Analysis of Gene Expression Data **Spring Semester, 2005**

Lecture 1: February 25, 2005

Lecturer: Ron Shamir *Scribe: Karin Inbar and Anat Lev-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 revealed the basic inheritance mathematics (information that is carried between generation). 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. The basic biological units responsible for possession and passing on of a single characteristic are called *genes*. 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 335-337). In Febuary 2001, the first draft of the human genome was published (see [3]).

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 Figure 1.1(a)).

Structure

The structure of DNA is described as a *double helix*, which looks rather like two interlocked bedsprings (see Figure 1.1(a,c)). Each helix is a chain of nucleotides held together by phospho-diester bonds, that are considered as strong bonds. The two helices are held together by hydrogen bonds. These are considered as weak bonds, so the strands might be separated. Each base pairs (see Figure 1.1(b)) consists of one *purine* base (A or G) and

[1]Based on a scribe by Eran Balan and Maayan Goldstein, March 2004 and on a scribe by Dana Torok and Adar Shtainhart, March 2002.

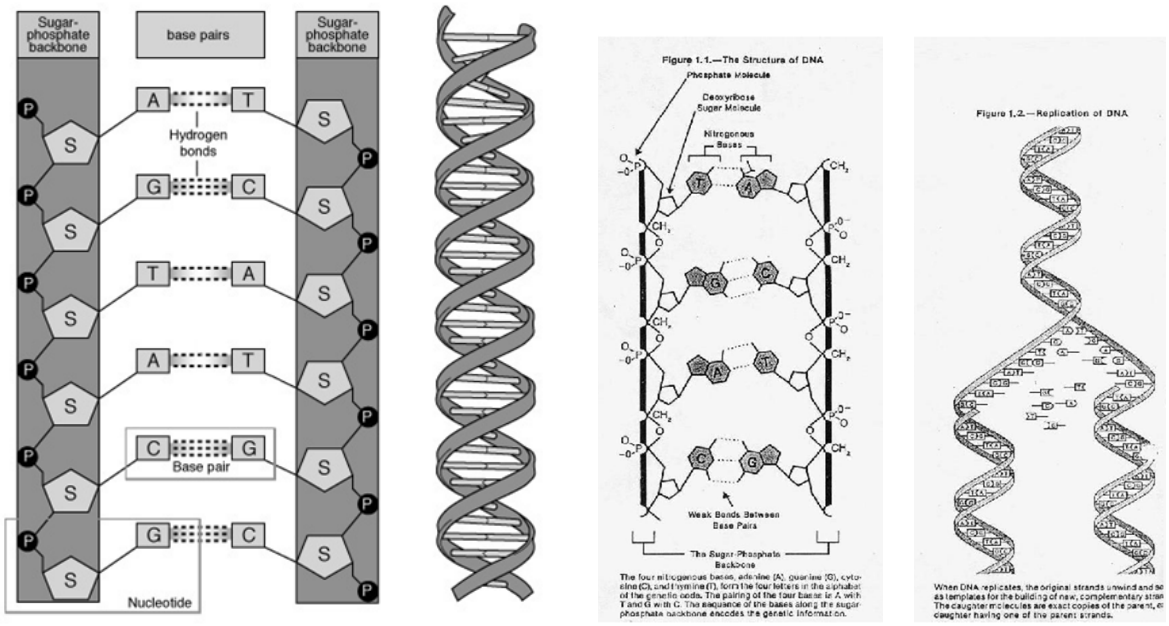

Figure1.1a Figure1.1b Figure1.1c

Figure 1.1: a: DNA structure (Source: [8]); b: 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 (Source: [5]); c: DNA Replication (Source: [5])

one *pyrimidine* base (C or T), paired according to the following rule: $G \equiv C, A = T$(each '-' symbolizes a hydrogen bond), actually defining complimentary strands. As each strand uniquely defines its complementary strand, it carries whole genetic information. The direction of the strand has a meaning, and the information is always read in the same direction. The length of human DNA is about 3×10^9 base pairs of nucleotides (abbreviated *bp*). DNA allows duplication. The term *Genome* refers to the totality of DNA material.

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 1.2). 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 histones, a form of DNA binding proteins. Each unit of DNA wrapped 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*. In human the length of a single chromosome might be 100-150 million bp.

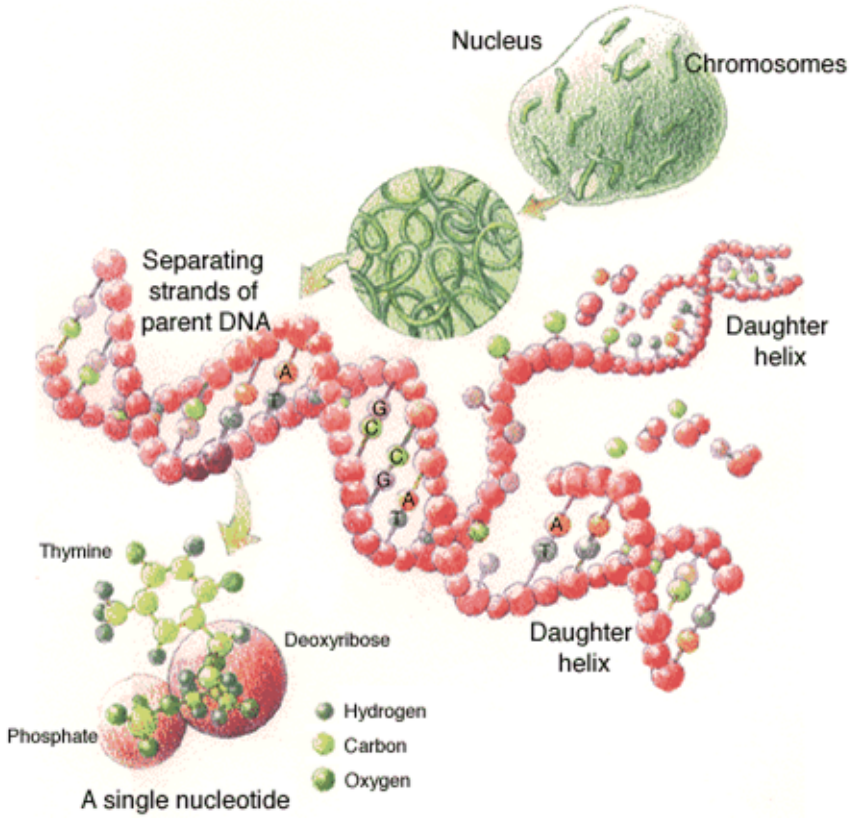

Figure 1.2: Source: [14]. Chromosomes.

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. In 1977 molecular biologists discovered that most Eukaryotic genes have their coding sequences, called *exons*, interrupted by non-coding sequences called *introns* (see Figure 1.3). 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.

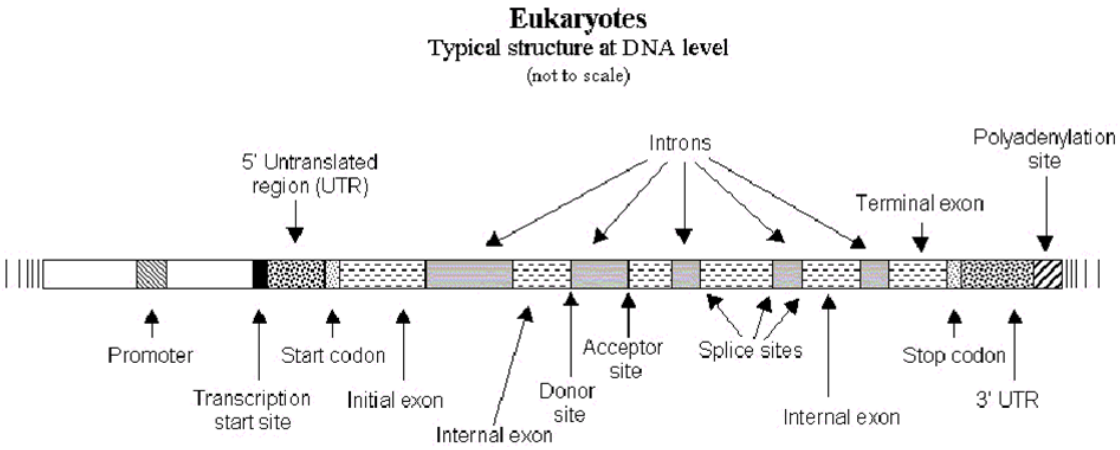

Figure 1.3: Gene structure in Eukaryotes.

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 a co-linear sequence of amino acids in proteins.
The flow is: DNA→ RNA → Protein (see Figures 1.4 and 1.5). 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 302-313)(see Figure 1.6). 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. Only one of the DNA strands is coding for each gene, meaning: the transcription source strand matters, and for each gene is always done from same strand (different genes might be transcribed from different strand). 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) in-

stead 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 1.4). 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.

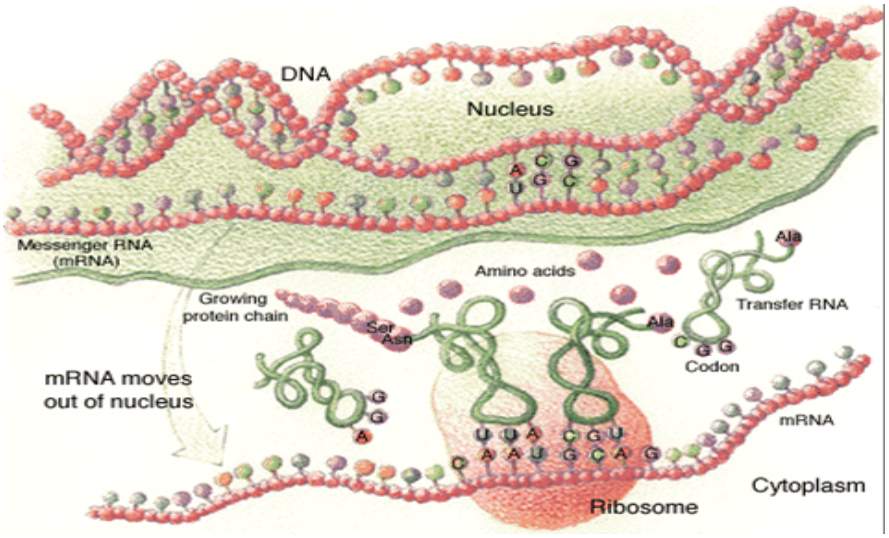

Figure 1.4: Source: [15]. From gene to protein.

A computer world analogy: DNA can be viewed as a hard disk, the RNA as a single program saved in it and the Protein as the output of that program (see Figure 1.5).

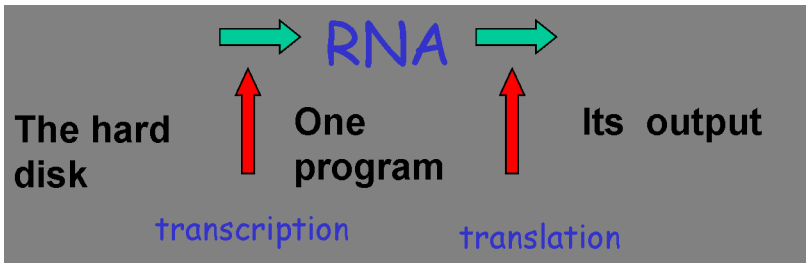

Figure 1.5: From DNA to Protein.

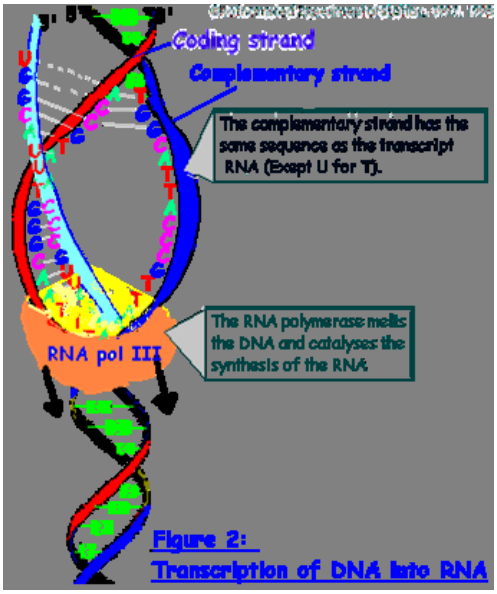

Figure 1.6: Source: [11]. Transcription of DNA into RNA.

Replication

The double helix could be imagined as a zipper that unzips, starting at one end. We can see that if this zipper analogy is valid, the unwinding of the two strands will expose single bases on each strand. Because the pairing requirements imposed by the DNA structure are strict, each exposed base will pair only with its complementary base. Due to this base complementarity, each of the two single strands will act as a template and will begin to re-form a double helix identical to the one from which it was unzipped (see [1], pages 238-266). 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 1.1(c)).

The Genetic Code

The rules by which the nucleotide sequence of a mRNA 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 336). 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

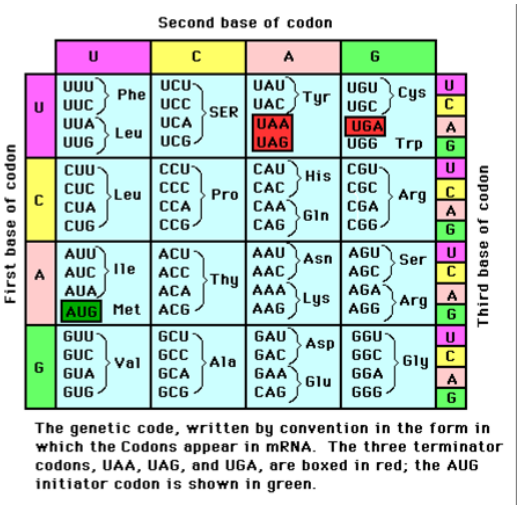

Figure 1.7: Source: [4]. The genetic code table.

is a linear polymer of four different nucleotides, there are $4^3 = 64$ possible codon triplets (see Figure 1.7). 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

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 317-325), thereby producing a much shorter RNA molecule (see Figure 1.8). 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 1.9)

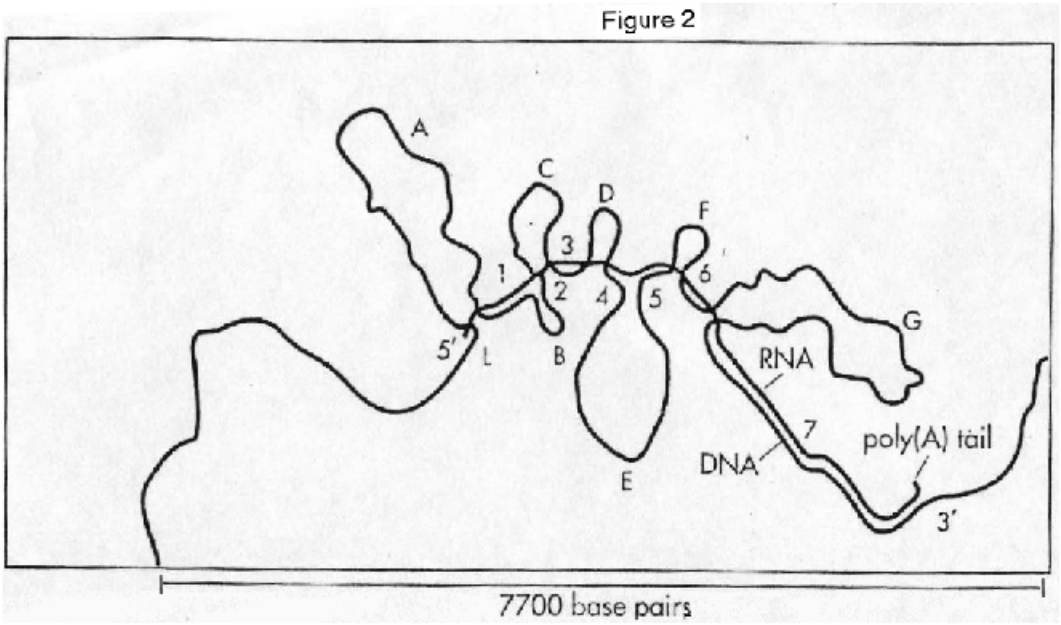

Figure 1.8: Introns are spliced out to form the mature mRNA.

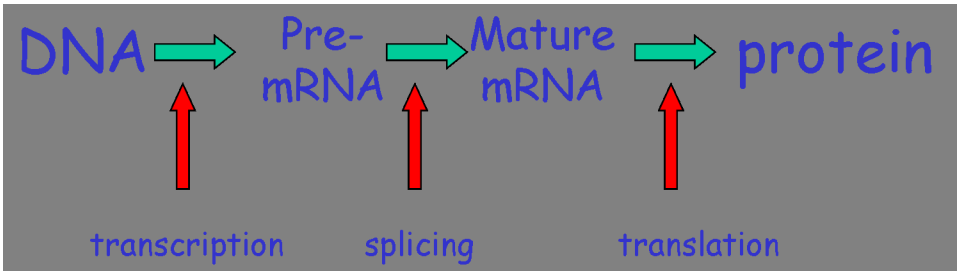

Figure 1.9:

Translation

The *translation* of mRNA into protein (see [1], pages 335-352) 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 1.10). 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.

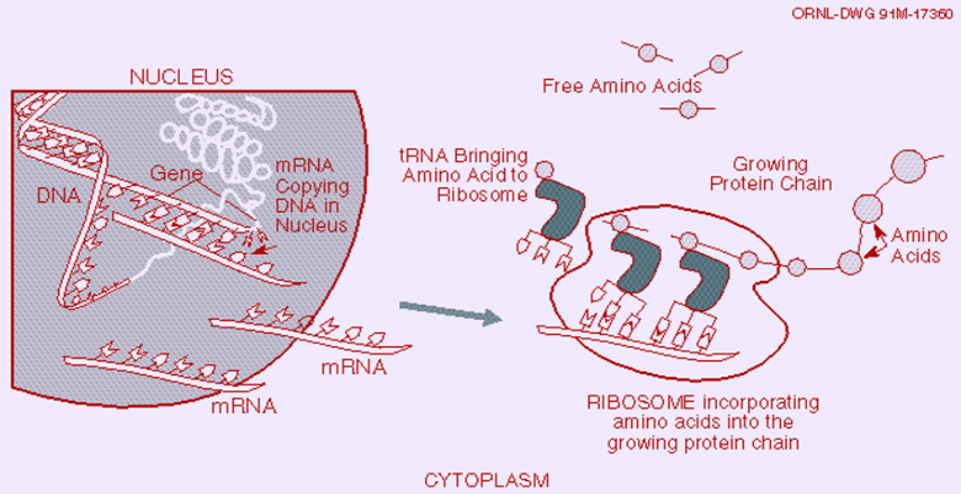

Figure 1.10: Source: [6]. Transcrption of DNA into RNA.

Proteins

A protein is linear polymer of amino acids linked together by peptide bonds (see [1], pages 7-8, 129-134). 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 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 134-140). 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. The life cycle of protein in normally measured in few hours. Therefore, there is a need to re-construct proteins dynamically.

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 (because of changes in protein's structure during their life cycle in the cell, and also due to alternative splicing).
- There is a small difference between the genome of two people (about 0.1%).

Human DNA length and its number of chromosomes aren't the biggest. Rice, for example has longer DNA. Virus's DNA length, on the other hand, is only dozens of thousand, since its life cycle is short and the replication should be efficient.

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 [16] and [9].

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 *in-vitro DNA synthesis*, which is catalyzed by DNA polymerase, and they determine the ends of the final DNA fragment that is obtained.

Each cycle of the reaction requires a brief heat treatment to separate the two strands of the genomic DNA. The success of the technique depends on the use of a special DNA polymerase isolated from a thermophilic bacterium that is stable at much higher temperatures than normal, so that it is not denatured by the repeated heat treatments. A subsequent cooling of the DNA in the presence of large excess of two primer DNA oligonucleotides allows these oligonucleotides to hybridize to complementary sequences in the genomic DNA. The annealed mixture is than incubated with DNA polymerase and an abundance of the four nucleotides (A, C, T, G), so that the regions of DNA downstream from each of the two 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 [17].

1.2.3 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 [12] and [18]).

- 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 1.11).

For a more detailed timetable of the Human Genome Project see [19].

1.2.4 After the HGP, the Next Steps

The first step in the research of the human genome, was the HGP, which discovered the full DNA sequence of the human genome. Next steps in the research are:

- Functional Genomics - a main research area - seek to find out function of genes/proteins. HGP revealed the building blocks, but the next challenge is to understand their role. So Far, there is some knowledge of the functionality of about 5000 human genes.

- Understand gene regulation - how proteins production is controlled.

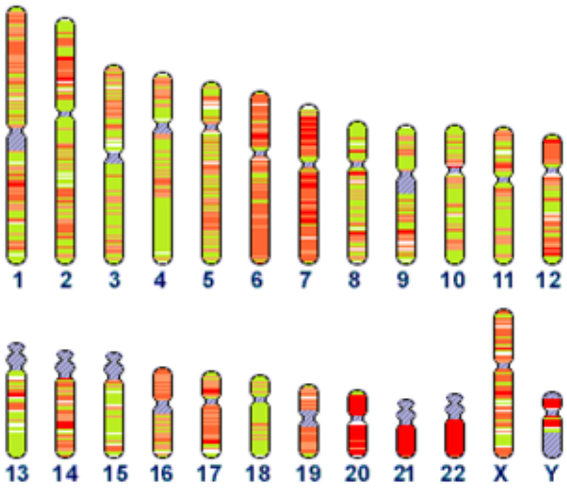

Figure 1.11: Source: [13]. The Public Human genome sequencing progress as of 31/12/2001, scaled from green = draft to red = finished, 63% finished, 34.8% draft, 97.8% total.

- System Biology - new research area - figure out the protein networks, and learn how proteins interact with each other, rather than understand the role of single protein. The functionality of some proteins has a meaning only when interacting with others - proteins can function together or disturb each other.
- Identify individual differences in sequences - two individuals' genomes differ in 1 of 1000 nucleotides. This could be the base for finding genetic diseases or other characteristics that appear in human beings.
- Genome-wide high throughput technologies - exploring all genes at once (vast data):
  - Transcriptome - global gene expression profiling - exploring mRNA.
  - Proteome - wide-scale protein profiling.
- Paradigm shift - Reductionist (Hypothesis driven) to Holistic (exploratory, Hypothesis generating). More of this will be elaborated in "Functional Genomics".

1.2.5 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 non-achievable 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 provide 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.

1.2.6 Measuring Genes Expression

There are technologies for measuring the mRNA existence in a living cell for understanding the functionality of the genes. These methods can measure all mRNA at once, and check how many mRNAs from each gene exist in the cell. These technologies enable comparing approximately 10,000 full length genes expression in human cells (one decade ago we were able to measure up to 10 genes at once):

- Different tissues in the same human may express different genes, according to their role in the human body (an eye cell and a liver cell don't express same genes).
- The same cell may express different genes under different circumstances (stress, nutrition, etc.).
- Cells express different genes during lifetime (for instance, embryonic's genes expression differs from adult's genes expression).

Technologies for measuring mRNA for learning about gene functionality assume the following:

- The level of mRNA in the cell is a direct indication of the protein level in the cell, since the major regularity is on the subscription process, and not the transcription process.
- Genes are expressed only when needed.

So, detecting changes in gene expression level provides clues on its product function.

A basic rule on which many of these technologies rely on is *hybridization*. Hybridization occurs when 2 single-strand DNA sequences are bonding to each other according to the complementary rule, forming hydrogen bonds. Thus, when trying to discover existence of a particular DNA sequence, one can use a recognizable probe with the complementary sequence. A *probe* is the tethered cDNA with known sequence which we use in order to discover information about the *target* which is the free mRNA sample whose identity/abundance is being detected.

1.3 DNA Chips and Microarrays

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 1.12).

A DNA chip is a dense matrix of DNA probes. Each spot contains a probe (or many copies of it) that match a certain mRNA. When putting colored single-strands DNA/RNA in the chip, thousands of hybridization are formed simultaneously. Then according to the color levels in each spot, one can determine the amount of each strand. There are many variant of DNA chips.

- For accuracy we should mention that some of the technologies are actually using cDNA rather than mRNA. cDNA is a DNA strand that was synthesized according to the mRNA (its actually the DNA sequence of the gene, without the introns). For our purposes this is a technical non-important difference, and we might use the term mRNA where we should have used cDNA. Also, when referring to DNA sequence in probes an hybridization, we sometimes also mean RNA sequence.

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.

1.3.1 Technologies

Oligonucleotide Arrays

The basic idea, developed (and patented) by a company named Affymetrix, is to generate probes of oligos (a sequence of nucleotides) that would capture coding region on mRNA. The length of the oligos used depends on the application, but they are usually no longer than 20 bases. The density of these chips is very high (and is increased over the years), for instance,

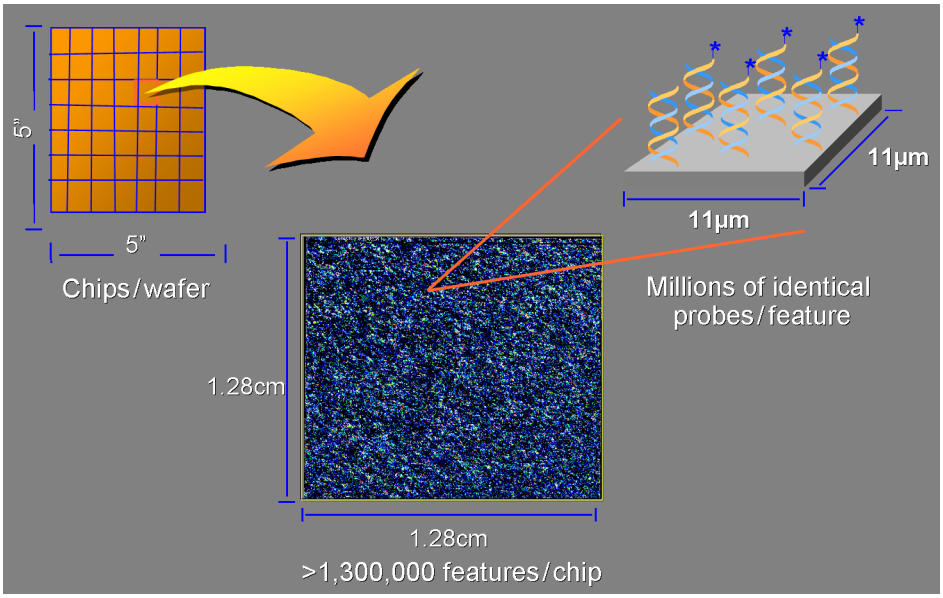

Figure 1.12: Wafers, Chips, and Features.

a chip with size of 1cm by 1cm can contain 1,000,000 oligo types, about 11 microns for each spot (see Figure 1.12).

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 a *terminator*, 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 (for creating probes of 20 bases, there is a need of 80 masks). Figures 1.13 and 1.14 demonstrate the production process.

One of the problems with this method, is that the length of the probe might be 20 at maximum due to increasingly inaccuracy in manufacturing of long probes. However, hybridization of 20 bases isn't accurate enough - mis-hybridizations can occur , and also a probe might match more than one mRNA. To address this issue, the DNA chip contains about 20 different probes for each gene. In addition, for each synthesized probe , a second

mismatch probe (identical to it, except for the central oligo-nucleotide) is also synthesized. This comes to imply of mis-hybridizations, since for perfect hybridization we expect that there won't be hybridization with the mismatch probe. So, when analyzing the results and computing mRNA level, we take into account the mismatch number for each probe (see figure 1.15).

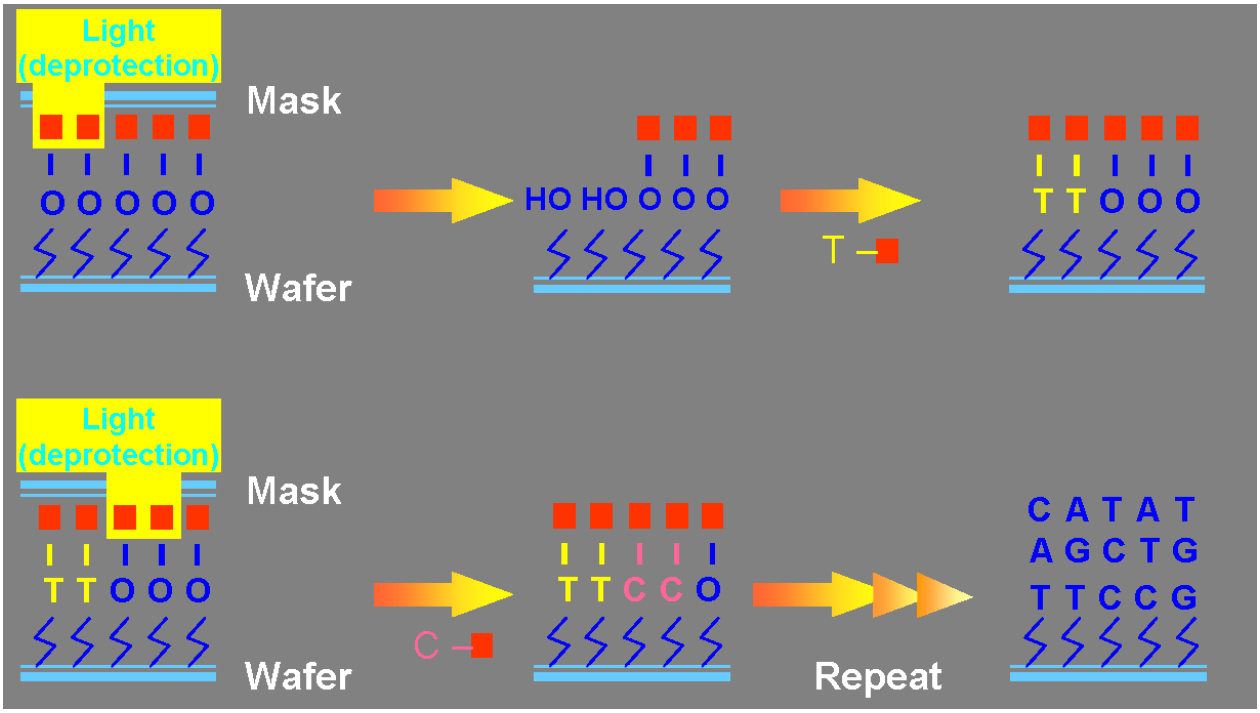

Figure 1.13: Manufacturing DNA chips. 1-2) The light removes the terminator from the chains not covered by the mask, creating hydrogen bonds instead. 3) Bonds are formed with a nucleotide base. 4-6) The process is repeated with a different base.

When using a DNA chip, we should extract mRNA from cells (we use thousands cell tissue, since we can't work with a single cell; it also helps overcoming an inaccuracy problem), and mark it (usually by synthesizing cDNA with fluorescent nucleotides). Then we should warm the RNA, make the hybridization with the chip, wash, scan and check the signals. At the end we should analyze the results by computing the number of mRNA expressed in a single cell.

cDNA Microarrays

In this approach, which was developed in Stanford, each spot in the chip contains a long DNA sequence that are the complement of the cDNA (typically few hundreds of bases), instead of

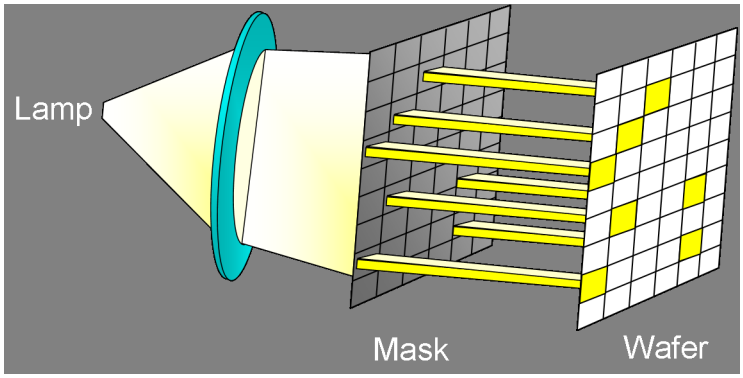

Figure 1.14: GeneChip Manufacturing Process. A typical experiment with an oligonucleotide chip. Labeled RNA molecules are applied to the probes on the chip, creating a fluorescent spot where hybridization has occurred.

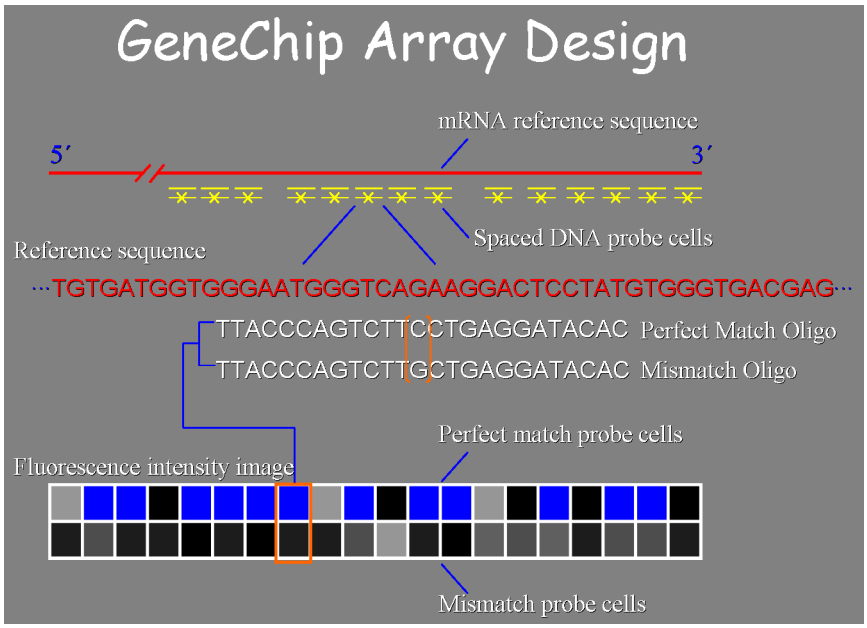

Figure 1.15: GeneChip Array Design

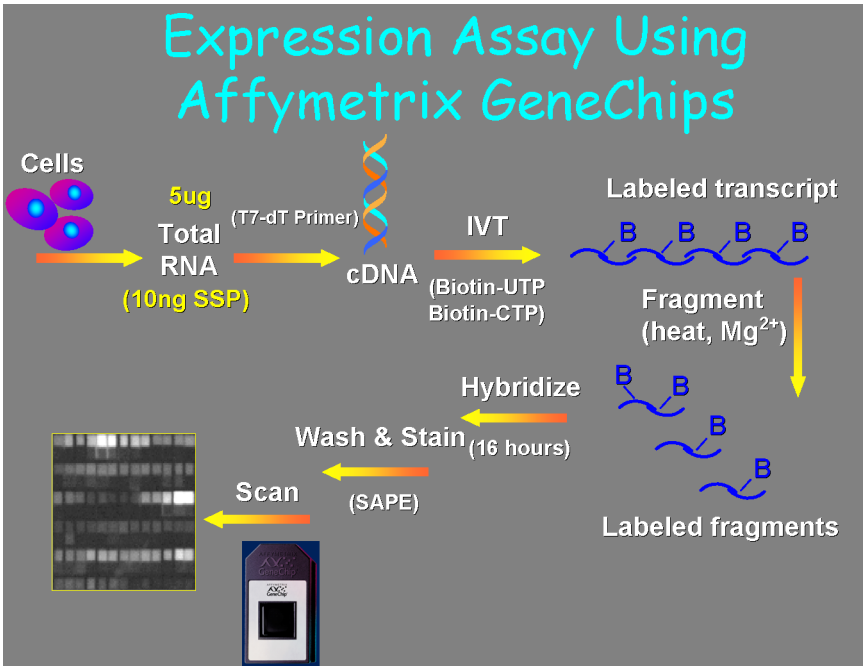

Figure 1.16: Expression Assay Using Affymetrix GeneChips

short oligos. Since cDNA clones are much longer than oligos, a successful hybridization with a clone is an almost certain match for the gene. However, the manufacturing process of this technology can't control the amount of probes in each spot (since its based on extraction of RNA from the cell). So, the technology cannot be used for determining the quantity of mRNA in the cell (absolute number), but for comparing levels of gene expressions in different cells (relative number). For example, extracting mRNA from two different tissues cells, and marking each with difference marker. Then merge all mRNA and make hybridization with the chip. Eventually, analyze the ratio and the differences of gene expression between these tissues. Another example is using a healthy tissue sample as a reference and comparing it with a sample from a diseased tissue (like tumor). (see Figure 1.17)

Using Mirrors for DNA probe synthesis

This technology - done by NimbleGen - is similar to Affymatix' DNA chip, but differs in its manufacturing process. NimbleGen builds its arrays using photo deposition chemistry with its MAS system. At the heart of the system is a Digital Micromirror Device (DMD), similar to Texas Instruments' Digital Light Processor (DLP), employing a solid-state array of miniature aluminum mirrors to pattern up to 786,000 individual pixels of light. The DMD creates "virtual masks" that replace the physical chromium masks used in traditional arrays. These "virtual masks" reflect the desired pattern of UV light with individually addressable

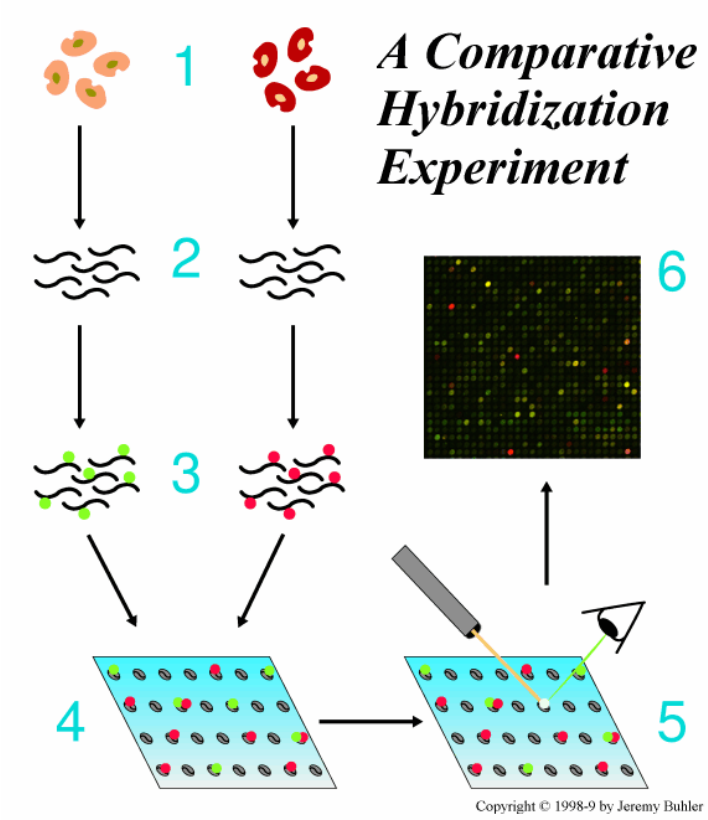

Figure 1.17: cDNA Microarray. 1) Two cells to be compared. On the left is the reference cell and on the right the target cell. 2) The mRNA is extracted from both cells. 3) Reference mRNA is labeled green, and the target mRNA is labeled red. 4) The mRNA is introduced to the Microarray. 5) According to the color of each gene clone the relative expression level is deduced. 6) cDNA chip after scanning.

aluminum mirrors controlled by the computer. The DMD controls the pattern of UV light on the microscope slide in the reaction chamber, which is coupled to the DNA synthesizer. The UV light deprotects the oligo strand, allowing the synthesis of the appropriate DNA molecule (see Figure 1.18).

This method is chipper, faster and simpler than Affymatrix's method. The mirrors direction might be changed hundreds of times in second. Designing new specific chip with Affymatrix' method may cost 1/4 million dolar (for building the masks). With this method it may be done over night. However, Affymatrix's method is still wider (Affymatrix recently bought NimbleGen).

Agilent's SurePrint Technology

Agilent, which was part of HP a decade ago, gained a hugh knowledge in inkjet. They use a cassete with 4 nucleotides (A,C,G,T) instead of colors, and use it for inkjeting nucleotides to the desired places (see Figure 1.19). They can produces probes 60 bases long normally, and even 100 long in special ways. Synthesis are done over glass slides. A DNA chip usually contains 22,000 probes, from which 18,000 are for genes, and the rest for regulatory and repeats. Another 3,000 are blanks for customized needs.

This technology synthesizes probes on the chip, but with no accuracy on the amount of probes in each spot, so it may be used for comparing levels of expression, similarly to cDNA microarrays.

1.3.2 Raw Data

The outcome of DNA chips, in all technologies, 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 either ratio values or absolute values. In the future, hopefully, each entry will contain distribution of values (for example, 10% that the value is 7, and 50% that the value is 50, etc.).

1.3.3 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

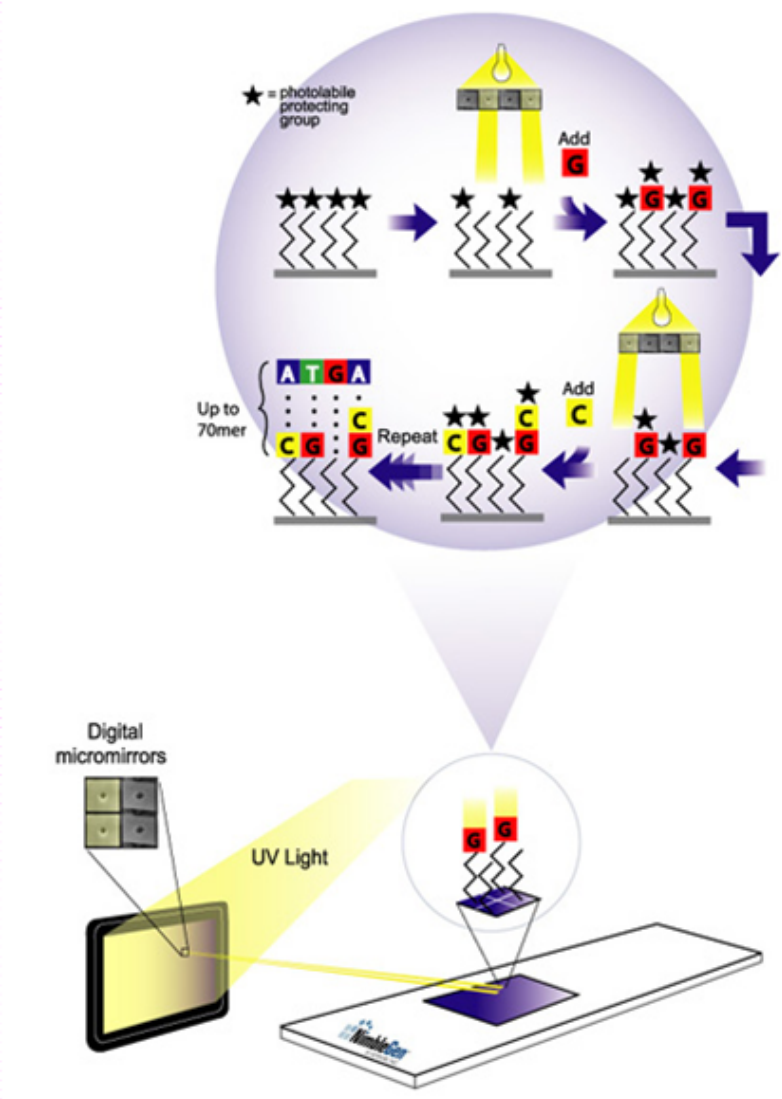

Figure 1.18: Digital Micromirror Device's (DMD) micromirrors

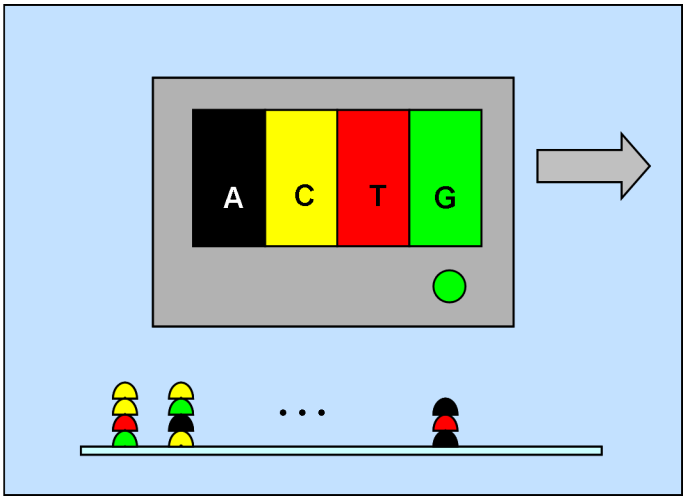

Figure 1.19: SurePrint process

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 DNA sequence (e.g. gene). The general idea is using a chip that contains all the possible sequences of a given length. Target samples (about 20 times longer than the probe length) are marked and hybridized, and their sequence is deduced from the hybridization results. More information on this subject is available in former years' scribes: scribes 1,2,3 of the years 2002 and 2004.

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.

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.

1.4 Biological 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.

1.4.1 Pathways

The Figure 1.20 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 many types of molecular networks which are all combined into complex biological systems. An example of one type of network, a *Metabolic Pathway* can be seen in Figure 1.20. The Metabolic Pathway involves a chain of catalyzed biochemical reactions. 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 biological network networks.

The following two figures (1.22 and 1.23) 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. In this pathway the final product is part of a complex that inhibits part of the pathway steps. In other pathways the products might involve other paths. Also, an inhibition might be indirect, when inhibiting an expression of gene of a protein that catalyzes the process, etc.

1.4.2 Signal Transduction

Figure 1.23 demonstrates signal transduction - a complex cellular process initiated by signaling molecule arrived from outside of a cell (a result of outside changes). A typical signal transduction will be with a transmembrane protein with some part outside of the cell, and

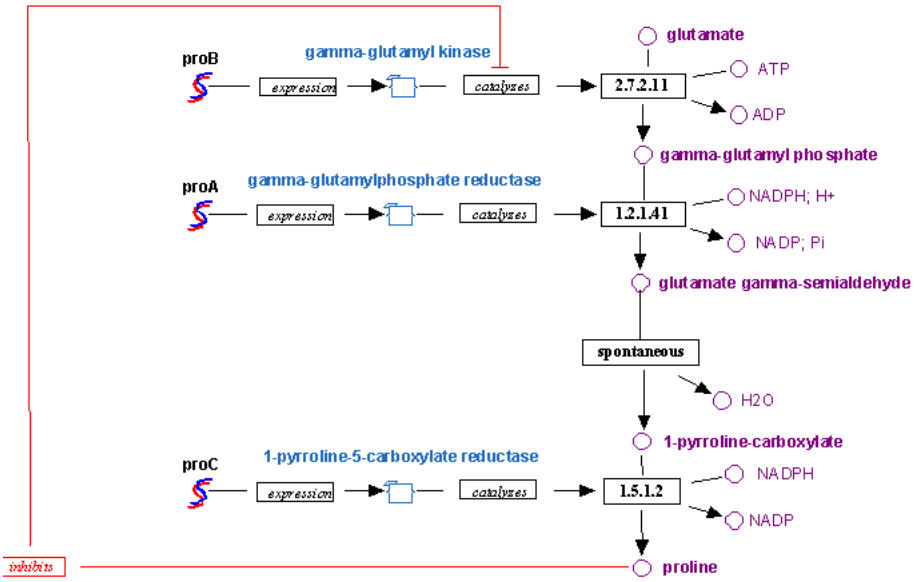

Figure 1.20: Source: [10]. An example of a metabolic pathway: Proline biosynthesis. The role of gene expression in catalyzing chemical reactions.

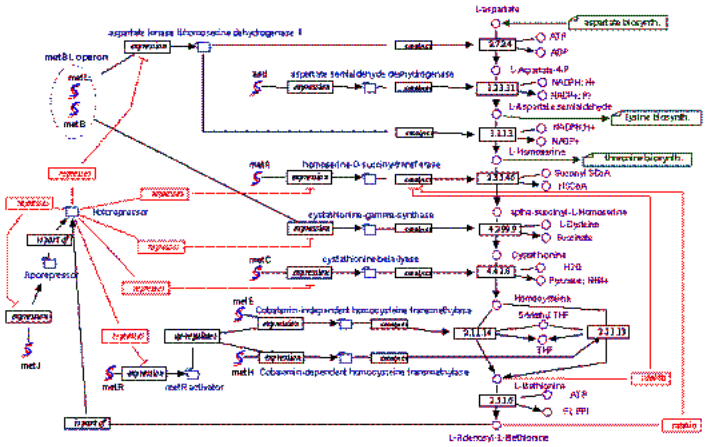

Figure 1.21: Source: [10]. Methionine biosynthesis network in E-coli.

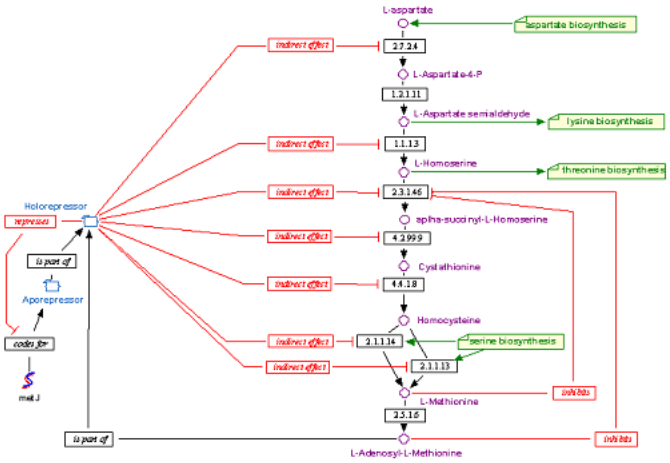

Figure 1.22: Source: [10]. Shortcut representation of the biosynthesis pathway presented in Figure 1.21.

some part inside. When the signaling molecule connects the outside part, a structural change occurs in the inside part, which makes it connect to some other protein in the cell, and make some interaction with a third protein etc. The last product of the interaction catalyzes a process, which product is a protein that can enter the cell nucleus and regulate another protein transcription. This network is an example of second network type called Protein Network.

Figure 1.24 demonstrates signal trasduction in sea-urchin, which regulates the evolvement of a tissue that should be a part of the intenstine.

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.

1.4.3 Reverse Engineering

Some of the regulations are logical (either the regulator exist or not), and some are by edge limit - when the regulator's amount crosses some limit. There are compound networks. Some genes are regulated by 6 other proteins. Such network was revealed by biological technologies. Hopefully, we'll be able to use the new technologies for revealing such networks, and have models the biological networks.

The examples shown above present the goal of reverse engineering of generic networks, in order to determine the network topology and compounds, and the network logic (regulatory, functionality) at system level. This produces 2 main challenges:

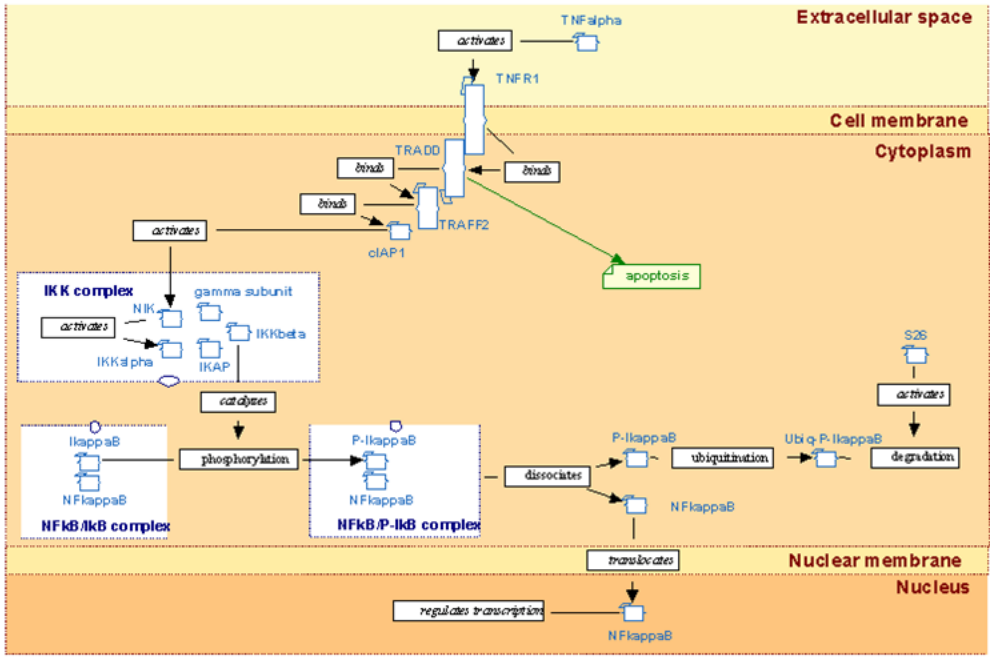

Figure 1.23: Source: [10]. A gene network that performs signal transduction from outside the cell into the nucleus.

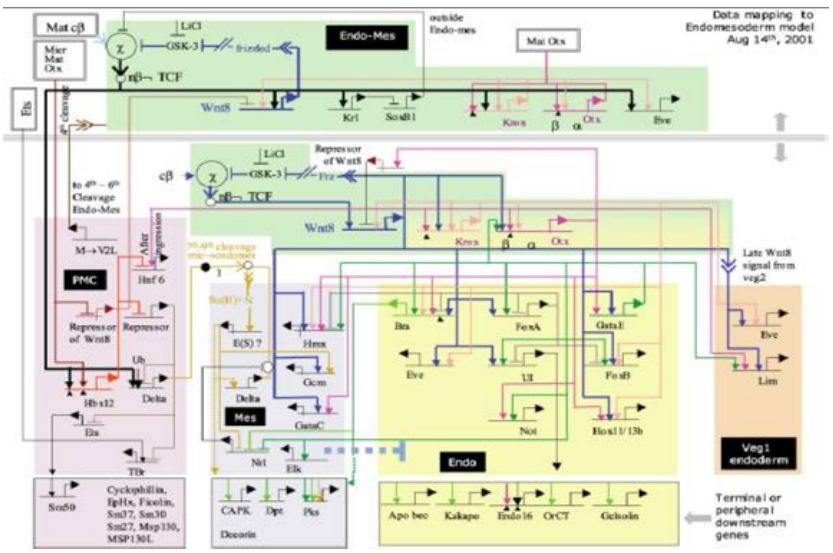

Figure 1.24: Source: [10]. A Genetic network controlling early development of sea urchin endomesoderm.

- Exploit the emerging vast, heterogeneous data sources (gene sequences, knowledge of proteins, interaction between proteins).
- Develop computational tools that are robust, realistic and rigorous.

1.4.4 Functional analysis

Functional Analysis - discovering and modelling gene networks from experimental data. Tools:

- DNA sequences
- Monitoring gene expression levels, via e.g., DNA microarrays, 2-hybrid systems
- In the future - monitoring protein expression (proteomics 2D-gels, protein chips)
- Causing perturbations:
  - Genetic - knockout, over-expression. For example, knockout of one gene of yeast. In most cases it will be fatal, but in some cases it will live with some functional problems. Another example is down syndrome, where one chromosome is duplicated. These things can be done in mice too
  - Biological - one or more non-genetic factors are altered (like: nutrition, environment, temprature...)

Methods presented above supply biological data in terms of expression levels of many genes at different time points and in different conditions.

1.4.5 Goals

- Being able to adequately model the network
- Construction of a knowledge-base of gene regulatory networks
- Verification of pathways or gene networks hypotheses
- Refine/improve a given network
- Discover gene network from experimental data
- Optimize experiments to verify/reconstruct network
- Incorporate highly heterogeneous genome-wide data
- Infer some plausible model of the network from the observations with minimal number or cost of biological experiments

Bibliography

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson. *Molecular Biology Of The Cell*. Garland Publishing, Inc., 2002.

[2] A. L. Lehninger. *Biochemistry*. Worth Publishers, Inc., 2000.

[3] http://www.sciencemag.org/cgi/content/full/291/5507/1304/.

[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.dnaftb.org/dnaftb/23/concept/index.html/.

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

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

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

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

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

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

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

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

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

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