

# Analysis of Biological Networks: Introduction to Molecular Biology and Biological Networks\*

Lecturer: Roded Sharan

Scribe: Liat Perlman and Oded Magger

Lecture 1, March 4, 2009

## 1 Introduction to the course

In classical biology, each biological lab focused on the little details of the whole biological system. It mostly investigated a single gene or a number of elements that constitute a part of a large system.

With the progression of the technology, and the collection of large-scale data, a new trend began in biological research called **Systems Biology**. This new doctrine didn't focus on the details of the system but rather examined the system as a whole. It models the behavior of an entire biological system.

In this course we'll deal mainly with systems biology - looking at the biological systems as a whole rather than its individual components. We'll review the elements of the system and the connections between them. In addition, we will learn how the system operates in the course of time, as a dynamic system.

Why is this important?

1. Provides better understanding of life.
2. Numerous applications in medicine, such as:
  - Disease diagnosis
  - Development of drugs
  - Personalized medicine -It is the concept that managing a patient's health should be based on the individual patient's specific characteristics, such as his genomic sequence.

### 1.1 Example of personalized medicine

The company 23andMe took biology into personal grounds. For each person willing to pay the price, it sequences his genome. The company utilizes this data to link a person with disease and traits. In addition, it provides insights into the ancestry of the customer.

---

\*Based on scribes by Moran Cabili, Elad Donsky, Elena Kyanovsky and David Hadas.

## 2 Basic Biology

### 2.1 The cell

The basic unit of life is the cell. All living creatures are made of cells which are small membrane-bounded units filled with a concentrated aqueous solution of chemicals, the **cytoplasm**.

The membrane functions as a selective barrier to substances that enter the cell and exit from it.

Each cell is an independent entity, capable of creating copies of itself by growing and dividing into two identical daughter cells. The complete characteristics of an organism is carried by each of its cells. This hereditary information is stored within the DNA molecule. In higher multicellular organisms, each cell carries the same DNA content, storing the complete biological information essential for life [1]. All cells transform DNA to proteins, which determine cells' structure and function.

Organisms can be divided into two classes:

- **Prokaryotes** - Organisms whose cells do not contain a nucleus. These are simple unicellular organisms such as bacteria and archaea.
- **Eukaryotes** - Organisms whose cells' DNA is stored within a nucleus and also contain other inner-membrane elements (Figure 1). The eukaryotes include more complex multicellular organisms such as plants, animals and fungi. However, they also include unicellular organisms such as yeast and amoebas. The unicellular organisms are the most widespread creatures in the world.

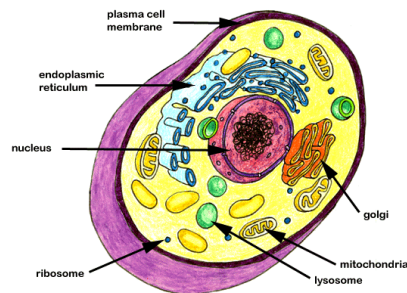


Figure 1: Source: [2]. Eukaryotic cell

Two major discoveries occur in biology:

- Mendel discipline:

Gregor Johann Mendel was an Augustinian priest and scientist, and is often called the father of genetics for his study of the inheritance of certain traits in pea plants. Mendel showed that the inheritance of these traits follows particular laws, which were later named after him. Its rediscovery prompted the foundation of the discipline of genetics [3].

One of Mendel's experiments on peas was as follows:

**A** is an attribute of a high pea while **a** is an attribute of a low pea. Mendel crossbred a pure species of high pea with a low one. Surprisingly, all the hybrid offsprings (F1 generation) were not of medium height as was expected, but were all high. This phenomenon led Mendel to conclude that **A** is a **dominant** trait. A dominant trait is one that will always be manifested, no matter which other trait is present in the organism. When he crossbred peas of F1 generation with each other, he obtained

peas that maintain a 1:3 ratio, three were high and a single one was low. Mendel stated that attributes **A** and **a** are two variants of a factor, that determines the length of the pea.

He claimed that each individual has two factors for each trait, one from each parent. The two factors may or may not contain the same information. The factors Mendel referred to were later called **genes**. Alleles are specific variants of a gene.

In this experiment **A** is the **dominant** trait while **a** is a **recessive** one, both are variant of the gene that determined the length of the pea. A dominant trait is one that express by itself and silences the expression of the recessive trait. This experiment is illustrated in Figure 2.

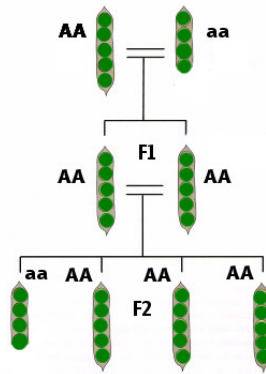


Figure 2: Mendel's pea experiment

- Watson and Crick:

In 1953 Watson and Crick characterized the DNA properties. They discovered that the DNA is a long double stranded molecule composed of 4 types of bases A,G,C and T (Figure 3).

## 2.2 DNA characteristics

Life depends on the ability of cells to store, retrieve and translate the genetic instructions required to make and maintain a living organism. This hereditary information is passed on from a cell to its daughter cells at cell division. In the 1940s, **deoxyribonucleic acid (DNA)** was identified as the likely carrier of genetic information [1]. The building blocks of the DNA molecule are called **nucleotides**. The nucleotides are linked together into a chain by covalent bonds (a strong chemical bond involving electron sharing). This chain is referred to as a **DNA strand**. Each nucleotide is composed of three parts: a five-carbon sugar, a phosphate group and a base. There are 4 different bases: **adenine (A)**, **cytosine (C)**, **guanine (G)** and **thymine (T)**. C and T bases are composed of one carbon ring which is called **pyrimidine**. A and G bases, however are composed of two coupled carbon rings, a structure termed **purine**.

The DNA molecule is composed of two DNA strands held together by hydrogen bonds between the bases on the different strands. The bases do not pair at random: A always pairs with T and G always pairs with C. The G-C base pair is stronger due to the formation of 3 hydrogen bond instead of 2 hydrogen bonds formed by the A-T base pair. The hydrogen bond is 20 times weaker than the covalent bond, which enables a relatively easy separation of the two strands [1].

Each DNA strand has a chemical polarity, meaning, each of its two ends are chemically different. This polarity is indicated by referring to one end of the chain as the 5' end (pronounced "**five prime**") and the other as the 3' end (pronounced "**three prime**"). The 5' end terminates with a phosphate group attached to the fifth carbon on the sugar-ring. The 3' end terminates with hydroxyl group (OH) on the third carbon of

the sugar-ring. The two strands of the DNA double helix are antiparallel, that is, the polarity of one strand is oriented opposite to that of the other strand (Figure 3).

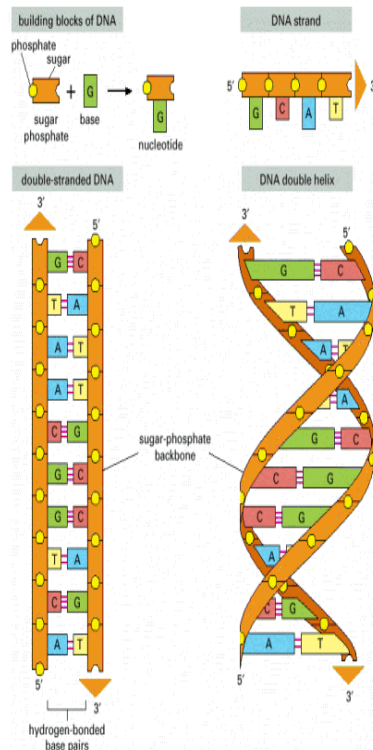


Figure 3: Source: [1]. DNA structure

## 2.3 DNA Packaging

The length of the outstretched human DNA is approximately 2 meters long. However, the diameter of the cell nucleus is only about  $6 \mu\text{m}$ . Therefore, the DNA must be packed compactly enough to fit inside the cell, but still organized enough (and not tangled up) to become uncoiled in order to be replicated and transcribed. The DNA is wrapped around proteins called **histones**, in a complex called **chromatin**. The continuous folding of chromatin results in a more compact structure - the **chromosome** (Figure 4). Some primitive organisms' cells have only one chromosome (meaning their entire DNA is in one piece), whereas human cells have 23 pairs of chromosomes. Humans (and all other animals) are **diploid** - meaning, each of their cells contains 2 copies of each chromosome, one from each parent. The entire genetic material of the organism is called its **genome**.

## 2.4 DNA Replication

DNA replication produces two identical copies of a single DNA molecule. In order to carry out this process, the participation of enzymes is essential. **Enzymes** are proteins that catalyze a biochemical reaction. DNA base pairing is the central principle on which the replication process relies. Each DNA strand contains a sequence of nucleotides that is exactly complementary to the nucleotide sequence of its partner strand. Therefore each strand can act as a **template** for the synthesis of a new complementary strand (Figure 5) [1]. The initial step of the replication process is the separation of the two DNA strands by the **helicase** enzyme.

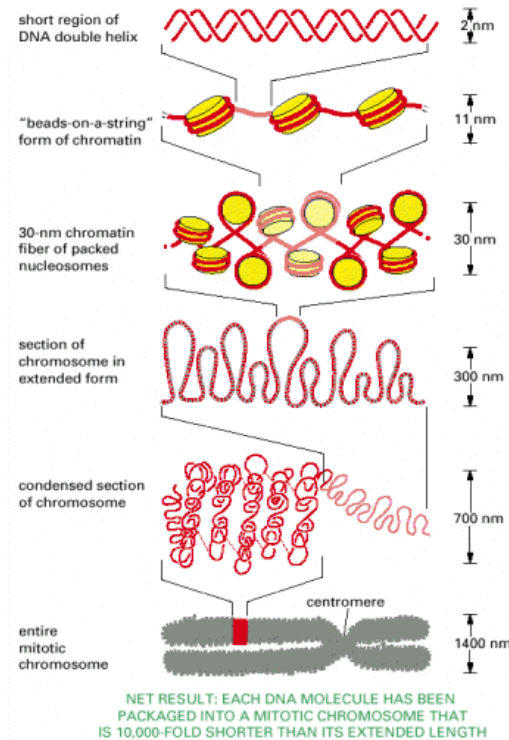


Figure 4: Source: [1]. Levels of DNA condensation in the cell

A Y-shaped junction formed by this separation is called the **replication fork**. Next, a short RNA segment, termed **primer** is base-paired to the template strands by the **primase** enzyme. Finally, **DNA polymerase** enzyme, the heart of the replication machine, synthesizes a new DNA strand by adding free nucleotides that are capable of forming base-pairs with the template strand. DNA polymerase can only join a free nucleotide to a sequence (primer) which is already base-paired to the template. Hence, primers are crucial for the initiation of the replication. Other enzymes participate in this process in order to complete the final formation of the two DNA double helices, each containing one original strand and one newly constructed strand (Figure 5) [1].

DNA polymerase can only synthesize a new DNA strand from 5' to 3' direction, and there is no enzyme capable of synthesizing the other direction. As a result, the replication fork is asymmetrical. Meaning, one new DNA strand is formed on a template running from 3' to 5' (the **leading** strand), whereas the other is formed on a template running on the opposite direction (the **lagging** strand). The leading strand is synthesized continuously, while the lagging strand must be made initially as a series of short DNA fragments that are later joint together to a single continuous strand (Figure 6). The lagging-strand fragments are also known as **Okazaki fragments**. After the generations of the Okazaki fragments all the primers within them are removed by enzyme called **nuclease**. Once the primers are removed DNA polymerase complements them. Finally, an enzyme called **DNA ligase** joins all the DNA fragments together, to produce a single continuous DNA strand identical to the original one.

During the replication process, errors might occur. For example, adding to the newly constructed strand a nucleotide that can not be based paired with the template strand will cause an error. An error in the new

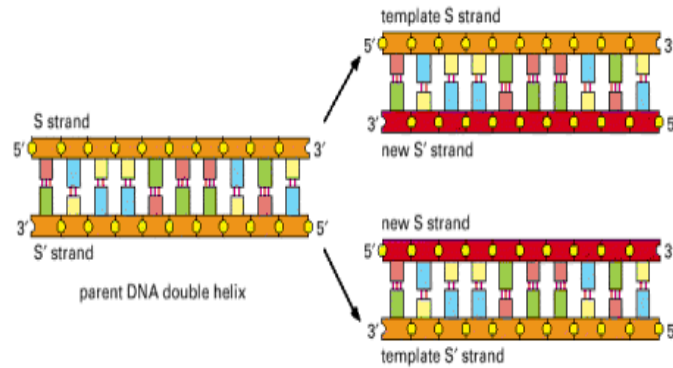


Figure 5: Source:[4]. DNA replication

formed DNA molecule is called **mutation**. A mutated DNA might harm the proper function of the cell carrying it. However, the DNA polymerase has a very efficient error-correcting activity called **proofreading**, and the rate of errors is very low, 1 per  $10^9$  bases.

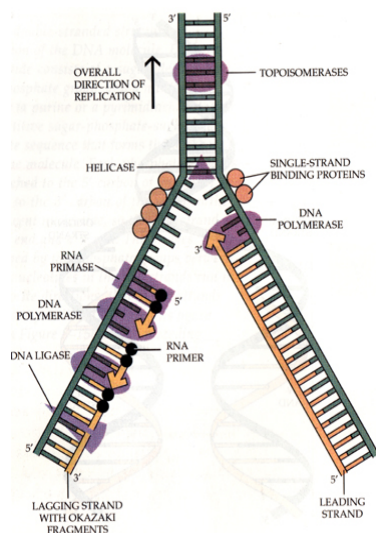


Figure 6: Source: [1]. Lagging strand with okazaki fragments in DNA replication

## 2.5 Genes

A gene is the basic unit of heredity in a living organism. Genes are segments of DNA that determine the characteristics of a species as a whole and of function of the cells within it [1]. They do that by encoding a functional unit called **protein**.

The transformation of a gene into a protein is called **expression**.

Surprisingly, genes constitute only a very small part of the DNA - only 2%-3% of the human DNA are genes. The rest of the DNA, is non-coding and is termed **"junk DNA"**. Among the junk DNA there are:

- Regulatory regions - sequences that determine the expression level of different genes

- RNA elements
- Retrotransposons - are genetic elements that can amplify themselves in a genome and are ubiquitous components of the DNA of many eukaryotic organisms [5].
- Pseudogenes - non-functional relatives of known genes that have lost their protein-coding ability.

## 2.6 Proteins and the Central Dogma of molecular biology

Proteins are crucial elements to the existence of each organism as they build the cell and drive most of its functions. They are sequences composed of 20 basic components called **amino acids**. Each amino acid contains an **amino group** (marked N) and a **carboxyl group** (marked C). The amino acids that constitute the protein are linked together by **peptide bonds**. The amino group at one end of the protein's sequence and the carboxyl group at the other end, determine its orientation from N to C.

Protein folds into a specific 3D structure with the lowest energy (Figure 7). The fold of the protein plays a central role in its function to such an extent that a misfolded protein might not perform its function properly.

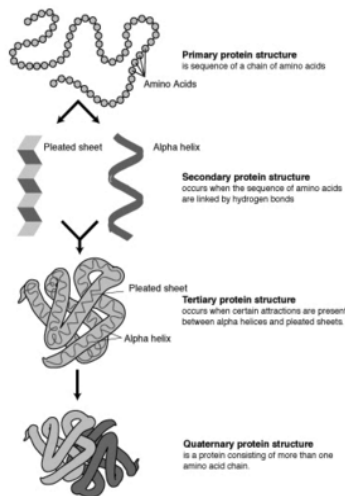


Figure 7: Source: [6]. Protein folding

The **central dogma** is a framework for understanding the transition from the instructions coded in the DNA (genes) to the production of proteins (Figure 8).

The stages of the process are:

- **Transcription** - Producing RNA sequence from DNA template in the nucleus.  
After transcription the newly synthesized RNA is transferred by other proteins to an inter-cellular medium called the **cytoplasm**.
- **Translation** - The synthesis of a protein from RNA template in the cytoplasm.  
An enzyme called **ribosome** attaches to the RNA and uses the information in it as a template for the synthesis of the protein.

The information can be thought of "being written" in two languages: one where nucleotides function as characters, and the other where amino-acids (from which proteins are made) function as the characters.

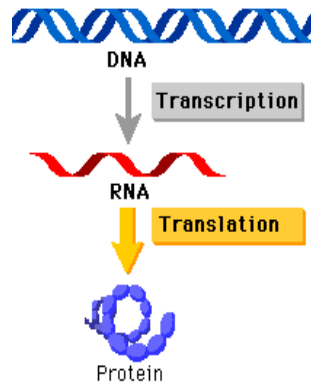


Figure 8: Source: [7]. The Central Dogma of molecular biology

As mentioned previously, DNA replication is a very accurate process. The frequency of a mistake in this procedure is  $1:10^9$  bases. However, the process of transcription (in which DNA is transformed to RNA) has much greater tendency toward errors. This raises the question - Why is there such a remarkable difference in the frequency of mistake between RNA transcription and DNA replication?

1. DNA is a heredity information - it is carried from a parent to his offsprings. As a result, alterations in the DNA of an organism will not only affect it but may also affect his descendants. RNA , however is not passed from a parent to his offsprings.
2. RNA is a single sequence that encodes a protein. Generally, many RNA sequences are transcribed from each gene, forming multiple copies of the same protein. Hence, if a single copy of RNA is impaired then other same source proper RNA sequences may compensate for the damaged one.

In computer science nomenclature the DNA is the hard disk - it contains the exact necessary instructions to generate a machine. Thus, the maintenance of the DNA is crucial to the function of the machine. RNA, however, is a single a program that its output is a protein.

## 2.7 RNA

RNA (ribonucleic acid), similarly to DNA is also a nucleic acid. However it has some differences:

- The sugar part of the nucleotide is ribose rather than deoxyribose (DNA), meaning there is an additional oxygen atom on the sugar ring.
- Both DNA and RNA contain the bases adenine (A), guanine (G) and cytosine (C). However, RNA contains the base uracil (U) instead of thymine (T) in DNA. U is base-paired with A and is chemically similar to T.
- RNA is usually single stranded.

RNA is oriented from 5'(phosphate) to 3'(sugar). It can fold on itself in a variety of ways (Figure 9), one of which is the hairpin structure. In addition, RNA can function as a catalytic molecule.



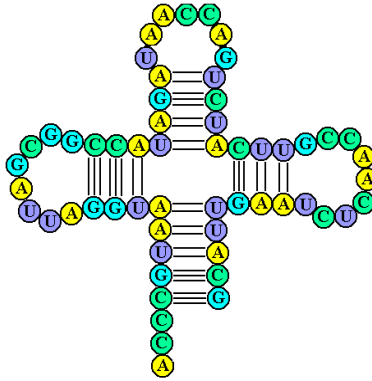


Figure 9: Source: [8]. Example of RNA that folds on itself

## 2.8 RNA transcription

The first step of creating a new protein is transcribing a new RNA molecule. The process that produces a new RNA molecule based on DNA is termed transcription. The most dominant protein complex that participate in this process is RNA polymerase. In order for transcription to take place a small portion of the DNA unwinds to two separate strands. After the unwinding, one of the two strands serves as a template for the synthesis of RNA. RNA Polymerase enzyme binds the incoming ribonucleotides into the RNA chain according to complementary base-pairing with the DNA template. As opposed to DNA replication, only one strand serves as a template, therefore the RNA transcription from 5' to 3' is always continuous.

How does RNA polymerase know where to initiate and terminate the transcription?

There are special sequences in the gene sequence, that enable RNA polymerase to identify where to initiate and terminate the transcription. There are proteins named **transcription factors** that identify these special sequences such as the TATA-box on the promoter region of the gene and then they recruit RNA polymerase to the location in the genome that should be transcribed. The transcription terminates when the RNA polymerase encounters a second signal sequence, the terminator. Then it halts and releases both the RNA molecule and the DNA template. Note that unlike DNA polymerase RNA polymerase does not need an oligonucleotide primer and initiates the transcription process after attaching itself to the promoter sequence on the DNA template strand. The newly synthesized RNA sequence is called mRNA (messenger RNA).

The mRNA must undergo a number of processes before it is translated into protein:

1. **Capping.** Capping is the addition of a modified guanine nucleotide to the 5' end of the mRNA.
2. **Polyadenylation.** Polyadenylation is the binding of a sequence consisting of consecutive A's (poly A tail) to the 3' of the mRNA.
3. **Splicing.** The primary mRNA transcript can be divided into exons sequences and introns sequences. Splicing is the process in which the introns are being removed from the initial transcript, leaving only the exons to be translated.

A primary mRNA molecule that has undergone all these stages is called **mature mRNA** to designate that it is prepared to be translated into protein. All the processing stages that transform the primary mRNA into mature one occur in the nucleus. Once the mRNA is mature it exits the nucleus and enters the cytoplasm where protein translation takes place.

## 2.9 The genetic code

As previously described, the information encoded in the DNA is transcribed to RNA and later translated into protein. How is the information in the linear sequence of ribonucleotide translated into a chemically different set of subunits, the linear sequence of amino-acids?

The rules by which this translation occurs are known as the **genetic code** (Figure 10). The principles of the genetic code are quite simple. Each triplet of consecutive nucleotides referred as a **codon** specifies a single amino-acid. Since there are 64 different sets of codons and only 20 different amino-acids, the genetic code is redundant and some amino-acids are specified by more than one triplet [1].

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

Figure 10: Source: [9]. The genetic code

## 2.10 RNA translation

The translation of the mRNA sequence into proteins is done by **ribosomes** found in the cytoplasm. The ribosomes are large complexes of proteins that decrypt the code on the mRNA molecule and synthesize a protein accordingly. The tRNA molecules are the elements that match each triplet on the mRNA to its corresponding amino-acid, and elongates the new polypeptide chain synthesized on the ribosome (Figure 11). In principle there are 3 different ways to read a sequence of triplets - 3 different **reading frames**. The reading frame depends on which of the 3 first characters the decoding process begins. In order to resolve this matter, special punctuation signals exist. A unique initiation codon, AUG, determines the beginning point of translation, and 3 different termination codons determine where the translation process ends - UGA, UAA and UAG. Note that the sequences before the initiation signal and after the termination signal aren't translated as seen in Figure 12.

The order of the reading frame is of great importance. A deviation of a single base may totally distort the original sequence of the amino acids which results in an impaired protein.

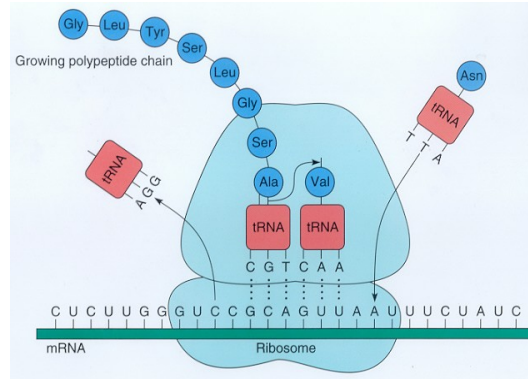


Figure 11: Source: [10]. mRNA translation

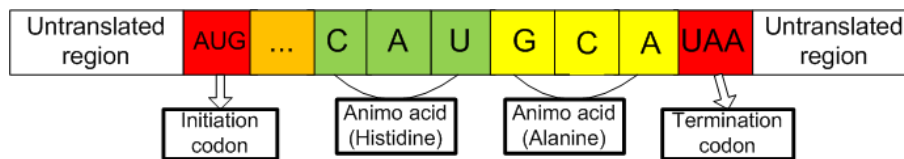


Figure 12: Example of mRNA codon sequence translation

How does a cell specify which of its genes to express and when?

There are several types of regulations in the cell. The most prominent one is transcription regulation. The transcription is an exceedingly regulated process in the cell. It is mostly regulated by transcription factors. Transcription factors (TFs) control transcription by binding to specific DNA sequence motifs. They determine whether or not transcription will take place (they could be repressors or activators of transcription). Note that there are other types of regulations such as translation regulation and post-translation regulation which we will not discuss in detail.

## 2.11 The human genome and model organisms

The human genome consists of 23 pairs of chromosomes that include  $\sim 3,200,000,000$  bases. A typical gene is 1000-3000 bases long, spanning 30K-40K bases. There are  $\sim 25,000$  genes.

There are several model organisms that are thoroughly explored. The following is their order from the least similar to the most similar to human:

- **Budding yeast** - A unicellular eukaryote. Has about 6000 genes.
- **Nematode worm** - Consist of 959 cells and has 19,000 genes. Has a very developed nerve system that is widely studied.
- **Fruit fly** - Holding 14000 genes. It resembles vertebrates though it is not considered to be one.
- **Mouse** - A mammal that is genetically close to humans (98% sequence identity) and has about 30000 genes.

### 3 Basic Biotechnology

The field of Computational Biology owes a great debt to breakthroughs in Biotechnology from the recent decades. These innovative technologies and methods enabled researchers to perform measurements and obtain data at the Molecular level. Many of these methods are system-wide, providing information on the entire activity of the cell - Gene Expression, Protein-Protein Interactions, etc. This data can then be used as input for Bioinformatic tools, or be used as a basis for Biological networks modeling.

In this chapter we will survey several fundamental Biotechnological methods which are at the basis of contemporary Computational Biology.

#### 3.1 Hybridization

Hybridization is a process where complementary nucleic acid strands are paired into a double-stranded hybrid. This process occurs naturally when two DNA/RNA strands with complementary segments meet (it can also occur if one long strand will "fold" on itself). As previously described, the base pairs A-T, G-C form hydrogen bonds which hold the strands "glued" together. The stability of the double-stranded hybrid depends on the length of the complementary paired strands. The longer the strands are the more stable the hybrid pair of strands will be since more base-pairs are formed.

Note that a perfect complementary match is not required for hybridization to occur. For example, two 20-bp-long complementary RNA strands with A in the 10th. position at the first strand and C at the 10th. position in the second might hybridize. The probability that a 'partial' hybridization will take place is dependant on the match proximity (how close we are to perfect match) and environmental conditions.

Hybridization existence helps the identification of segments in the DNA. A specific short single stranded RNA or DNA **probe** is made in order to identify a specific complimentary strand. For example, if one holds a probe of DNA containing a known mutation causing a disease, he can examine whether this probe hybridizes with a DNA extract from a human cell in order to determine whether it carries this disease. The hybridization technique is very useful and is used in many biotechnology methods.

#### 3.2 PCR - Polymerase Chain Reaction

PCR is a method to amplify a DNA sequence - to replicate many copies of the same sequence in a short time.

The procedure starts with a single DNA sequence of up to 10Kbp, which we want to replicate many times. This sequence is typically a gene, part of a gene, or even a non coding region. It will be used as a template for the reaction.

A PCR assay also requires the following:

- **DNA Polymerase** - an enzyme which synthesizes DNA based on a DNA sequence (See the section about DNA replication). Usually this would be a Polymerase brand that operates optimally under heat conditions, such as the Taq Polymerase taken from Thermophilic bacteria.
- **Two types of RNA Primers** - these are short sequences of RNA which are designed to be complementary to sequences at the 5' and 3' edges of the DNA template. These primers are required for the initialization of the replication by the DNA Polymerase. The Primer sequences define the boundary of the sequence that will be replicated from the original template.
- **Deoxynucleoside triphosphates (dNTPs)** - used as substrate for the DNA Polymerase, which will chain them together according to the template DNA sequence.

- **Buffer solution** - provides a chemical environment in which the DNA Polymerase can operate optimally.
- **Cations** - Divalent and Monovalent, also required for the reaction.

PCR is an iterative procedure. It usually starts with the system being maintained at high temperature (90°C) to prevent catalytic activity.

The process, when launched, is composed of multiple cycles. Each PCR cycle in turn is composed of three stages, which are the heart of the chain reaction:

1. **Denaturation** - Increasing the heat of the system causes the double-stranded DNA to separate to two strands. The bond to any RNA primers attached to the DNA from a previous cycle also breaks. This requires 94 – 96°C for approximately 20-30 seconds.
2. **Annealing** - The heat is reduced, allowing the Primers to bind to the single-stranded DNA template. Stable DNA-DNA hydrogen bonds are only formed when the primer sequence very closely resembles the complement to the template sequence. The DNA-Polymerase binds to the primer-template hybrid and begins DNA synthesis. The heat remains high enough to prevent a full re-hybridization of two full-length strands. This requires 50 – 65°C for approximately 20-40 seconds.
3. **Extension** - The DNA polymerase synthesizes a new DNA strand complementary to the DNA template strand. The system's heat at this stage depends on the optimal temperature for the DNA-Polymerase, and the time required depends on the Polymerase's processivity (DNA synthesis rate) and the length of the fragment being replicated.

After  $N$  PCR cycles, we will obtain  $2^N$  copies of the original DNA fragment.

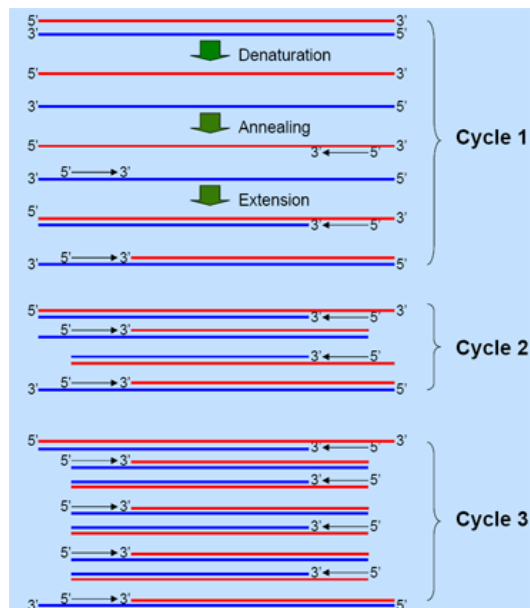


Figure 13: The PCR process.

### 3.3 Gene Expression analysis

The system-wide expression levels of genes are crucial to understanding cellular processes.

This is important because the final product of gene expression - Proteins, tRNAs, small RNAs, etc. - are, in 'Computer Science language', the 'logic' of how the cell operates. Different levels and concentrations of the expression products affect every aspect of cellular life - Metabolism, response to stress, regulation, communication with the environment and many more. It is the different levels of expression for each gene that differentiate cells from each other and derive their activity.

#### 3.3.1 Microarrays

One method to monitor gene expression is by using **DNA chips** or **microarrays** assays. The Microarray enables us to measure the expression of many genes simultaneously.

Two main fundamental biological assumptions are made (though in some cases, are not always true):

- Transcription level indicates genes' expression levels.
- Changes in Gene Expression are caused mainly by environmental effects (Specific conditions, etc).

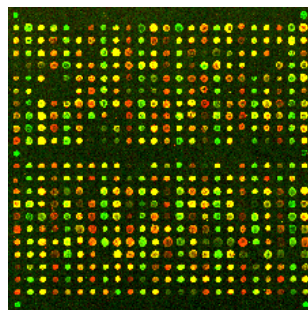


Figure 14: Source: [11]. A typical microarray.

Most genes have different transcription levels at different conditions. According to those differences, one can deduce their functionality and essentiality in different environmental conditions.

The DNA chip is a 2D array, which is divided to many tiny cells. In each cell, complementary strands of a known gene are embedded. A generic procedure for using Microarrays to measure Gene Expression includes:

- The mRNA extract of the tested cell is obtained.
- The tested mRNA (or complementary DNA fragments synthesized accordingly by Reverse-Transcription) are colored and added to the microarray plate.
- The tested nucleic acid strands bind to their complementary probes in the array blocks.
- The excess strands are washed away from the array, leaving each cell with matching strands only. The intensity of the color in each array block is the indication of the expression level of a specific gene.

There are several types of Microarrays. In **2-dye Microarray**, the color of the results is derived from introducing two samples of RNA (or cDNA) to the Microarray: one is colored Red, the other Green. The

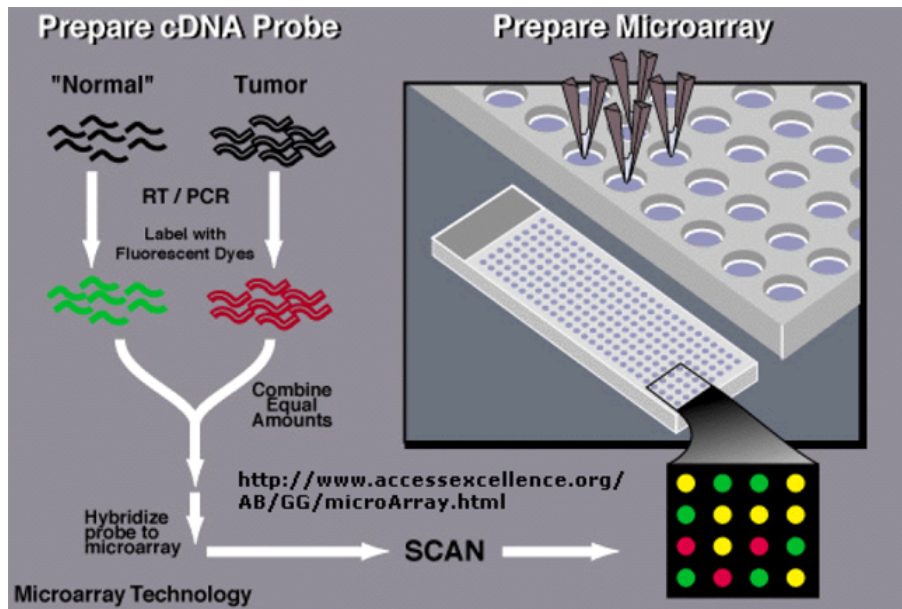


Figure 15: 2-dye Microarray technology.

color of each cell depends on how many 'green' strands bind to it in comparison to how many 'red' ones bind to it. Often, this comparison is done between a 'regular' sample and a sample taken from a unique condition (Environmental stress, disease, different nutrition conditions, etc.). In this way, one can identify genes that are affected by these conditions since their expression is expected to be significantly different than that in the optimal condition. For example genes that are highly expressed in cancer cells in comparison with healthy tissue cells.

### 3.3.2 Applications of Gene Expression analysis

The first application example we present allows doctors to spare women who have Breast Cancer from undergoing unnecessary treatments. In Breast Cancer, about 20% of the ill population will require adjuvant treatment in order to prevent them from developing metastases. The other 80% do not require the extra treatment. However, until 2002, due to the existing medical diagnosis methods, about 80% of the ill population were given the adjuvant treatment.

At 2002, Van't Veer et al[12] took cell samples from various Breast Cancer patients, and observed the Gene Expression. They saw that a considerable group of genes were expressed differently in patients that developed Metastases compared to those who did not. They developed a diagnosis kit that they used to predict whether a patient required adjuvant treatment or not, and succeeded in 17 out of 19 cases.

Van't Veer's method passed FDA authorization, and the diagnosis tool based on it dramatically reduced the amount of Breast Cancer patients receiving unnecessary adjuvant treatment.

The second application is reversing Chemotherapy resistance phenotype.

The **Connectivity Map**[13] is a collection of genome-wide transcriptional expression data from cultured human cells treated with bioactive small molecules and simple pattern-matching algorithms that together enable the discovery of functional connections between drugs, genes and diseases

In acute lymphoblastic leukemia, there is a problem of genetic resistance to a Chemotherapy substance called **Dexamethasone**. In an unrelated experiment, 100 genes appeared to change their expression in

response to treatment with Dexamethasone. These results were compared to the expression profiles of all of the substances in the Connectivity map, and a strong correlation was identified with **Sirolimus**, which is a known immunosuppressant.

Carcinogenic Dexamethasone-resistant cells that were introduced to Sirolimus and then treated with Dexamethasone displayed loss-of-viability similar to that of non-resistant cells, confirming that Sirolimus indeed reverses the Dexamethasone-resistance trait.

### 3.4 Protein-Protein Interactions

Protein-protein interactions refer to the physical association of protein molecules. These interactions are critical to all cellular processes. For example, signals from the exterior of a cell are mediated to the inside of that cell by the interactions between signalling proteins. Other proteins are able to induce some reaction only while interacting as a complex (for example, the DNA polymerase is active only as a complex of several proteins). Finding interactions between proteins which are involved in common cellular functions, is a way to get a broader view of how they work cooperatively in a cell.

#### 3.4.1 Yeast Two-Hybrid

One technique that can be used to study protein-protein interactions is the "yeast two hybrid" system. This technique uses the GAL4 Transcription Factor (TF) in yeast. This protein contains 2 domains: a DNA-binding domain (BD), that is capable of binding to the DNA, and an activation domain (AD), that is capable of activating the transcription of the DNA. In order to transcribe a gene, the binding domain of the TA binds the promoter region of the gene (Figure 16). Then, the activating domain binds the binding domain and enables the transcription of the gene. Hence, the activity of the TF requires both a DNA binding domain and an activation domain. However, the binding domain and the activation domain do not necessarily have to be on the same protein. In fact, a protein with a DNA binding domain can activate transcription when simply bound to another protein containing an activation domain; this principle forms the basis for this technique.

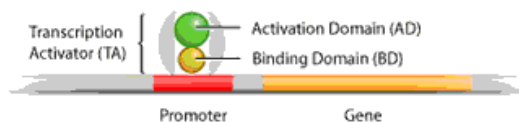


Figure 16: Source: [14]. Complete interaction of a transcription factor and a promoter.

In the two-hybrid assay, two fusion proteins are created: the protein of interest (X), called the "bait", which is attached to the binding domain, and its potential binding partner (Y), called the "prey", which is fused to an activation domain. If protein X interacts with protein Y, the binding of these two will form an intact and functional transcriptional activator. This newly formed transcriptional activator will then go on to transcribe a gene. In order to measure the gene product, the GAL4 promoter is followed by a reporter gene. A **reporter gene** is a gene which its product is easily traceable. A gene which its protein product spread fluorescent light is a good example for a reporter gene. Finally, the amount of the gene produced can be used as a measure of interaction between our protein of interest and its potential partner (Figure 17).



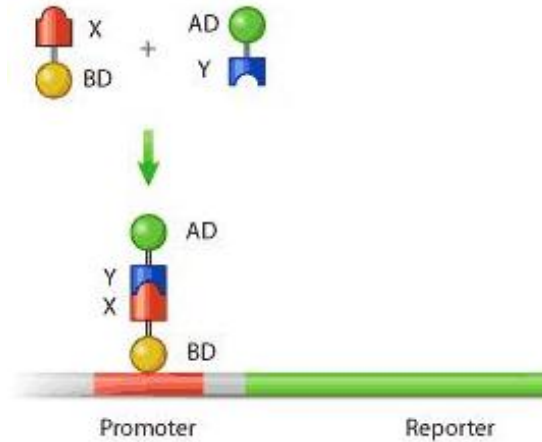


Figure 17: Source: [14]. Yeast Two-Hybrid.

### 3.5 Transcriptional Interactions

Transcriptional interactions are physical interactions between transcription factor proteins and DNA sequences in order to control the transcription of genes. Such an interaction can either induce the transcription of a gene or suppress it.

#### 3.5.1 Chromatin immunoprecipitation - ChIP

Chromatin immunoprecipitation (ChIP) assay, is an experimental method used in order to determine whether proteins, particularly transcription factors, bind a particular region on the DNA. The basic steps of this procedure are:

1. The DNA-binding proteins are crosslinked (bind) to the DNA by using the substance formaldehyde. At this point the proteins bind the DNA regions tightly.
2. The DNA is isolated and cut in to small fragments. The DNA and protein are still attached together in a complex.
3. Specific **antibodies**, which are molecules able to specifically bind the DNA-binding protein examined, are added. These antibodies will bind the DNA-protein complex.
4. The DNA-protein-antibody complex is isolated by a method called precipitation.
5. The DNA from the isolated complexes is separated and cleaned.
6. Special hybridization techniques are used in order to identify the DNA fragments.

#### 3.5.2 ChIP-on-chip assay

ChIP-on-chip, also known as genome-wide location analysis, is a technique for isolation and identification of the DNA sequences occupied by specific DNA binding proteins. These binding sites may indicate functions of various transcriptional regulators. The ChIP-on-chip assay follows the procedure of the ChIP assay. After obtaining the purified DNA fragments, the DNA is fluorescently labeled. The other fraction of DNA

fragments that was not bound by the DNA-binding protein is fluorescently labeled by a second color. Next, both sets of labeled DNA are combined to a microarray chip. Finally, based on the ratios of the microarray signals the binding regions on the DNA are identified [15].

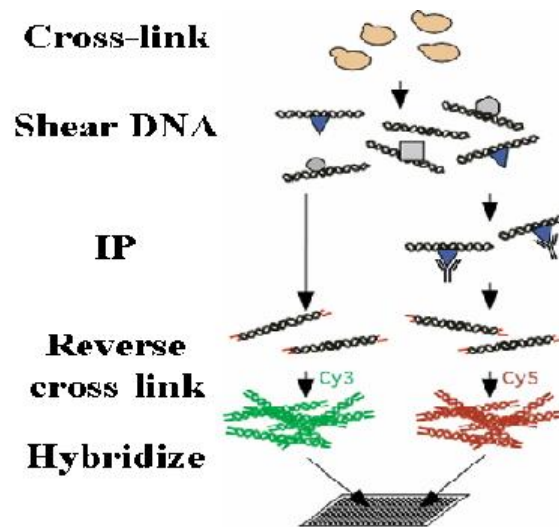


Figure 18: Source: [16]. Transcriptional interactions process.

## 4 Biological Networks

### 4.1 What are networks?

A network is a set of elements connected by links. Networks enable a simple and uniform representation of complex structures and processes in many domains such as: computer science (more commonly known as **graphs**), sociology, communication, and more.

A network consists of two sets:

- **vertices** - represent the elements themselves. For example, every town in a road map will be represented by a vertex.
- **edges** - represent the relations between pairs of elements. For example, a road between two towns in a road map will be represented by an edge.

An example for a real-world entity that can be represented as a network is Facebook. In Facebook each user or profile would be a vertex, and two vertices will be connected by an edge if the two relevant users are friends in Facebook.

### 4.2 Biological networks

Networks also enable the characterization of biological processes. Here are some of the common biological networks:

- **Co-Expression network** - Describes the relation between the expression pattern of different genes. In this network vertices represent genes. Two vertices are connected if the corresponding genes have similar expression patterns, meaning they are expressed under the same conditions.

- **Transcriptional network/gene regulatory network** - a graph with two sets of vertices. One set represents transcription factors and the other represents genes. Each edge represents binding of a transcription factor to a gene. This type of networks reflect the cell's regulatory circuitry.
- **Protein-protein interaction network** - vertices represent proteins and edges represent interactions, that is, two proteins are connected if they interact with each other. This network reflects the cell's machinery and signaling pathways.
- **Metabolic network** - a directed graph, where each vertex represents a metabolite (a molecule), and an edge represents a biochemical reaction. A **biochemical reaction** is a chemical process that results in the conversion of one biological molecule to another. Metabolism is the biochemical modification of chemical compounds in the cell. The two main processes in metabolism include the biosynthesis of complex organic molecules and their breakdown. The metabolic network describes the cell's metabolic pathways which consist of a sequence of reactions in which molecules interact and convert to a different molecule or compound.
- **Disease-Gene network** - a bipartite graph with two sets of vertices. One set represents Diseases of a given organism (e.g. Humans) and the other set represents the organism's genes. A gene and a disease are connected by an edge if the gene is involved in the disease (e.g. Causal Gene).

Biology is often more complicated than what appears in a network model. For example, protein-protein interactions can be dependent on the location within the cell.

Another such complexity level is the time dimension. For instance, in the Galactose Utilization pathway, protein **GAL80** can be at one time bound to Transcription Factor **GAL4**, suppressing its activity, and at another time bound to **GAL3**, in which case it can't bind to **GAL4**. Both interactions will appear in the Protein-Protein Interaction network, although they do not occur simultaneously.

### 4.3 Computational challenges in biological networks

The study of biological networks gives rise to several major computational challenges:

- **Preprocessing:** data normalization and integration across different experiments.
- **Characterization of network attributes** (such as degree distribution) and construction of good network models.
- **Module finding:** Identification of sets of genes showing similar interaction and/or expression patterns.
- **Classification of biological conditions.**
- **Prediction of biological attributes:** Predict an attribute, such as protein function, according to the knowledge on the network.
- **Pathway reconstruction from cause-effect experiments.**
- **Characterizing disease and drug responses.**

#### 4.4 Degree distribution

Degree of a node A is the number of nodes to which node A is connected. One of the most basic features of a network is the degrees of its nodes, and the degrees distribution. The **Degree distribution** for a network is denoted by  $P(k)$ . It represents the probability that a node has degree  $k$ . Denote  $N$  the number of vertices in the network.

The degree distribution can be calculated the formula:

$$P(k) = \frac{\text{Number of nodes with degree } k}{N} \quad (1)$$

Note that directed graphs would have two degree distributions, namely **In-degree distribution** and **Out-degree distribution**.

The average degrees in a graph is given  $d \equiv \sum_{k \geq 0} k \cdot P(k)$

The number of edges in a graph is calculated by  $\frac{N \cdot d}{2}$

#### 4.5 Random networks (Erdős/Rényi)

Random networks that were characterized by Erdős/Rényi have a few fundamental properties:

- Each pair of nodes is connected with probability  $P$ , meaning each edge has a probability for its occurrence.
- The mean degree of the network is given by  $d = (N - 1) \cdot p \approx N \cdot p$ . Where  $N$  is the number of nodes in the network.
- Degree distribution is binomial, asymptotically Poisson:

$$\text{(For large } N) P(k) = \frac{e^{-d} \cdot d^k}{k!} \quad (2)$$

In this type of networks most of the vertices are concentrated around the mean degree (Figure 19).

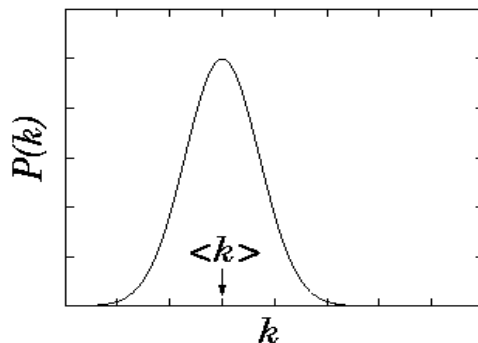


Figure 19: Source:[17]. Poisson distribution of random networks

## 4.6 Scale-free networks

Recent interest in **scale-free** networks started in 1999 with work by Albert-Lszl Barabasi and colleagues who mapped the topology of a portion of the Web , finding that some nodes, which they called **hubs** had many more connections than others and that the network as a whole had a **power-law** distribution of the number of links connecting to a node. Barabasi and collaborators coined the term **scale-free network** to describe the class of networks that exhibit a **power-law** degree distribution.

They showed in 1999 that many real networks behave like scale-free networks [18].

Power-law degree distribution is of the following form:  $P(k) \propto k^{-c}, k \neq 0, c > 1$ .

This distribution describes networks which contrasts with the Poisson and exponential distributions. Although it is characterized by many vertices with low degrees, it also includes a small but non-negligible number of vertices with high vertex degrees. Such vertices are highly important to the network connectivity and serve as hubs.

Figure 20 shows a plotting of the power-law distribution in log-log representation. Under a log-log representation, a Power-Law distribution is characterized by a linear line.

Power-law distributions are named **scale free** since they can be scaled without altering the distribution. If we denote the distribution by  $p(x)$ , scaling the distribution by a factor  $a$  results in  $p(ax) = g(a)p(x)$ . It has been proven that the only scale free distributions are power-law distributions. Thus, scale-invariability is an exclusive feature of scale free networks, no other type of networks obey this property.

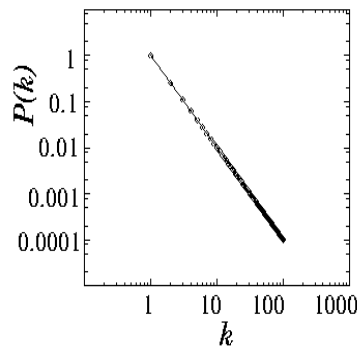


Figure 20: Source [18].Power law degree distribution in log-log representation

## 4.7 Degree distribution of real networks

In the following section we'll view a number of real networks which are scale-free.

### 4.7.1 The internet

The Internet can be described as a graph with routers as nodes and physical links between routers as edges. In 1999 C. Faloutsos [19] showed that the degree distribution of the Internet can be approximated as:  $P(k) \sim k^{-2.5}$  using a log-log scale.

Figure 21 shows the power-law distribution graph of the Internet on a log-log scale.From the graph we can infer that the Internet network behaves like a scale-free network.

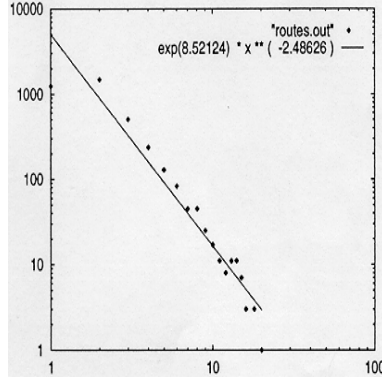


Figure 21: Source [19]. Degree distribution of the Internet

#### 4.7.2 Film actors

Barab'asi and Albert [18] analyzed in 1999 the collaboration graph of movie actors. This graph represents a well documented example of a social network. Each actor is represented by a node, two actors being connected if they were cast together in the same movie. The probability that an actor has  $k$  links (characterizing his or her popularity) has a power-law tail for large  $k$ , following  $P(k) \sim k^{-2.3}$ . Figure 22 shows the degree distribution of the actors social network.

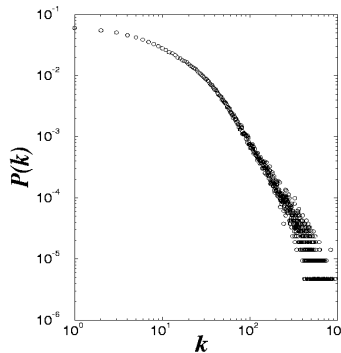


Figure 22: Source [18]. Degree distribution of the film actors collaboration network

#### 4.8 Protein interaction networks

Yook [20] analyzed in 2004 a variety of Protein Interaction networks and found the probability that a protein interacts with  $k$  other proteins (characterizing his or her popularity) follows a power-law of  $P(k) \sim k^{-2.5}$ .

Figure 23 shows the Power-law distribution graph of the protein interaction network. It is interesting to note that each protein network by itself as well as all taken together seem to all follow the power-law distribution with  $c = 2.3$ .

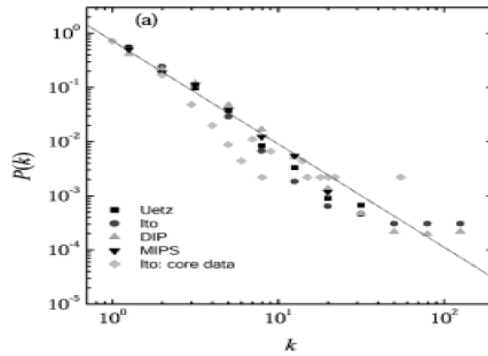


Figure 23: Source:[20]. Degree distribution of the protein-protein interaction network

#### 4.8.1 Metabolic Networks

In these networks, each Metabolite is represented by a vertex, two metabolites are connected if there is a biochemical reaction transforming one metabolite to the other.

The edges are therefore directed resulting in a directed graph. Surprisingly, despite significant variances in different species individual constituents and pathways the metabolic networks from all kingdoms of life are scale-free with the same value range of  $c$ ,  $c = 2.2 \pm 0.2$  [21].

Figure 24 shows some examples of the power-law distribution graph of the metabolic networks.

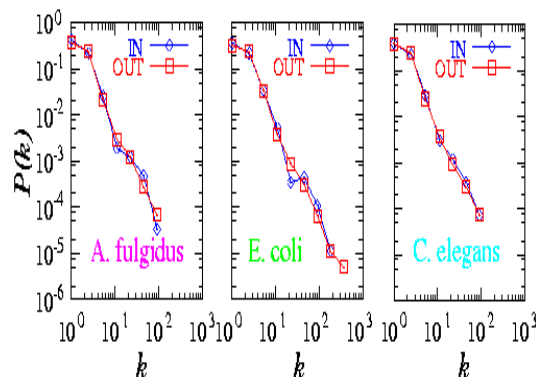


Figure 24: Source:[21]. Degree distribution of metabolic networks

We saw that scale free networks are quite common in real life. A question that raises is why are they so prevalent?

According to Barabasi, real networks aren't formed randomly, but gradually.

The growth of the network obeys a law called **preferential attachment**. Preferential attachment means that new vertices prefer to bind to the more connected vertices. Formally, the probability that a vertex A will connect to other vertex B is proportional to the degree of the vertex B.

In summary, we learned in a nutshell about the basic concepts in biology and about biological networks. In the following weeks we'll discuss the application of networks and how to deduce from them biological insights.

## References

- [1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*, pages 3–375. Garland Science, 4th edition, 2002.
- [2] <http://www.biologycorner.com/resources/>.
- [3] [http://en.wikipedia.org/wiki/Gregor\\_Mendel](http://en.wikipedia.org/wiki/Gregor_Mendel).
- [4] [http://www.ornl.gov/sci/techresources/Human\\_Genome/publicat/primer/fig4.html](http://www.ornl.gov/sci/techresources/Human_Genome/publicat/primer/fig4.html).
- [5] <http://en.wikipedia.org/wiki/Retrotransposon>.
- [6] <http://www.answers.com/topic/protein-structure-png>.
- [7] [http://www.phschool.com/science/biology\\_place/biocoach/translation/overview.html](http://www.phschool.com/science/biology_place/biocoach/translation/overview.html).
- [8] <http://www.slic2.wsu.edu:82/hurlbert/micro101/images/TransferRNA.gif>.
- [9] [http://www.mun.ca/biology/scarr/MGA2\\_03-20.html](http://www.mun.ca/biology/scarr/MGA2_03-20.html).
- [10] <http://library.tedankara.k12.tr/chemistry/vol1/biochem/trans100.htm>.
- [11] <http://www.biology.ucsc.edu/mcd/research/index.html>.
- [12] Laura J. van 't Veer et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.
- [13] Justin Lamb et al. The connectivity map: Using gene-expression signatures to connect small molecules. *Science*, 313(5975):1929–1935, 2006.
- [14] <http://www.bioteach.ubc.ca/MolecularBiology/AYeastTwoHybridAssay>.
- [15] <http://www.bio.brandeis.edu/haberlab/jehsite/chip.html>.
- [16] [http://jura.wi.mit.edu/cgi-bin/young\\_public/navframe.cgi?s=14&f=fig1a](http://jura.wi.mit.edu/cgi-bin/young_public/navframe.cgi?s=14&f=fig1a).
- [17] P. Erdos and A. Renyi. On the evolution of random graphs. *MTA Mat. Kut. Int. K ozl.*, 5:17–61, 1960.
- [18] R. Albert and A. L. Barabási. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.
- [19] M. Faloutsos, P. Faloutsos, and C. Faloutsos. *On power-law relationships of the Internet topology*. ACM Press, 1999.
- [20] S. H. Yook, Z. N. Oltvai, and A. L. Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4:928942, 2004.
- [21] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000.