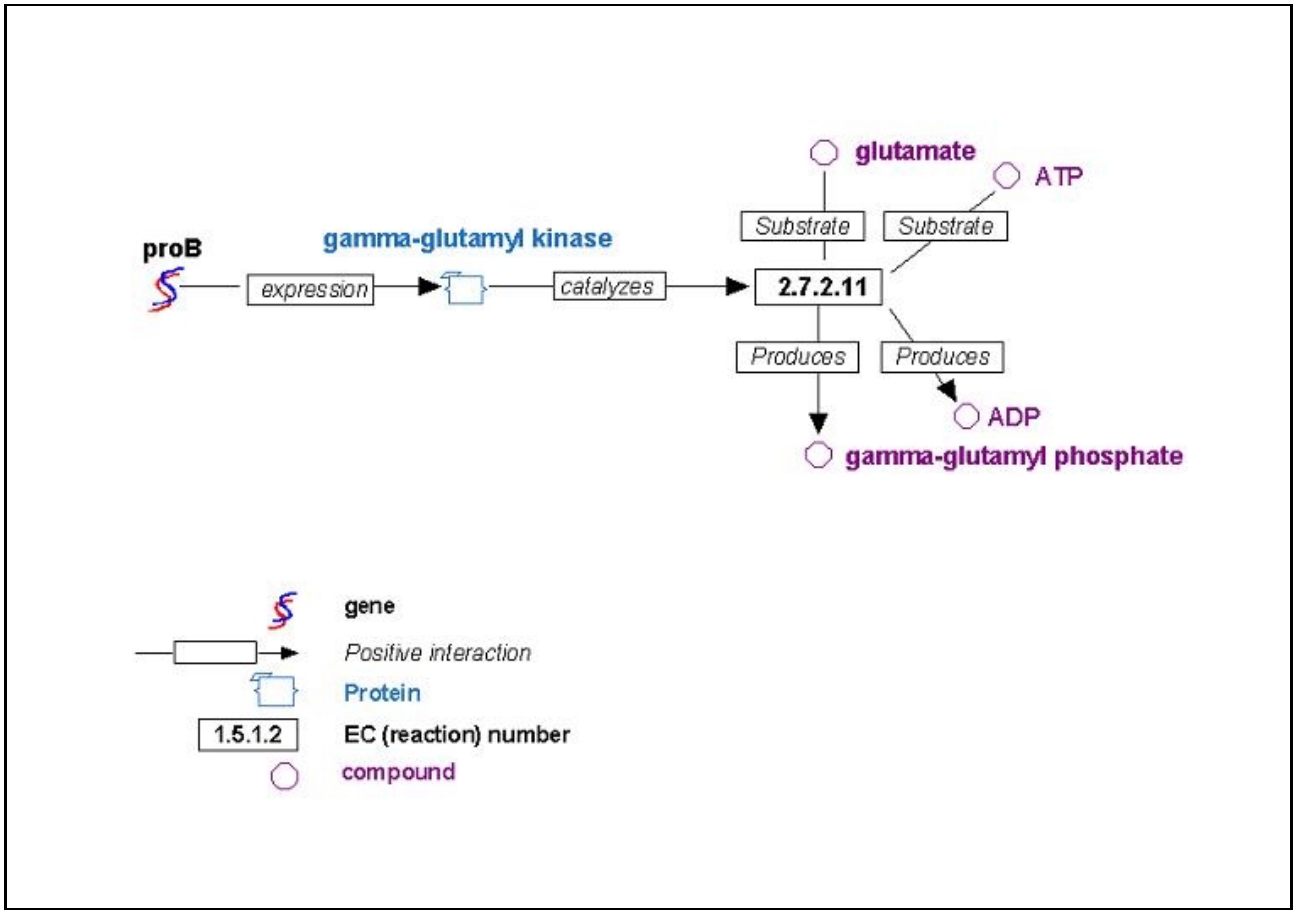


Figure 1.22: Source: [11]. An example for the role of gene expression in catalyzing chemical reactions.

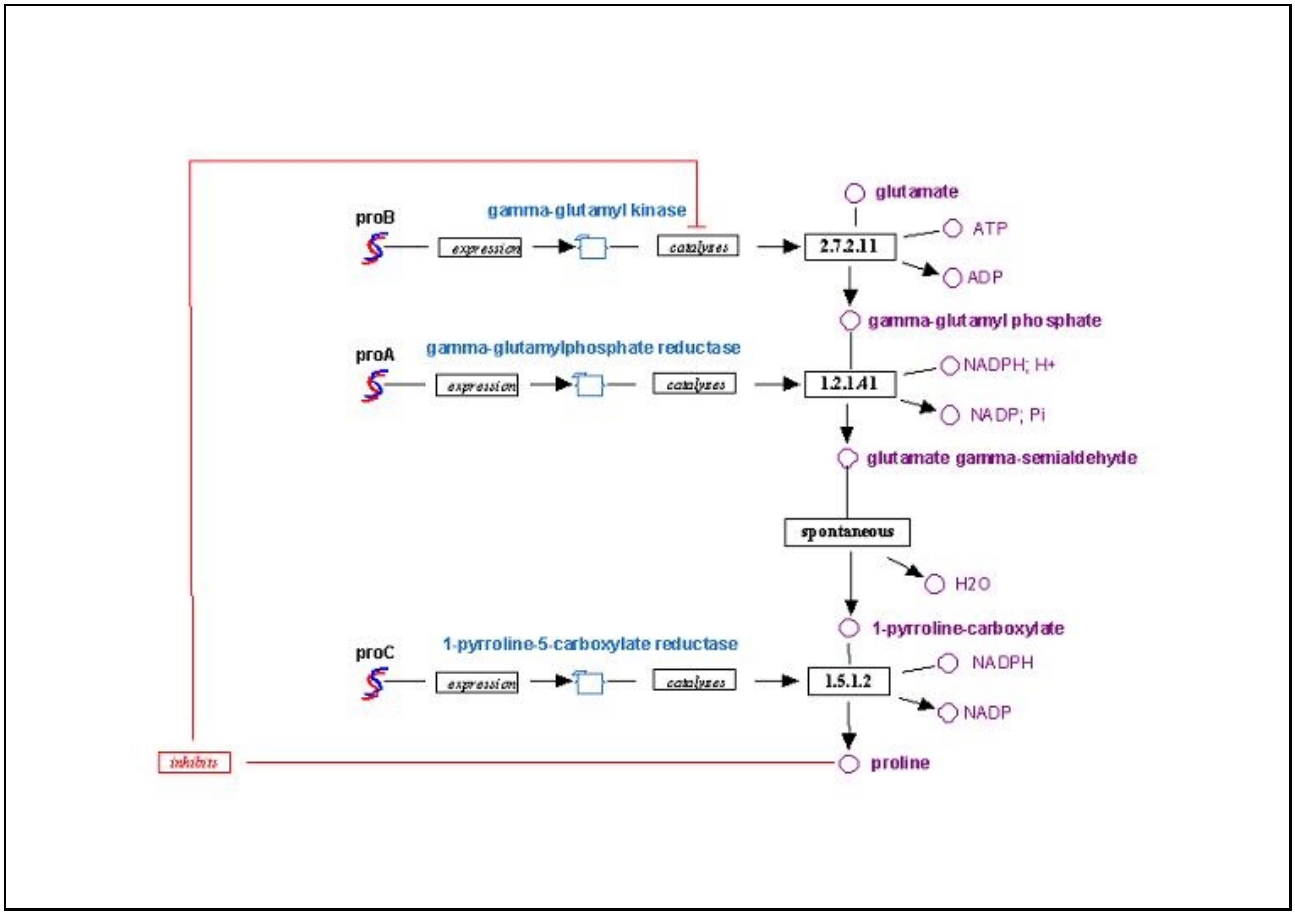


Figure 1.23: Source: [11]. An example of an metabolic pathway: Proline biosynthesis.

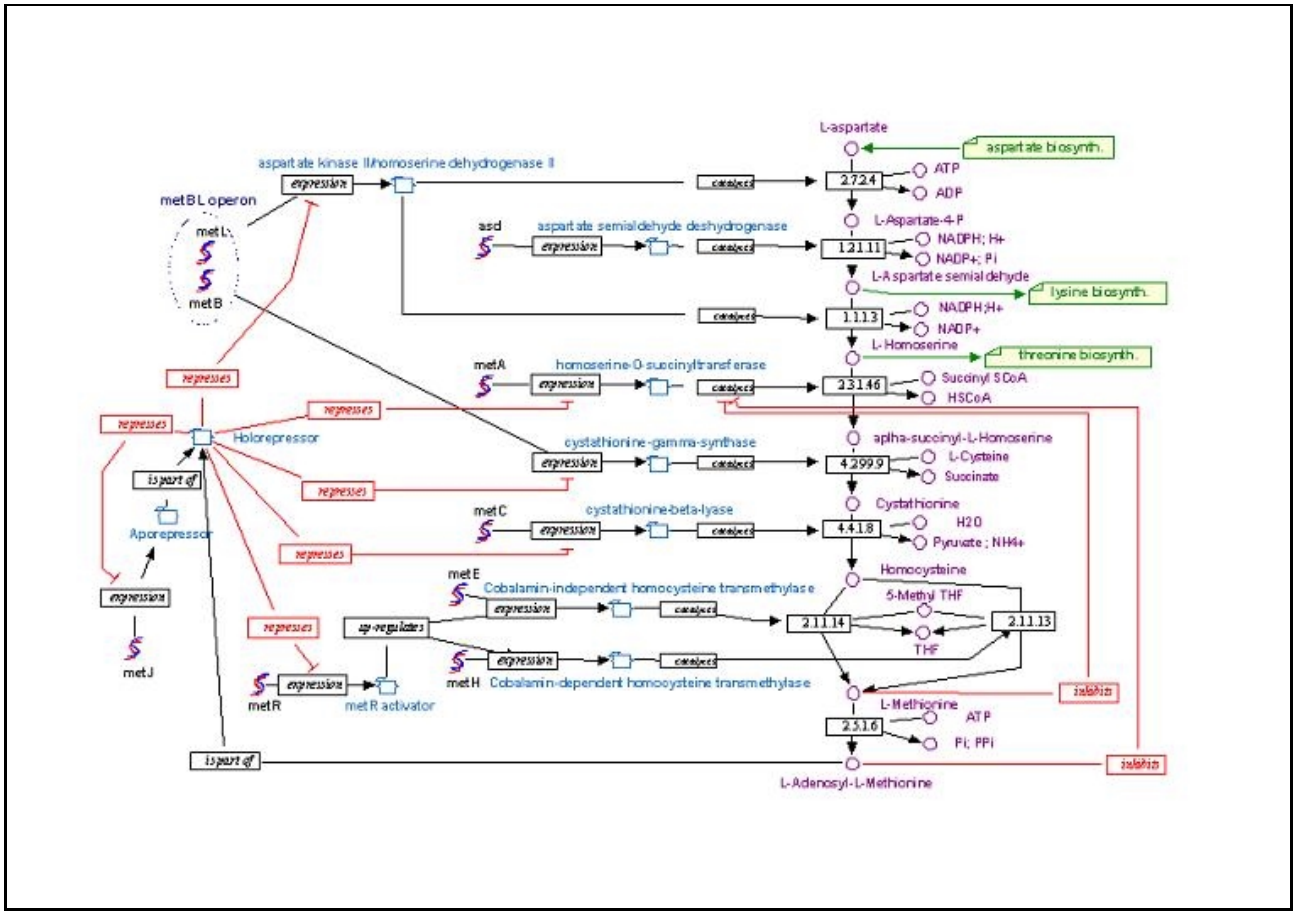


Figure 1.24: Source: [11]. Methionine biosynthesis network in E-coli.

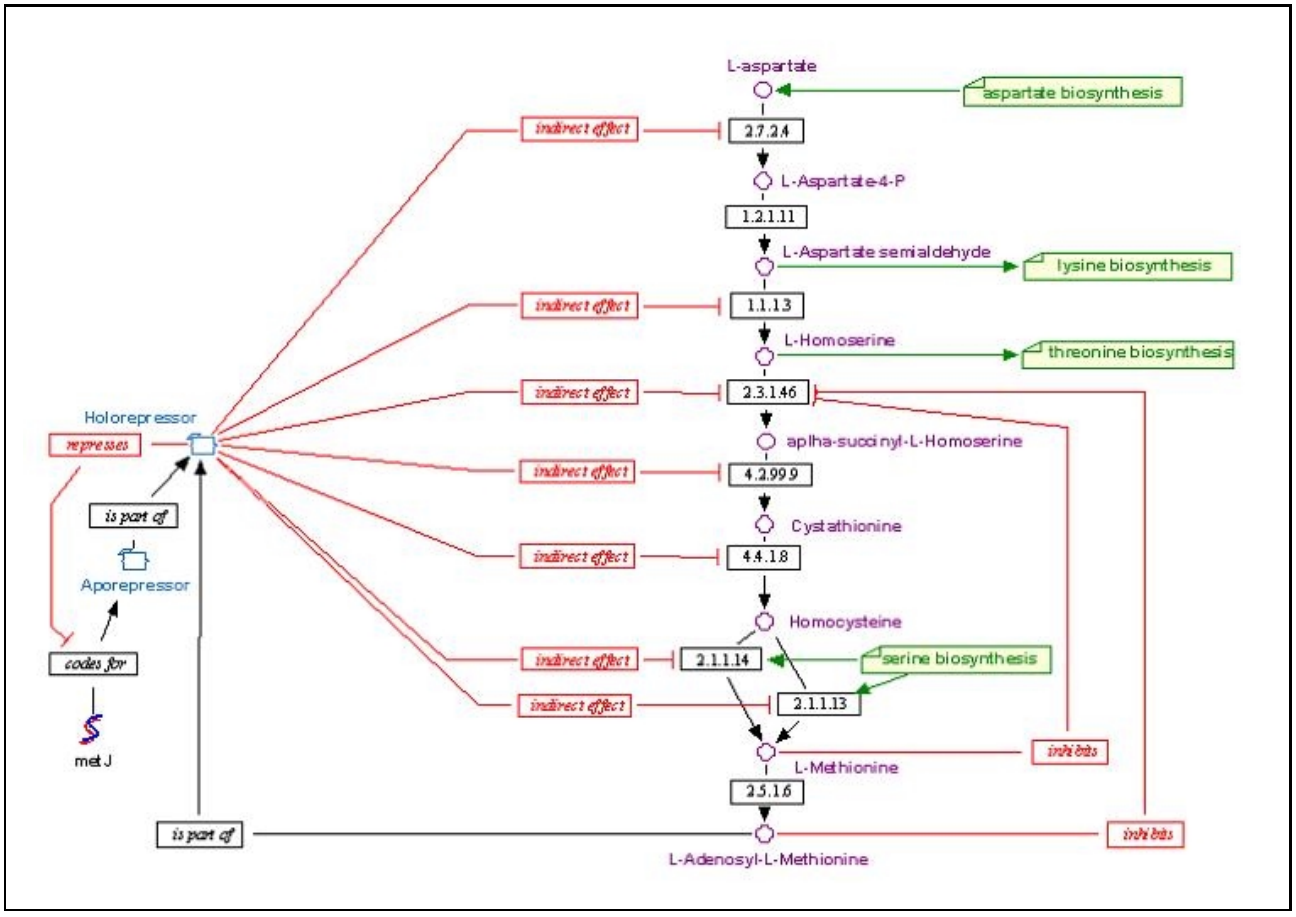


Figure 1.25: Source: [11]. Shortcut representation of the biosynthesis pathway presented in Figure 1.24.

The last example (see Figure 1.26) is that of signal transduction - complex cellular process initiated by signaling protein arrived from outside of a cell. This process eventually affects gene expression in the cytoplasm and inside the nucleus.

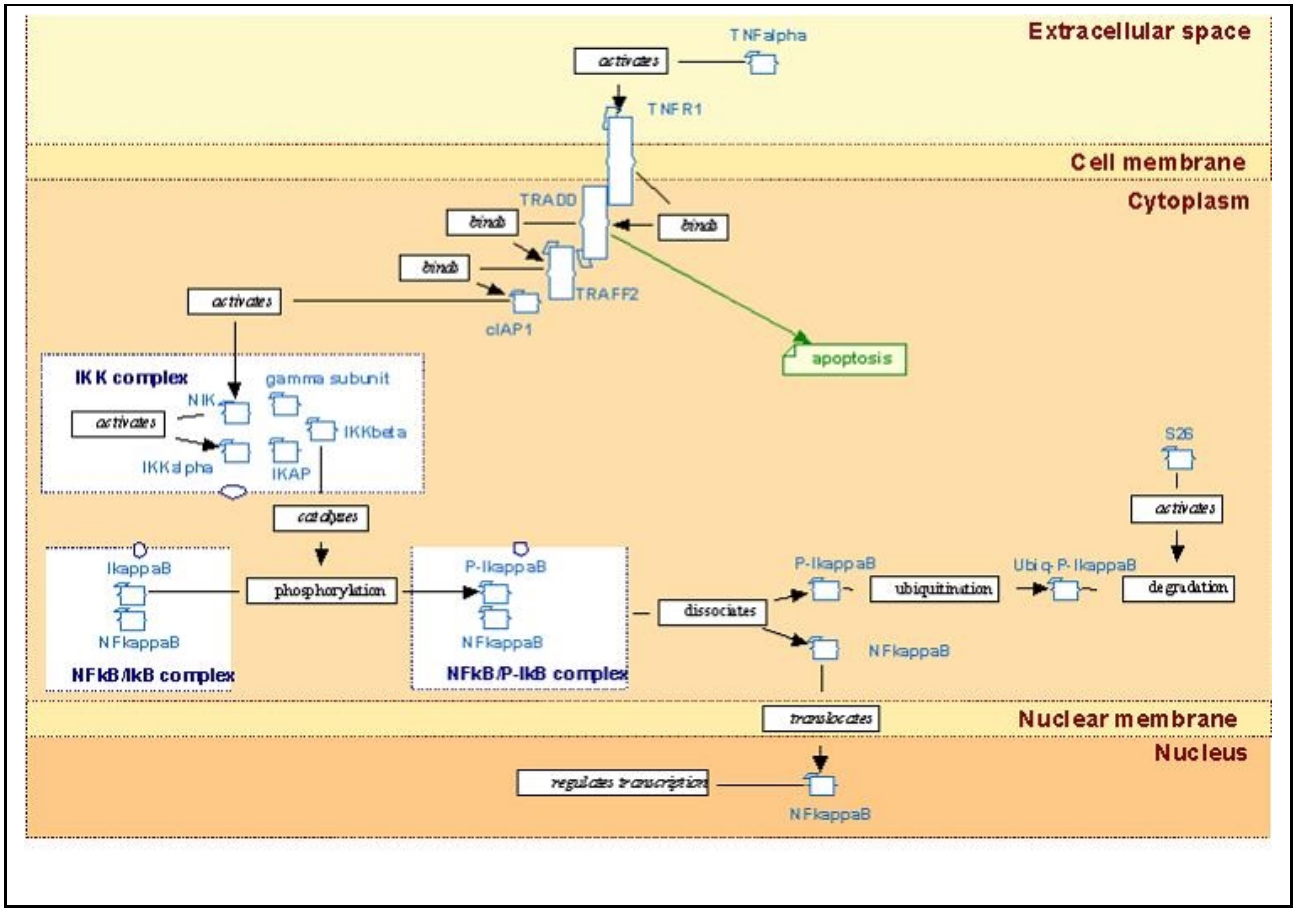


Figure 1.26: Source: [11]. A gene network that performs signal transduction from outside the cell into the nucleus.

1.5.2 Functional Analysis

Using a known structure of such networks it is sometimes possible to describe behavior of cellular processes, reveal their function and the role of specific genes and proteins in them. That's why one of the most important and challenging problems today in molecular biology is that of *functional analysis* - discovering and modeling gene networks from experimental data.

Addressing this problem has been made possible by recent advances in genetic sequencing

and development of a whole new generation of sophisticated biological tools. The most promising technique to date is based on the view of gene systems as a logical network of nodes that influence each other's expression levels. Consequently, one may obtain some information on gene interactions in the network by measurement of gene expression. The level of gene expression, i.e., the production rate of a given protein changes during execution of a genetic process involving it and can be monitored by several biological experiments. A variety of experimental tools have been developed recently with the ability to observe the expression of many genes simultaneously. At the forefront of these technologies lies the DNA microarray, commonly used to monitor gene expression at the level of mRNA abundance. Similarly, the rapid identification of proteins and their abundances is becoming possible through methods such as 2D polyacrylamide gel electrophoresis, 2-hybrid systems, protein chips, etc. The main contribution of all of these technologies is that numerous genes can be monitored in the same experiment, making it possible to perform a global expression analysis of the cell. Additional information about a genetic network may be gleaned experimentally by applying a directed perturbation to the network, and observing expression levels of every gene in the network in the presence of the perturbation. Perturbations may be genetic, in which the expression levels of one or more genes are fixed by knockout (removal of the gene) or over expression (higher than usual level of gene expression), or biological, in which one or more non-genetic factors are altered, such as a change in environment, nutrition, or a temperature increase. Such biological experiments are very costly and very few such perturbations may be performed at one time. Thus, reducing the number and cost of experiments is crucial. Methods presented above supply biological data in terms of expression levels of many genes at different time points and in different conditions. The functional analysis of the data can be defined as a computational problem, aiming to infer some plausible model of the network from the observations with minimal number or cost of biological experiments. The model should describe how the expression level of each gene in the network depends on external stimuli and expression levels of other genes. Additional goals include construction of a knowledge-base of gene regulatory networks, verification of pathways or gene networks hypotheses.

Bibliography

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

[2] A. L. Lehninger. *Biochemistry*. Worth Publishers, Inc., 1975.

[3] http://library.thinkquest.org/C004535/chromosomes.html/.

[4] http://morgan.rutgers.edu/MorganWebFrames/Level1/Page1/p1.html/.

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

[6] http://www.accessexcellence.org/AB/GG/dna_replicating.html/.

[7] http://www.bis.med.jhmi.edu/Dan/DOE/fig5.html/.

[8] http://www.celera.com/.

[9] http://www.cs.utexas.edu/users/s2s/latest/dna1/src/page2.html/.

[10] http://www.dnaftb.org/dnaftb/23/concept/index.html/.

[11] http://www.ebi.ac.uk/research/pfmp/.

[12] http://www.howstuffworks.com/cell4.htm/.

[13] http://www.lsic.ucla.edu/ls3/tutorials/gene_expression.html/.

[14] http://www.nature.com/nature/.

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

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

[17] http://www.pbs.org/wgbh/nova/genome/sequencer.html/.

[18] http://www.people.virginia.edu/~rjh9u/fish6.html/.

[19] http://www.people.virginia.edu/~rjh9u/pcranim.html/.

[20] http://www.sciencemag.org/.

[21] http://www.sciencemag.org/cgi/content/full/291/5507/1195/.