Analysis of Gene Expression Data **Spring Semester, 2002**

Lecture 1: March 4, 2004

Lecturer: Ron Shamir *Scribe: Eran Balan and Maayan Goldstein*[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). The first draft of the human genome was published in February 2001.

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*). DNA allowes duplication. The term *Genome* refers to the totality of DNA material.

Structure

The structure of DNA is described as a *double helix*, which looks rather like two interlocked bedsprings (see Figure 1.3). Each helix is a chain of nucleotides held together by phosphodiester bonds. The two helices are held together by hydrogen bonds. These are considered as weak bonds. Each base pairs consists of one *purine* base (A or G) and one *pyrimidine*

[1] Based on a scribe by Dana Torok and Adar Shtainhart, March 2002.

base (C or T), paired according to the following rule: $G \equiv C, A = T$ (each '-' symbolizes a hydrogen bond).

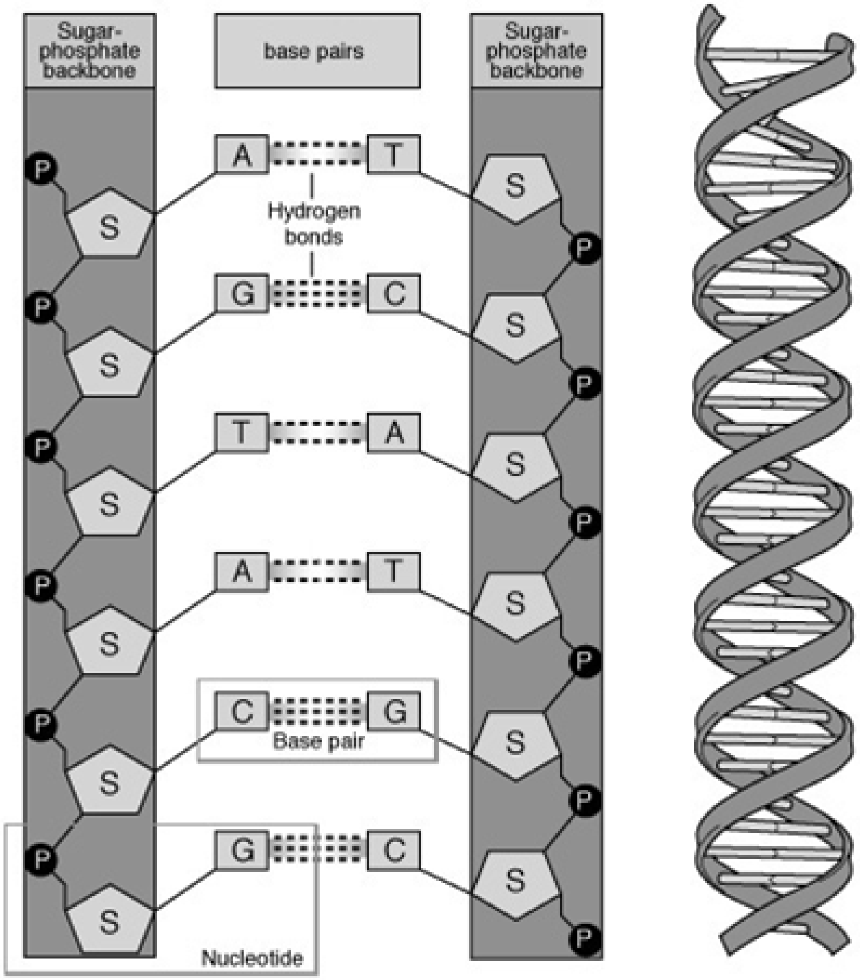


Figure 1.1: Source: [8]. DNA structure.

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 1.4). 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

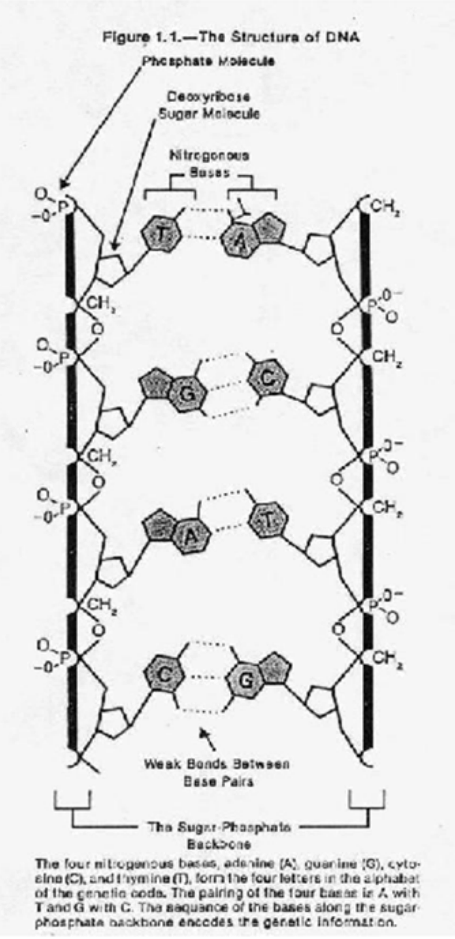


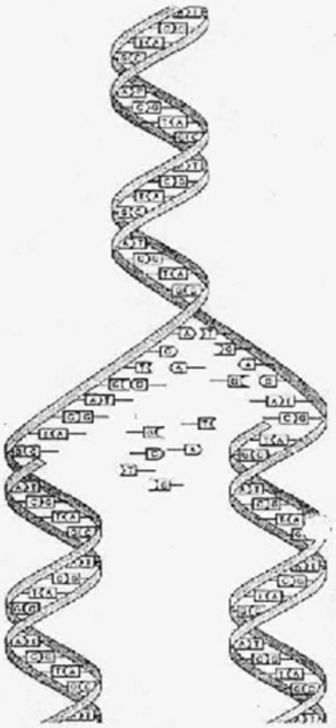
Figure 1.1.—The Structure of DNA

Phosphate Molecule
Deoxyribose Sugar Molecule
Nitrogenous Bases
T
A
G
C
A
T
C
G
O
-O-P
CH₂
O
O-
Weak Bonds Between Base Pairs
The Sugar-Phosphate Backbone

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

Figure 1.2: Source: [5] 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.

Figure 1.2.—Replication of DNA



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

Figure 1.3: Source: [5] DNA Replication.

around an *octamer* of histones molecule 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. Chromosmes are contiguous stretch of DNA. 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*.

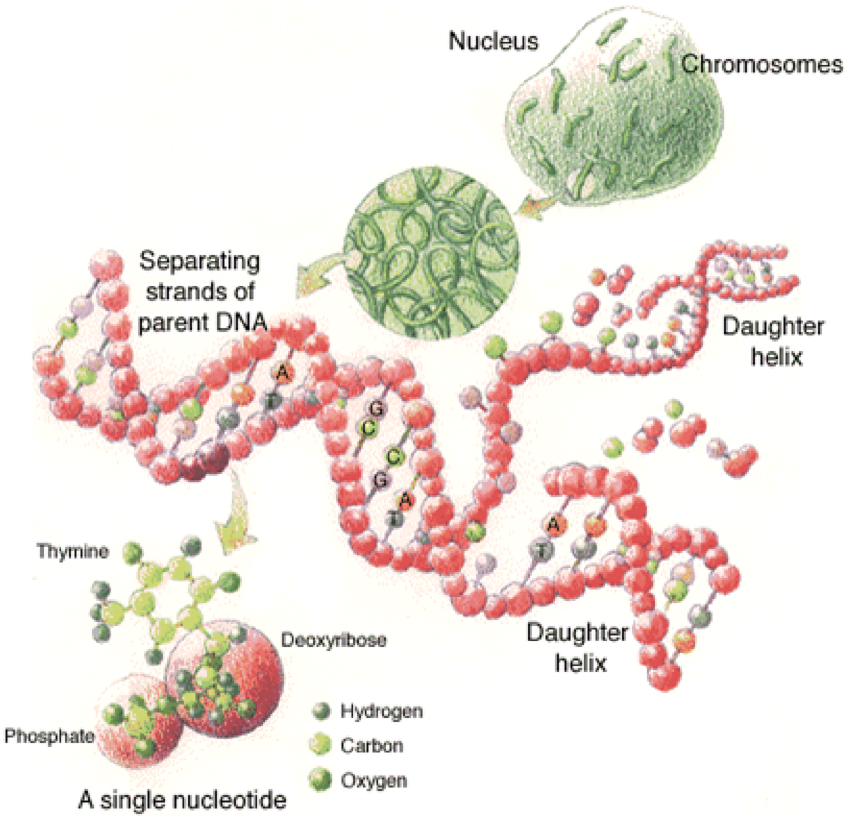
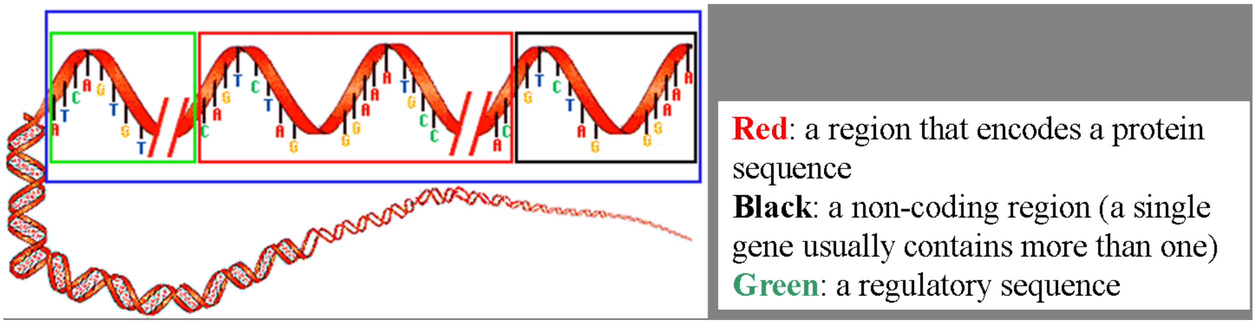


Figure 1.4: Source: [13]. Chromosomes.

1.1.4 Genes

A gene is a segment that specifies the sequence of a protein. It usually corresponds to a single mRNA carrying the information for constructing a protein. It contains one or more regulatory sequences that either increase or decrease the rate of its transcription (see Figure 1.5). In 1977 molecular biologists discovered that most Eukaryotic genes have their coding sequences, called *exons*, interrupted by non-coding sequences called *introns* (see Figure 1.6). 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. Every protein has a limited life time, that's why the genes are needed to produce new proteins in a supervised way.



Red: a region that encodes a protein sequence
Black: a non-coding region (a single gene usually contains more than one)
Green: a regulatory sequence

Figure 1.5: Source: [3]. Gene structure.

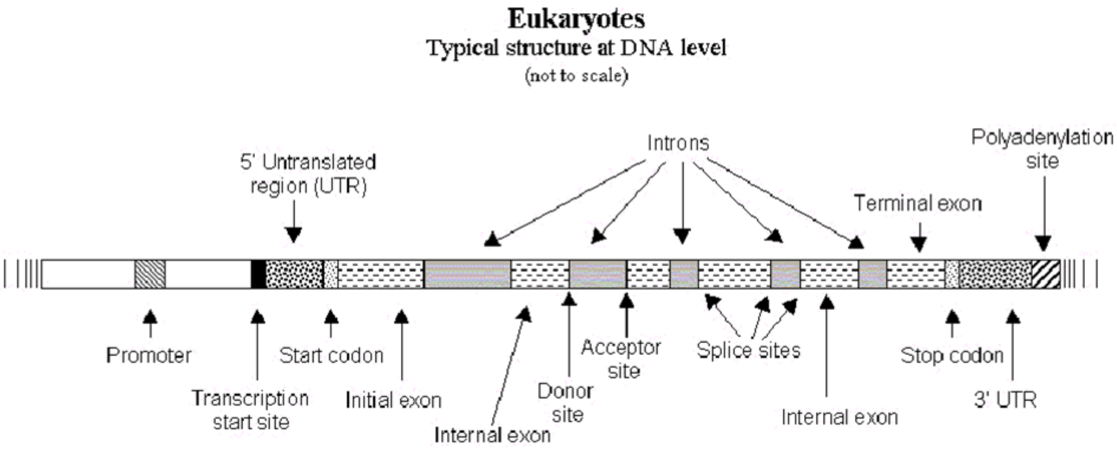


Figure 1.6: Gene structure in Eukaryotes.

1.1.5 From Gene to Protein

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 1.4 , 1.7 and 1.8). When genes are expressed, the genetic information (base sequence) on DNA is first transcribed (copied) to a molecule of messenger RNA in a process similar to DNA replication. This process called *transcription* is catalyzed by the enzyme *RNA polymerase* (see [1], pages 107, 200-202)(see Figure 1.9). 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 deoxyribose. 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.(see Figure 1.3) The mRNA molecules then leave the cell nucleus and enter the cytoplasm, where triplets of bases (codons) forming the genetic code specify the particular amino acids that make up an individual protein. This process, called translation, is accomplished by ribosomes (cellular components composed of proteins and another class of RNA) that read the genetic code from the mRNA, and transfer RNAs (tRNAs) that transport amino acids to the ribosomes for attachment to the growing protein.

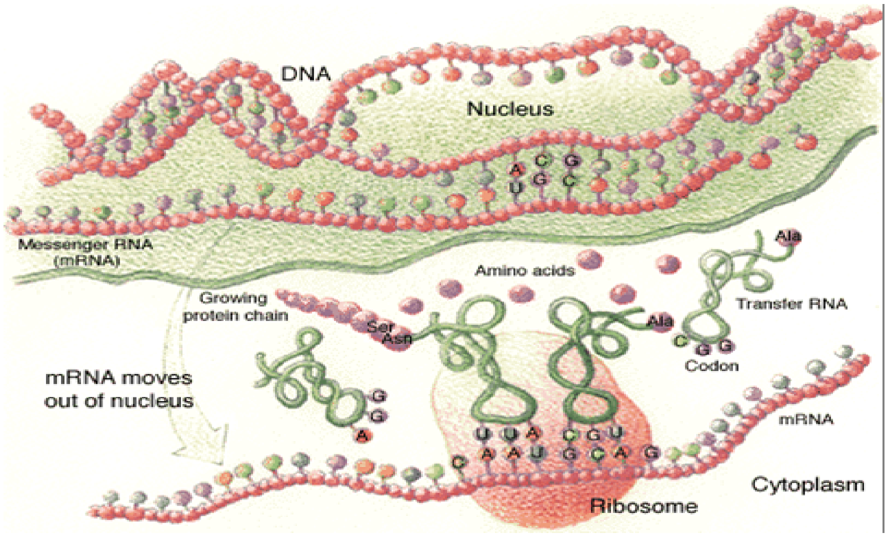


Figure 1.7: Source: [14]. From gene to protein.

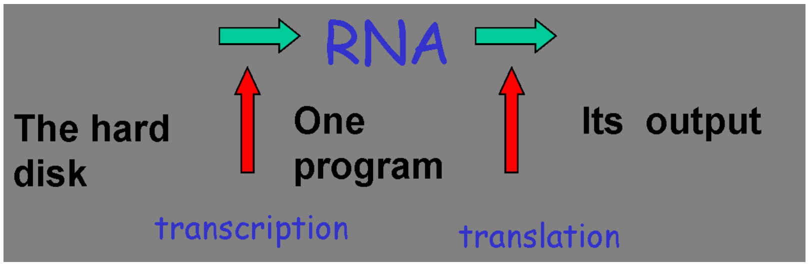


Figure 1.8: From DNA to Protein.

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).

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 t(see Figure 1.3)hat 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(see Figure 1.3) evolution: with a few minor exceptions, it is the same in organisms as diverse as bacteria, plants, and humans.

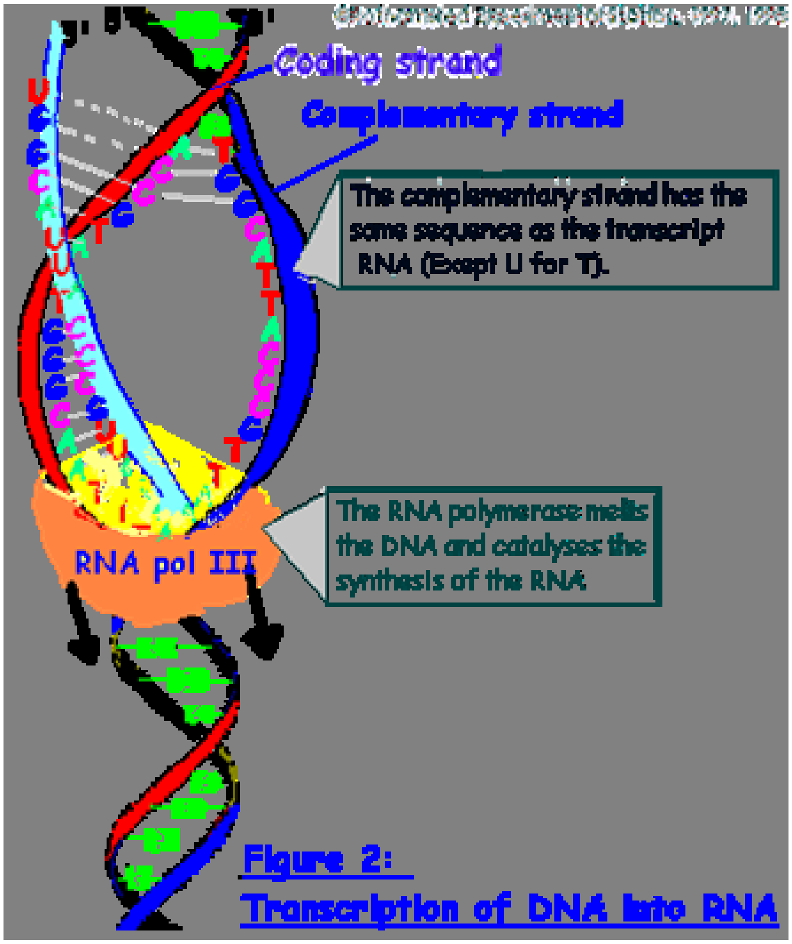
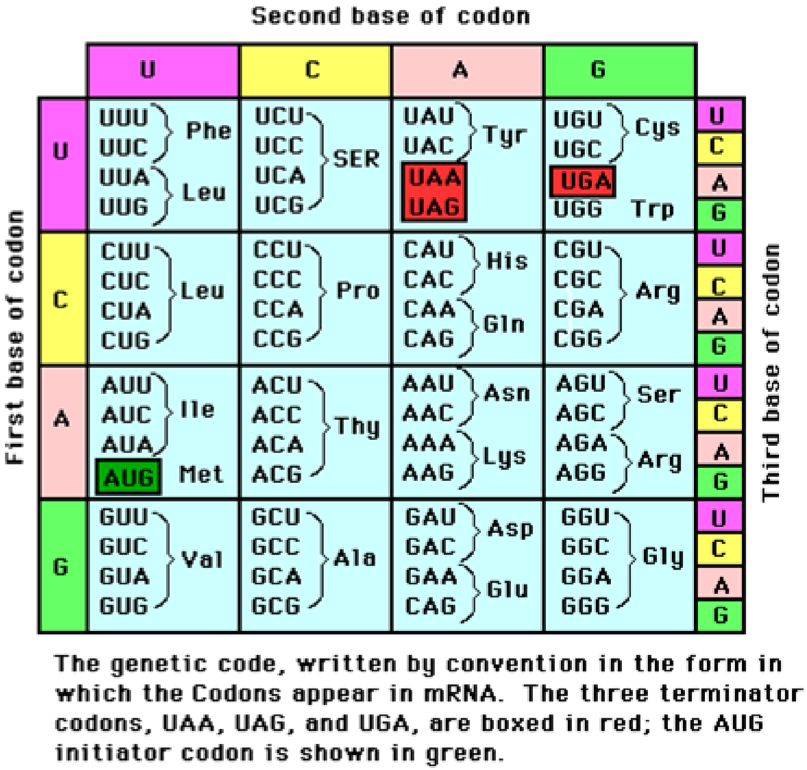

3'
5'
5'
3'
Coding strand
Complementary strand
The complementary strand has the same sequence as the transcript RNA (Exept U for T).
RNA pol III
The RNA polymerase melts the DNA and catalyses the synthesis of the RNA
Figure 2:
Transcription of DNA into RNA

Figure 1.9: Source: [10]. Transcription of DNA into RNA.



Second base of codon

First base of codon	U		C		A		G		Third base of codon
U	UUU	Phe	UCU	SER	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA		UGA		A
	UUG		UCG		UAG		UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thy	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		CAG		GGG		G

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 1.10: Source: [4]. The genetic code table.

Splicing

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. 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 1.11). 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. After the splicing the mRNA is referenced as mature mRNA whereas before the splicing it is referenced as pre-mRNA. (see Figure 1.12)

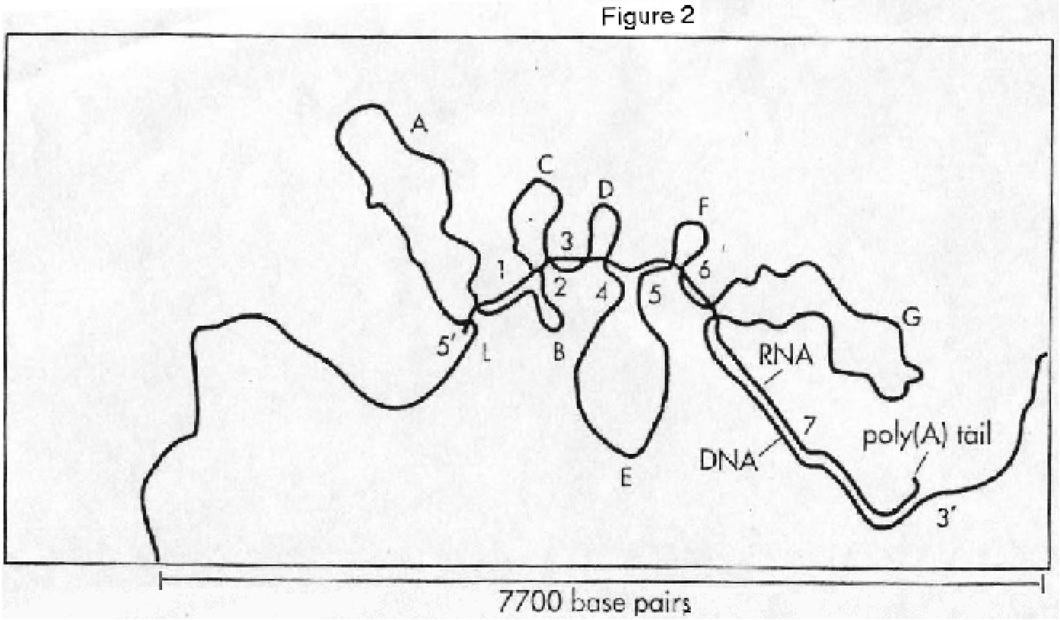


Figure 1.11: Introns are spliced out to form the mature mRNA.

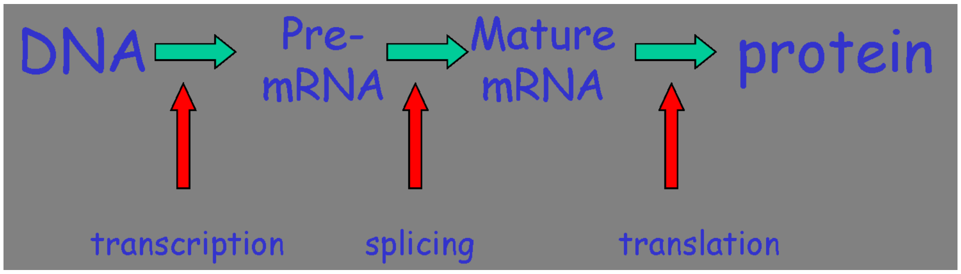


Figure 1.12:

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. 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 (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.

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 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

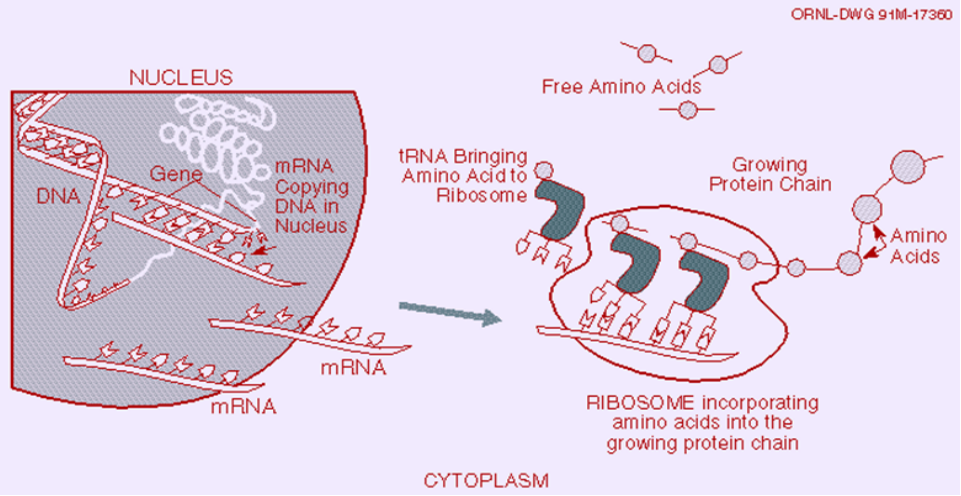

ORNL-DWG 91M-17360
NUCLEUS
Free Amino Acids
DNA
Gene
mRNA Copying DNA in Nucleus
tRNA Bringing Amino Acid to Ribosome
Growing Protein Chain
Amino Acids
mRNA
mRNA
RIBOSOME incorporating amino acids into the growing protein chain
CYTOPLASM

Figure 1.13: Source: [6]. Transcrption of DNA into RNA.

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

1.3 The Human Genome

Following are some statistics on the human genome:

- 23 pairs of chromosomes comprise the human genome.
- The human genome contains 3.2 billion nucleotide bases.
- Gene length: 1000-3000 bases, spanning 30-40,000 bases (because of intorns segments).
- The total number of genes is estimated at 25,000, 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.

- There is a small difference between the genome of two people (about 0.1%).

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 US budget for the project was 3 billion dollars. The project had the following goals:

- Create detailed maps of all chromosomes - to produce a single continuous sequence for each of the 24 human chromosomes.

- Form a dense set of markers, by delineating the positions of all genes, to help in gene and disease hunting. A small portion of each cDNA (the complementary DNA) sequence is all that is needed to develop unique gene markers.

- Obtain a complete (3,200,000,000 bases long) genome sequence.

- The Human Genome Project is expected to produce a sequence of DNA representing the functional blueprint and evolutionary history of the human species, and much more.

- To help discover the genome sequence in more primitive creatures in order to develop, in the future, technologies in DNA manipulations, which are much simpler than working with human DNA.

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”.

- 1997 - only 6.5% of the genome had been mapped.

- 1998 - Commercial players promise quick and dirty genome sequence by 2002. Craig Ventner forms a company with the intent to sequence the human genome within three years. The company, later named *Celera* (see [7]), introduced a new ambitious 'whole genome shotgun' approach. The idea was to map only the genes and not all the genome.
- 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 - Drafts of the human genome, public and private, were published in Nature and Science magazines (see [11] and [15]).
- 1-2 more years for real completion.
- April 2003 - The HG project was announced to be completed, when more than 99% of the human genome was sequenced, assembled into long pieces and reviewed (see Figure 1.14).

For a more detailed timetable of the Human Genome Project see [16].

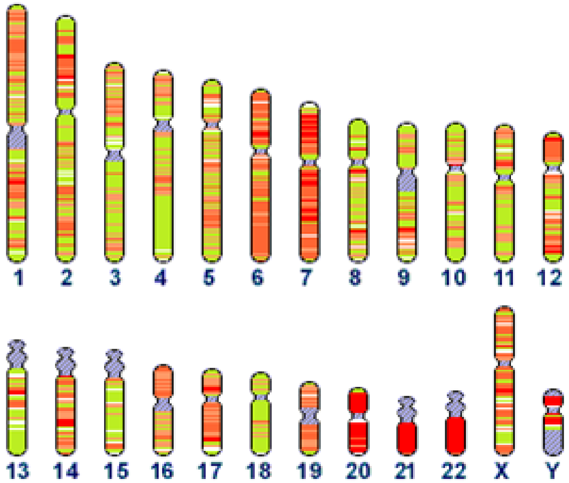


Figure 1.14: Source: [12]. The Public Human genome sequencing progress as of 31/12/2001, scaled from green = draft to red = finished, 63% finished, 34.8% draft, 97.8% total.

1.3.2 After the HGP, the Next Steps

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.

- Functional Genomics: seek to find out function of genes/proteins. 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.
- Identify individual differences in sequences. This could be the base for finding genetic diseases that appear in human beings.
- Genome-wide high throughput technologies:
  - Global gene expression profiling (Transcriptome).
  - Wide-scale protein profiling (Proteome).
- Understand gene regulation - how proteins production is controlled.
- Figure out how genes and proteins interact: gene networks, development.
- Paradigm shift: Reductionist (Hypothesis driven) to Holistic (exploratory, Hypothesis generating).More on this will be elaborated in ”Functional Genomics”.

1.3.3 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. Instead, a new paradigm, holistic and high throughput is emerging. 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 study of behavior of co-regulated gene networks.

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 DNA Chips and Microarrays

1.4.1 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 (see Figure 1.15). 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.

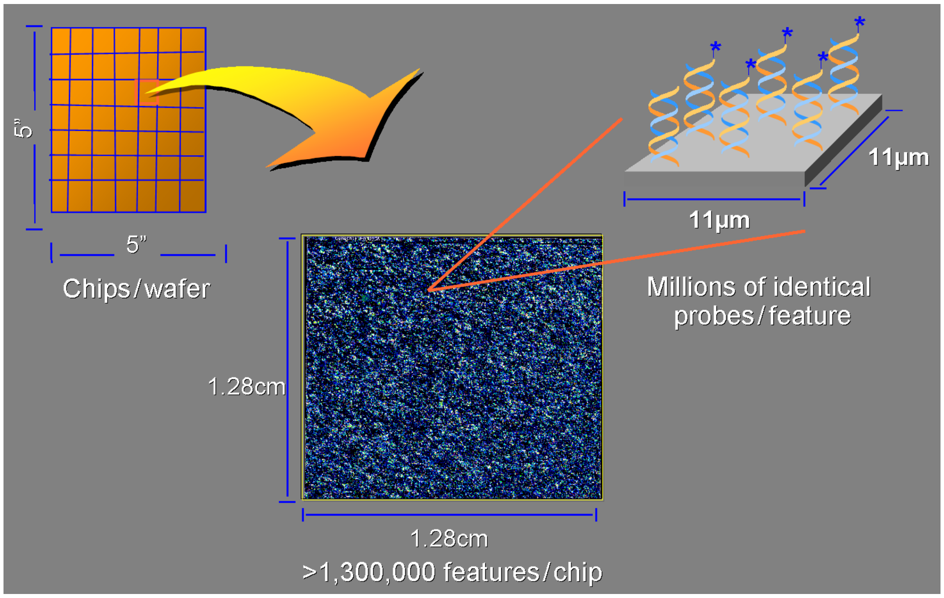


Figure 1.15: Wafers, Chips, and Features.

1.4.2 Technologies

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 they are usually no longer than 25 bases. Since the oligos are short, the density of these chips is very high, for instance, a chip with size 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 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. Figures 1.16 and 1.17 demonstrate the production process.

The GeneChip is used to detect a sequence of DNA based on comparing the ratio between the prefect match and the mismatch oligos fluorescented images intensity as demonstrated in figure 1.18. In addition, for each complementary, or perfect match oligo that is synthesized, a second mismatch oligo is also synthesized. There is a mismatch at the central (13th) position. This acts as an control for hybridization efficiency for each oligo. Analysis of perfect match/mismatch pairs allows low-intensity hybridization.

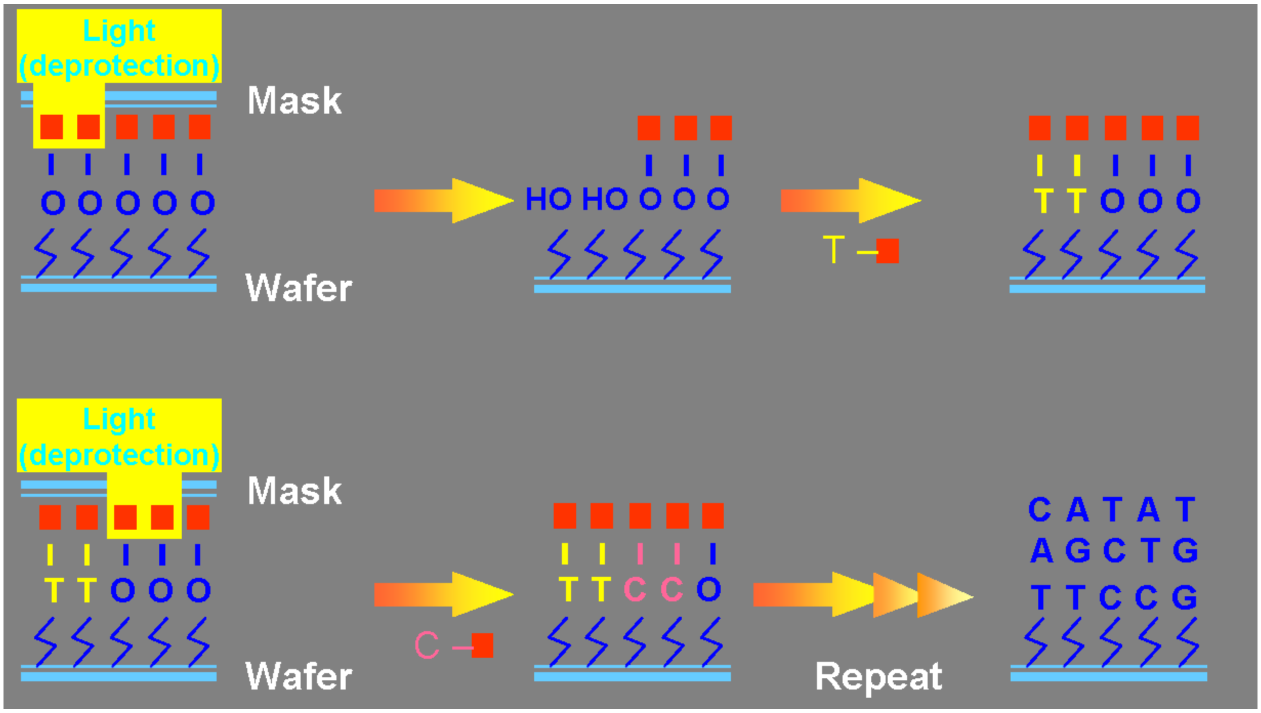


Figure 1.16: Manufacturing DNA chips. 1-2) The light removes the terminator from the chains not covered by the mask, creating hydrogen bonds instead. 3) Bonds are formed with a nucleotide base. 4-6) The process is repeated with a different base.

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 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

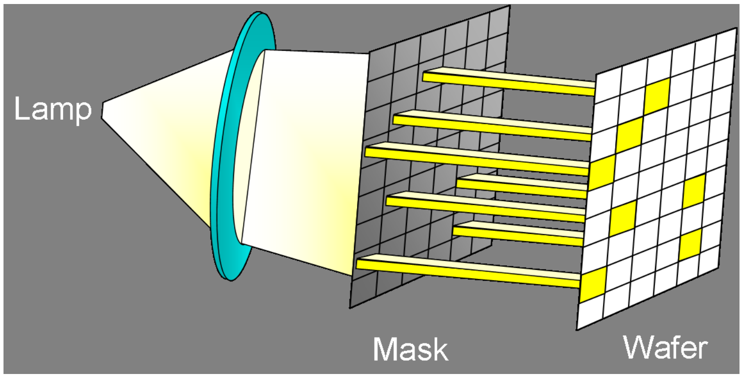
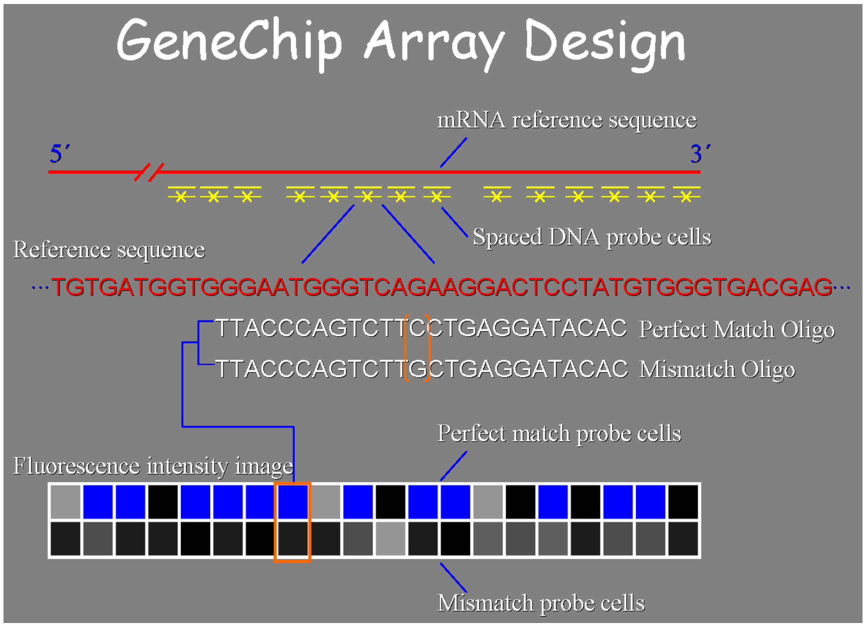


Figure 1.17: GeneChip Manufacturing Process. A typical experiment with an oligonucleotide chip. Labeled RNA molecules are applied to the probes on the chip, creating a fluorescent spot where hybridization has occurred.



GeneChip Array Design

mRNA reference sequence

5′ 3′

Spaced DNA probe cells

Reference sequence

···TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAG···

TTACCCAGTCTTCCTGAGGATACAC Perfect Match Oligo

TTACCCAGTCTTGCTGAGGATACAC Mismatch Oligo

Perfect match probe cells

Fluorescence intensity image

Mismatch probe cells

Figure 1.18: GeneChip Array Design

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.(see Figure 1.19)

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. Each row represents genes expression pattern or fingerprint vector. Each column represents experiment/conditions profile. Entries of the Raw Data matrix are ratio values, absolute values, distributions.

1.4.3 Biological Application

Monitoring Gene Expression

The goal is to simultaneously measure expression levels of all genes in one experiment. The monitoring is based on two fundamental biological assumptions: transcription level that indicates genes' regulation. By getting information on the mRNA quantity we can evaluate the control level on certain genes. Only genes which contribute to organism fitness are expressed in a particular condition. We assume that living organism won't produce unnecessary proteins, but rather those needed for its existence. Detecting changes in gene expression level provides clues on its product function.

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, ...) (see Figure 1.21). 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

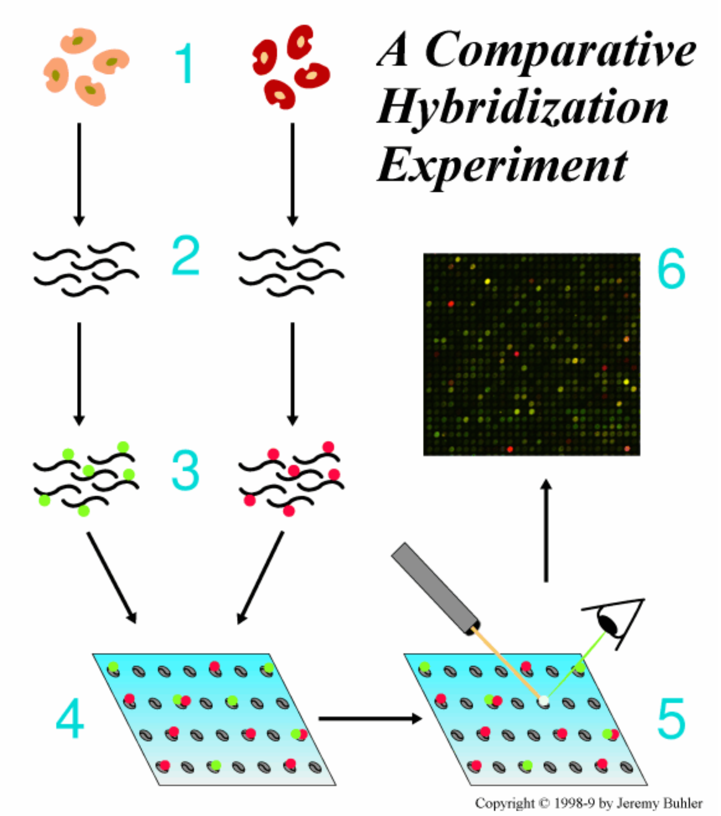

A Comparative Hybridization Experiment
1
2
3
4
5
6
Copyright © 1998-9 by Jeremy Buhler

Figure 1.19: 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 deduced. 6) cDNA chip after scanning.

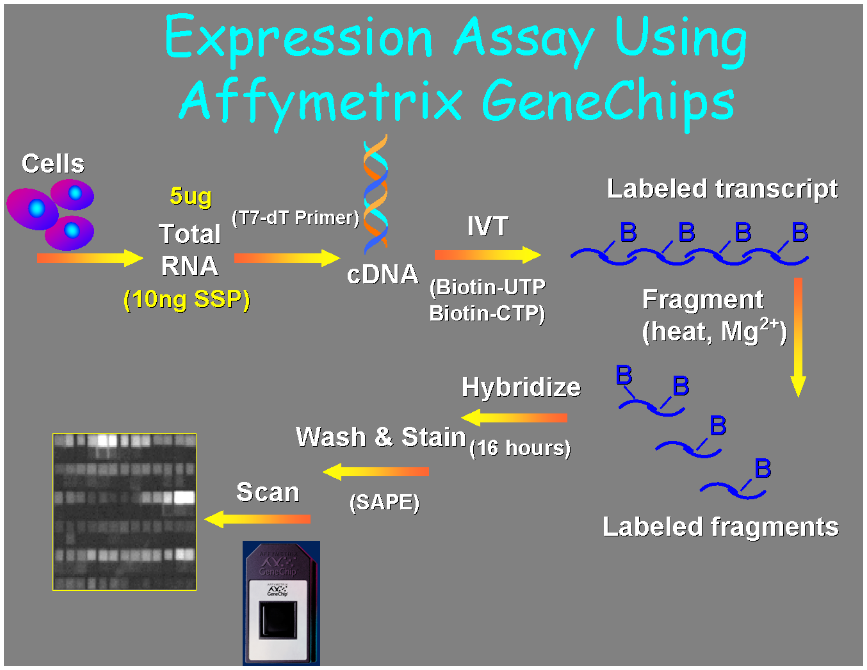
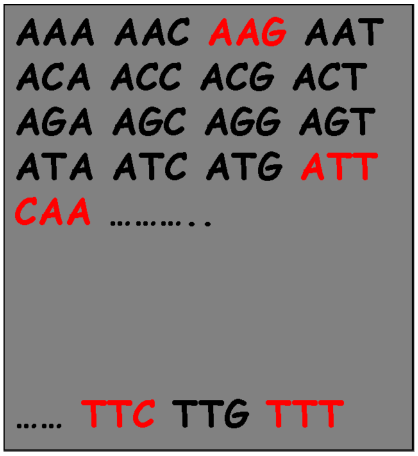


Figure 1.20: Expression Assay Using Affymetrix GeneChips

hybridization the spots with sequences that 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 a long 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.

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



AAA AAC AAG AAT
ACA ACC ACG ACT
AGA AGC AGG AGT
ATA ATC ATG ATT
CAA ………..

…… TTC TTG TTT

Figure 1.21: 3-mer oligos.

cluster with some biological semantics.

There are many other computational challenges. One of them is, given partition of the conditions into types, classify the types of new conditions and find a subset of the genes for each type that distinguishes it from the rest. These partitions could be also used to specify a clustering of genes that manifest similar expression pattern. Another challenge appears while designing experiments. One should 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. And finally, to assign statistical significance to the answer of the experiment.

1.5 Molecular Networks

A molecular 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.

1.5.1 Examples

The 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.

There are three types of molecular networks. An example for the first type, called Metabolic Pathways can be seen in Figure 1.22. The Metabolic Pathway involves a chain of generated proteins. 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.

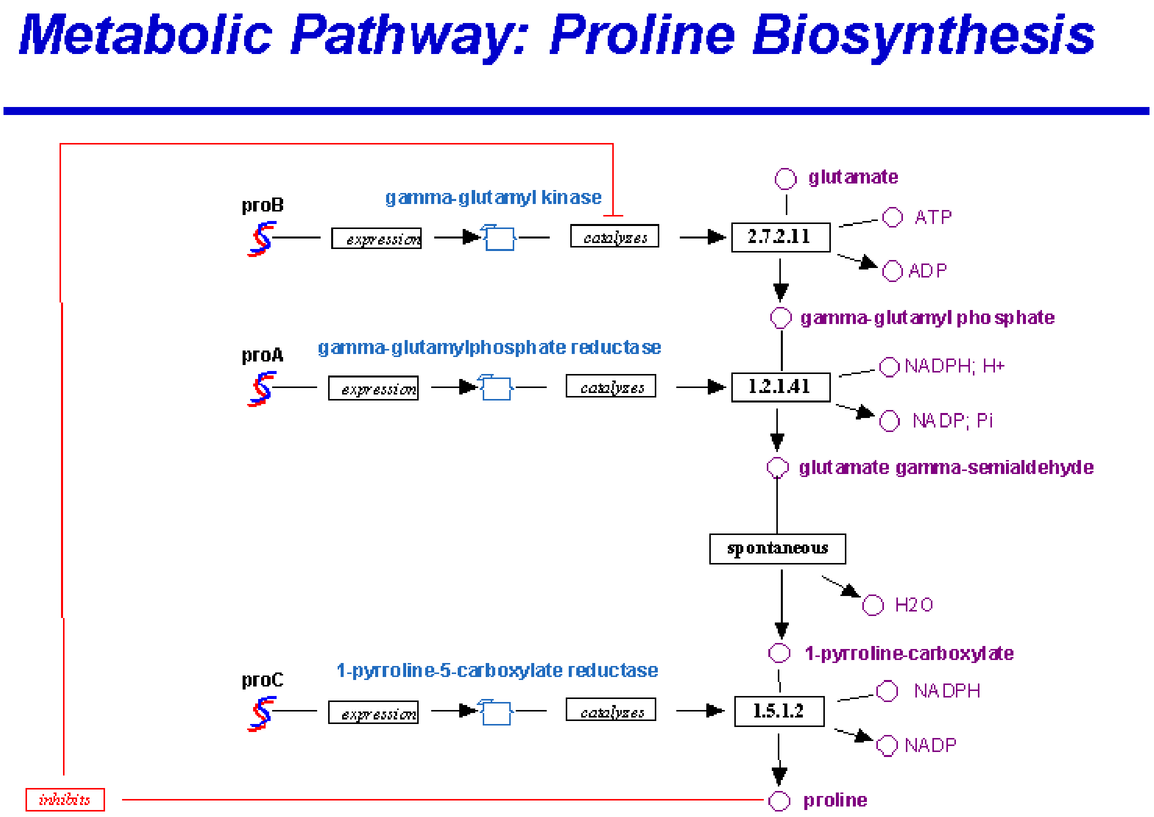


Figure 1.22: Source: [9]. An example of an metabolic pathway: Proline biosynthesis.the role of gene expression in catalyzing chemical reactions.

The following two figures (1.23 and 1.24) 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.

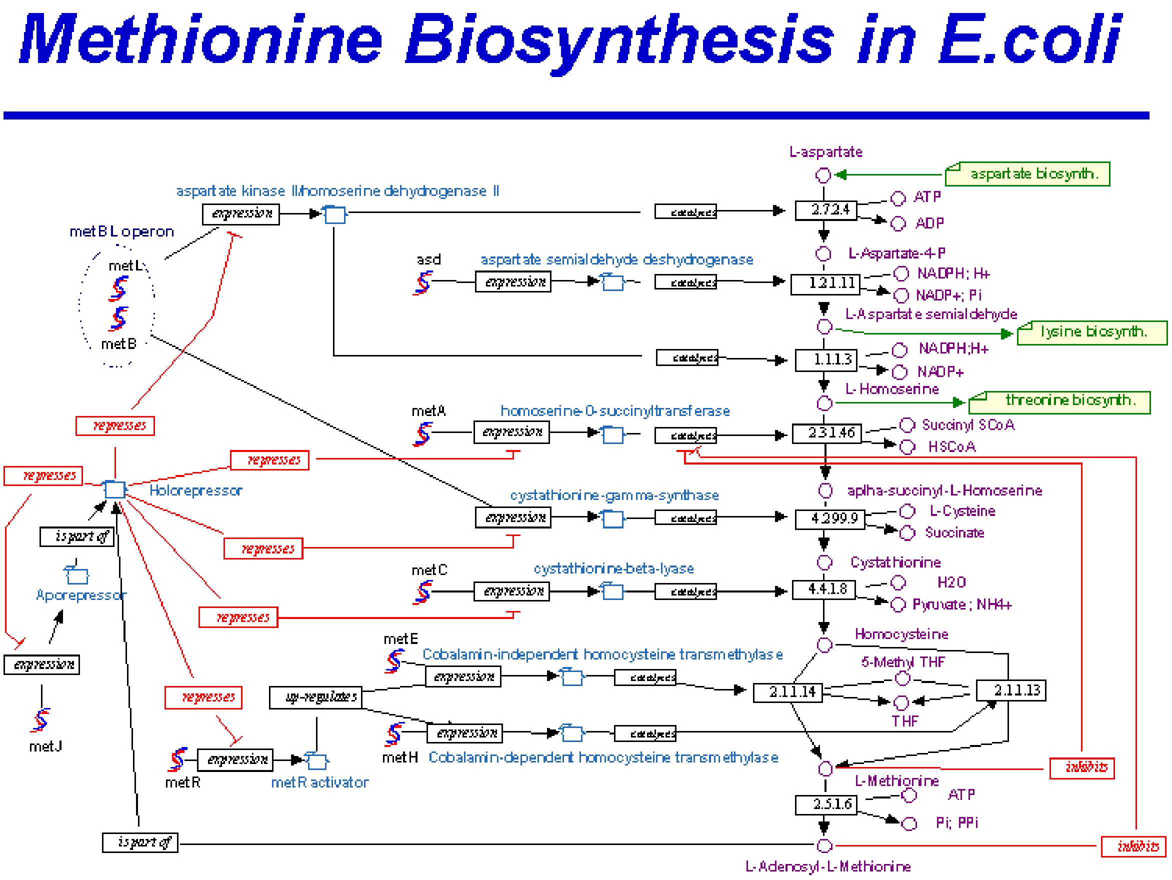


Figure 1.23: Source: [9]. Methionine biosynthesis network in E-coli.

The last example (see Figure 1.25) 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. This network is an example of second network type called Protein Network.

There's another type of network that is called Transcriptional Network that sometimes appears as a part of a bigger network that is classified as one of the other two types.

The examples shown above present the goal of reverse engineering of generic networks. This produces two main challenges: exploitable of the emerging vast, heterogeneous data sources and development of computational tools that are robust, realistic and rigorous.

Shortcut Representation

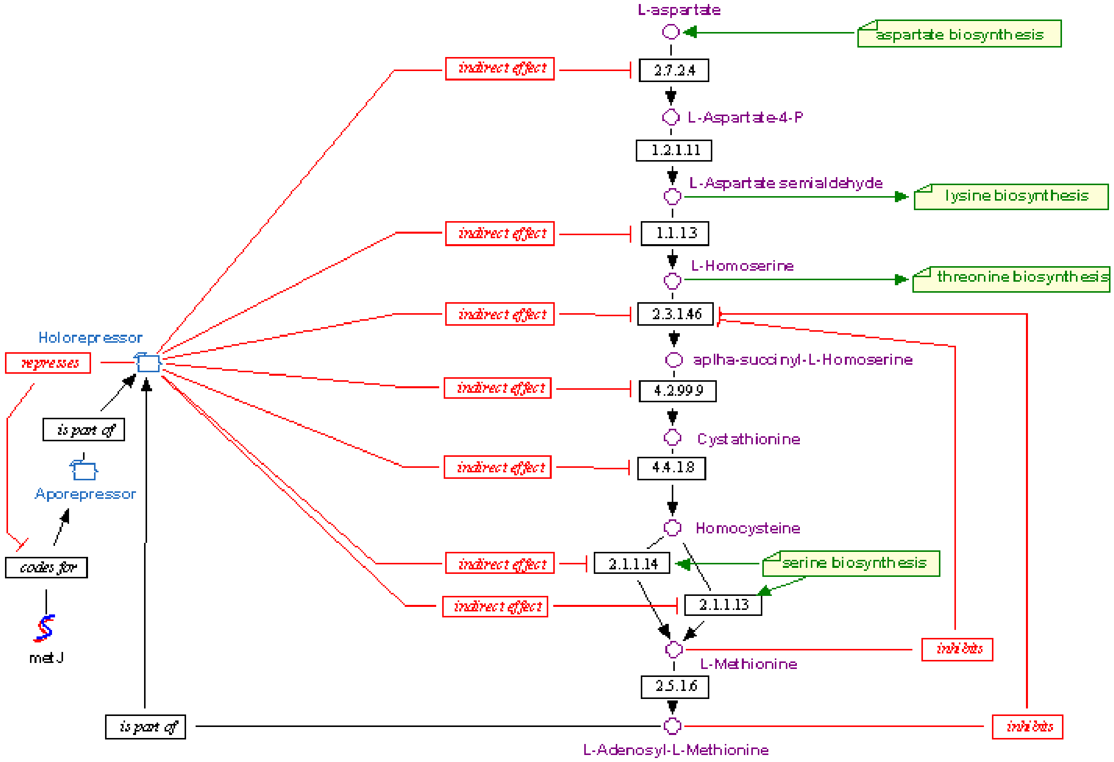
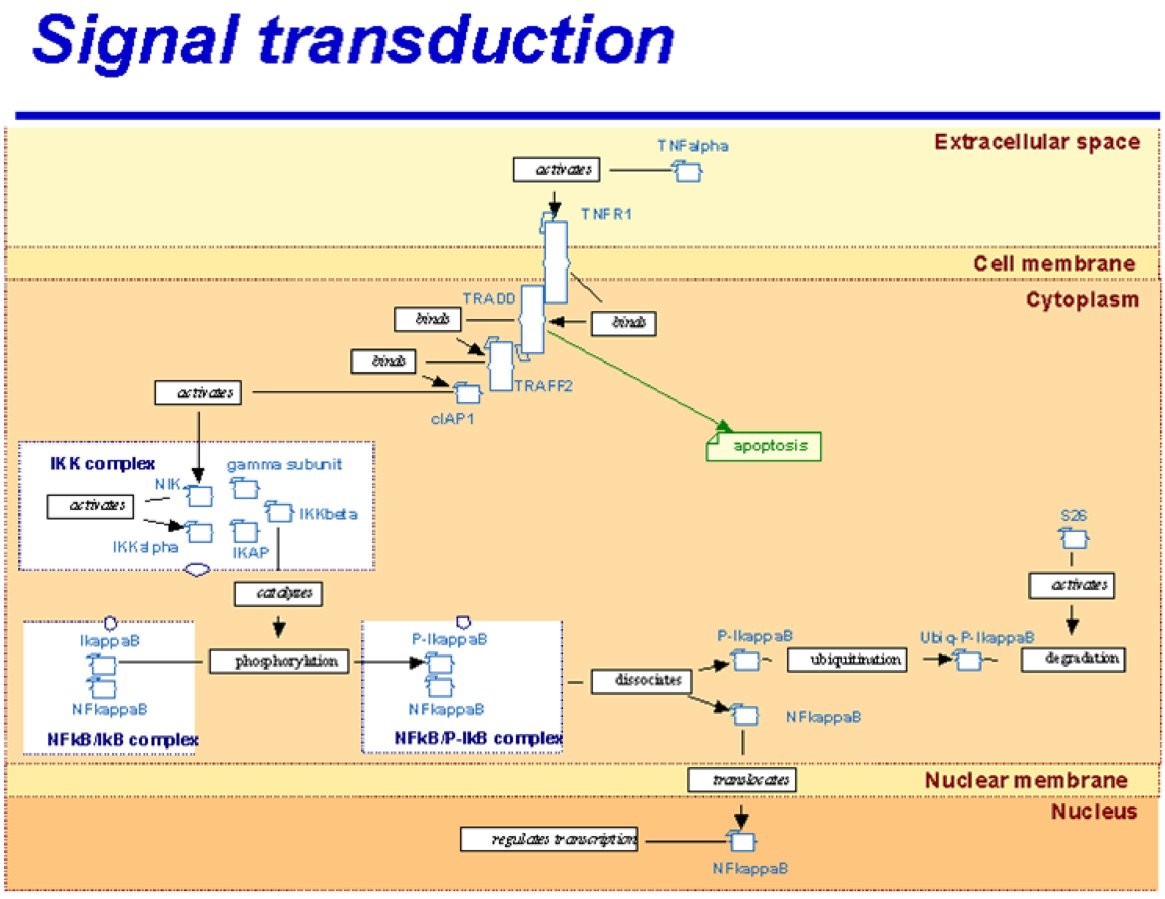


Figure 1.24: Source: [9]. Shortcut representation of the biosynthesis pathway presented in Figure 1.23.



Signal transduction

Figure 1.25: Source: [9]. 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. We can also use this data to define algorithms to study these models.

Furthermore, functional analysis strives to optimize experiments to verify and reconstruct network. And vice versa, integration between various types of networks allows to achieve better experimental result.

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://morgan.rutgers.edu/MorganWebFrames/Level1/Page1/p1.html/`.

[4] `http://ntri.tamuk.edu/cell/ribosome.html/`.

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

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

[7] `http://www.celera.com/`.

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

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

[10] `http://www.iacr.bbsrc.ac.uk/notebook/courses/guide/words/trans.html/`.

[11] `http://www.nature.com/nature/`.

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

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

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

[15] `http://www.sciencemag.org/`.

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