Molecular Cell Biology:
A Quick Introduction for non-Biologists

Eduardo D. Sontag

1 The Cell

The fundamental unit of life is the cell (Figure 1).

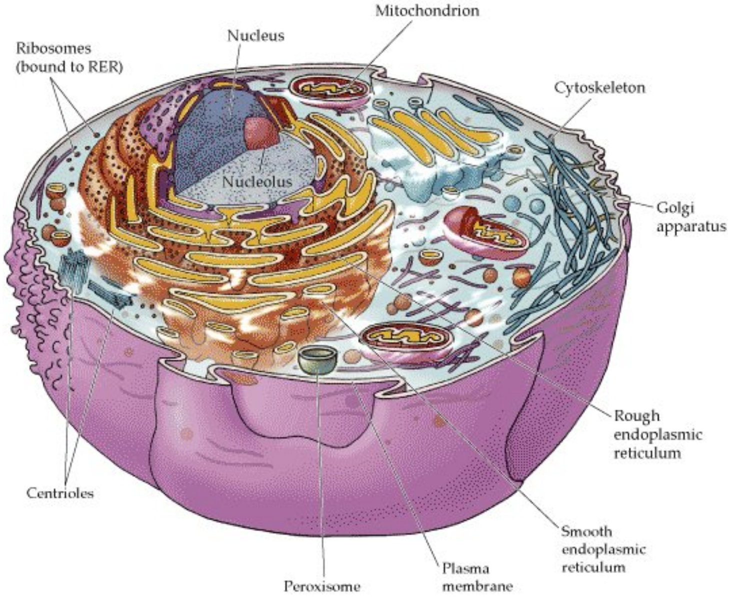

Figure 1: An eukaryotic cell

Organisms may consist of just one cell or they may be multicellular; the latter type are typically organized into tissues, which are groups of similar cells arranged so as to perform a specific function. (For example, humans have on the order of 10^{14} cells organized into roughly 200 tissues.)

One may view cell life as a collection of "wireless networks" of interactions among proteins, RNA, DNA, and small molecules involved in signaling and energy transfer. These networks process environmental signals, induce appropriate cellular responses, and sequence internal events such as gene expression, thus allowing cells and entire organisms to perform their basic functions. These control and communication networks can be relatively simple, such as the *two-component systems* found mainly in bacteria, which are cascades connecting sensors (proteins in the cell membrane, which detect outside signals) to actuators (typically transcription factors, which direct the expression of a gene). Or they may be incredibly sophisticated, as in higher organisms, involving multiple *signal transduction pathways* in which information is relayed among enzymes through chemical reactions (for instance, phosphorylation).

In addition to their own needs for survival and reproduction, cells in multicellular organisms need additional levels of complexity in order to enable communication among cells and overall regulation, as well as to direct differentiation from a single fertilized egg into the various tissues in an individual member of a species. We will focus on intracellular pathways, but these other aspects are no less exciting areas of study.

Before providing more details and examples, let us step back and review some of the basic concepts and terminology.

1.1 Prokaryotes, Eukaryotes, Archaea, and Viruses

At the highest level, biologists classify life forms into prokaryotes, and eukaryotes. *Prokaryotes* are organisms whose cells (Figure 2) do not have a nucleus nor other well-defined compartments; their genetic information is stored in chromosomes –typically circular– as well as in smaller circular DNA molecules called plasmids.

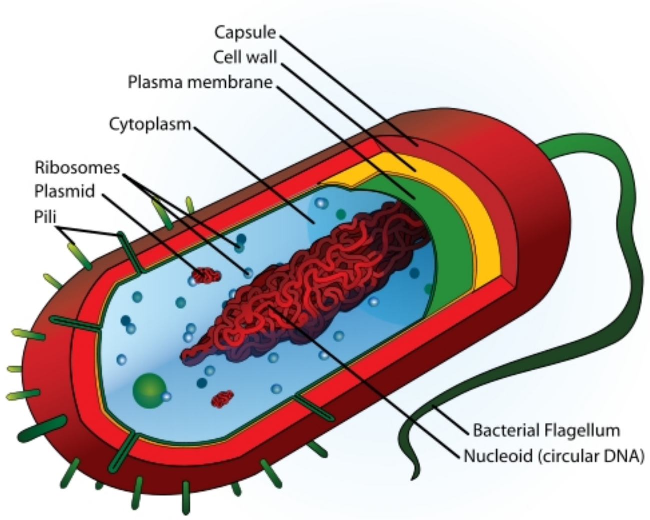

Figure 2: A prokaryotic cell

Eukaryotes have cells with organized compartments; their genetic material is stored in chromosomes – typically linear– that lie in the nucleus. Most prokaryotes, with few exceptions, are unicellular, and often sub-classified into bacteria and archaea. Eukaryotes might be unicellular (e.g., yeast) or multicellular (e.g., plants and animals). *Archaea* are single-celled microorganisms discovered in the mid-1970s and are considered by some as a third life form. They share characteristics with both prokaryotes and eukaryotes (structurally they resemble bacteria, but gene expression is similar to eukaryotes).

Eukaryotic cells are enclosed in a *plasma membrane*, which is made up of lipids and also contains proteins and carbohydrates, and acts as a protective barrier and gatekeeper, permitting only selected chemicals to enter and leave the cell. (In addition to membranes, plant cells also have a rigid cell wall.) Their interior is called the cytoplasm, and many types of organelles —specialized compartments— populate the cell (mitochondria, responsible for energy production through metabolism, and containing a very small amount of DNA; chloroplasts for photosynthesis; ribosomes, responsible for protein synthesis, and made up themselves of proteins and RNAs; endoplasmic reticulum; and so forth). The cytoskeleton, made up of microtubules and filaments, gives shape to the cell and plays a role in intracell substance transport. Prokaryotic cells, on the other hand, are surrounded by a membrane and cell wall, but do not contain the usual organelles.

Viruses consist of protein-coated DNA or RNA, and are not usually classified as living organisms, because they cannot reproduce by themselves, but rather require the machinery of a host cell in order to replicate. In particular, bacteriophages are viruses that infect bacteria.

1.2 Genomics and Proteomics

Research in molecular biology, genomics, and proteomics has produced, and will continue to produce, a wealth of data describing the elementary components of intracellular networks as well as detailed mappings of their pathways and environmental conditions required for activation.

DNA and Genes

The *genome*, that is to say, the genetic information of an individual, is encoded in double-stranded *deoxyribonucleic acid (DNA)* molecules, which are arranged into chromosomes. It may be viewed as a "parts list" which describes all the proteins that are potentially present in every cell of a given organism. Genomics research has as its objective the complete decoding of this information, both the parts common for a species as a whole and the cataloging of differences among individual members.

The key paradigm of molecular biology: "DNA makes RNA, RNA makes protein, and proteins make the cell" is called the *central dogma of molecular biology* (Crick, 1958).[1] A separate process, *replication*, occurs more rarely, and only when a cell is ready to divide (S phase of mitosis, in eukaryotes), and results in the duplication of the DNA, one copy to be part of each of the two daughter cells. See Figure 3. The term *gene expression* refers to the process by which genetic information gets ultimately transformed into working proteins. The main steps are transcription from DNA to RNA, translation from RNA to linear amino acid sequences, and folding of these into functional proteins, but several intermediate editing steps usually take place as well. (Sometimes the term "gene expression" is used only for the transcription part of this process.) At any given time, and in any given cell of an organism, thousands of genes and their products (RNA, proteins) actively participate in an orchestrated manner.

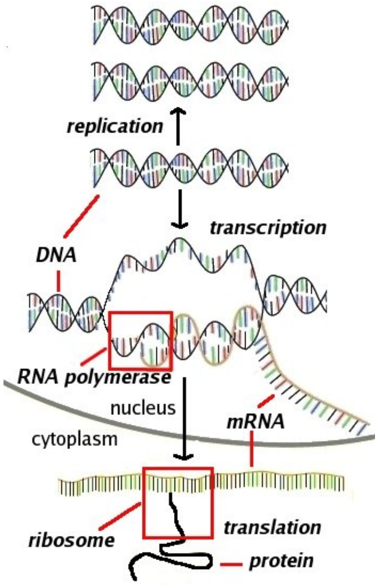

Figure 3: Central dogma of molecular biology

The DNA molecule is a double-stranded helix made of a sugar-phosphate backbone and nucleotide bases (Figure 4). Each strand carries the same information, which is encoded in the 4-letter alphabet $\{A, T, C, G\}$ (the nucleotides Adenine, Thymine, Cytosine, and Guanine), in a "complementary" form (A in one strand corresponds to T in the other, and C to G). The two strands are held together by hydrogen bonds between the

[1] We discuss limitations later

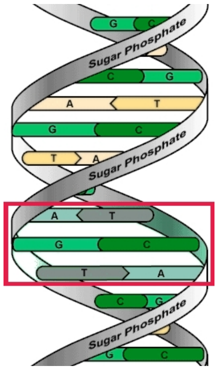

Figure 4: DNA; a codon shown in box

bases, which gives stability but can be broken-up for replication or transcription. One describes the letters in DNA by a linear sequence such as:

`gcacgagtaaacatgcacttcccaggccacagcagcaagaaggaggaatc...`

and genes (instructions that code for proteins) are substrings of the complete DNA sequence. (Besides genes, there are regulatory and start/stop regions that help delimit genes as well as determine if and when they should be "active". In addition, there are also regions that have other roles, such as coding for RNA that may not lead to proteins.) Because of its double-stranded nature, DNA is chemically stable, and serves as a good depository of information. One might think of DNA storage as a "hard disk" in a vague computing analogy.

Genomes vary in length; here are some typical sizes:

E.coli $\approx 5.44 \cdot 10^6$ bp, 5,400 genes

S.cerevisiae $\approx 1.2 \cdot 10^7$ bp, 5,800 genes

Drosophila $\approx 1.22 \cdot 10^8$ bp, 13,400 genes

C.elegans $\approx 10^8$ bp, 19,400 genes

Humans $\approx 3.3 \cdot 10^9$ bp, 20-25,000 genes

Arabidopsis $\approx 1.15 \cdot 10^8$ bp, 28,000 genes

RNA

The "read-out" of genetic information —bringing-in the instructions into working memory for execution, in our computer analogy— begins when DNA information is transcribed letter by letter into "RNA language." *Ribonucleic acid (RNA)* is a nucleic acid very similar to DNA, but less stable than DNA, and almost exclusively found in single-stranded form (with exceptions such as the RNA in some viruses). RNA language is basically the same as DNA's, with the minor (for us) detail that in RNA, the amino acid thymine is replaced with uracil, symbolized by the letter U. This process is known as *transcription*. The "copying-machine" is called *RNA polymerase* (Figure 5). A polymerase is, generally speaking, an *enzyme* —a type of protein that acts as a catalyst— that helps in the synthesis of nucleic acids. RNA polymerase is, thus, a polymerase that helps make RNA, more precisely *messenger RNA (mRNA)*.[2]

A *promoter region* is a part of the DNA sequence of a chromosome that is recognized by RNA polymerase. In prokaryotes, the promoter region consists of two short sequences placed respectively 35 and 10 nucleotides

[2] This description is over-simplified: in eukaryotic cells, an intermediate form of RNA called heterogeneous nuclear RNA (hnRNA) is produced first; then a process of "editing" gets rid of "introns" which are not part of the code for the desired protein, leaving the "exons" that are joined together to produce the actual mRNA, perhaps after insertion of some additional nucleotides.

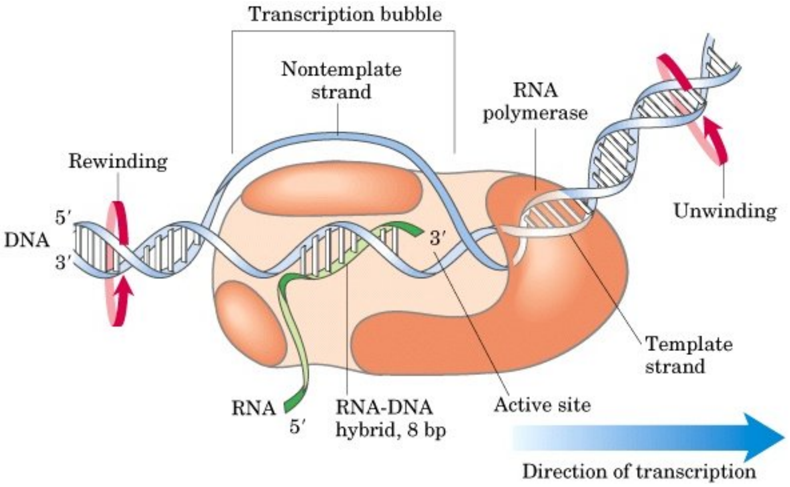

Figure 5: RNA polymerase

before the start of the gene. Eukaryotes require a far more sophisticated transcriptional control mechanism, because different genes may be only active in particular cells or tissues at particular times in an organism's life; promoters act in concert with enhancers, silencers, and other regulatory elements (Figure 6).

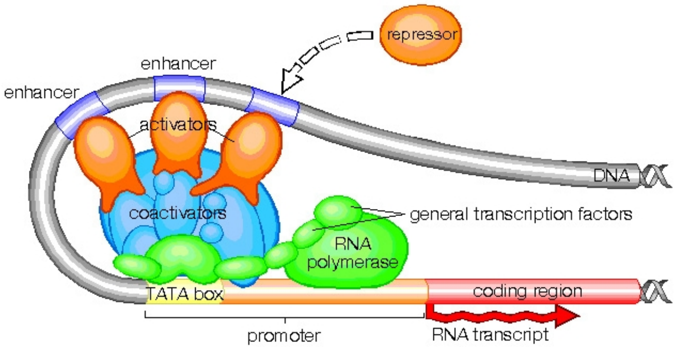

Figure 6: Transcriptional control

Proteins

Proteins are the primary components of living things. Among other roles, they form receptors that endow the cell with sensing capabilities, actuators that make muscles move (myosin, actin), detectors for the immune response, enzymes that catalyze chemical reactions, and switches that turn genes on or off. They also provide structural support, and help in the transport of smaller molecules, as well as in directing the breakdown and reassembly of other cellular elements such as lipids and sugars. Ultimately, one might say that cell life is about proteins and how and when they are produced.

After transcription, *translation* is the next step in the process of protein synthesis and it is performed at the *ribosomes* (Figure 7).

The information in the mRNA is read, and proteins are assembled out of amino acids (with the help of *transfer RNA (tRNA)*, which help bring in the specific amino acids required for each position). RNA language is translated into protein language by a mapping from strings written in the RNA alphabet $\Sigma_{\mathrm{n}} = \{U, A, G, C\}$

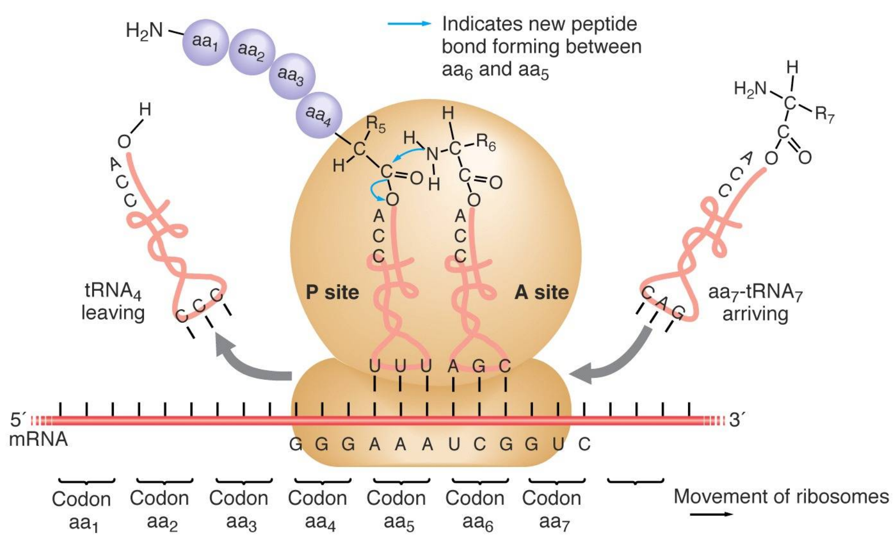

Figure 7: The Ribosome

into strings written in the amino acid alphabet:

$$\Sigma_{\mathrm{a}} = \{A, R, D, N, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V\}.$$

Every sequence of three letters in the RNA alphabet Σ_{n} is replaced by a single letter in the alphabet Σ_{a}. The genetic code explains how triplets (or *codons*, one of which is shown in Figure 4) of bases map into individual amino acids. The code, including full names and three and one-letter abbreviations, is shown in Figure 8. For example, the codon AUG translates into M (Methionine). Thus, the DNA string $TACTCATTGCGC$ would first get transcribed into the RNA string $AUGAGUAACGCG$ (note the complementation, and replacing T by U), and would be then translated into the sequence $MSNA$ (Methionine-Serine-Asparagine-Alanine) of amino acids. The string AUG codes for the amino acid Methionine but also serves as a "start" codon: the first AUG in an mRNA indicates where translation should begin.

Alanine Ala A	GCU, GCC, GCA, GCG	Leucine Leu L	UUA, UUG, CUU, CUC, CUA, CUG
Arginine Arg R	CGU, CGC, CGA, CGG, AGA, AGG	Lysine Lys K	AAA, AAG
Asparagine Asn N	AAU, AAC	Methionine Met M	AUG
Aspartic Acid Asp D	GAU, GAC	Phenylalanine Phe F	UUU, UUC
Cysteine Cys C	UGU, UGC	Proline Pro P	CCU, CCC, CCA, CCG
Glutamine Gln Q	CAA, CAG	Serine Ser S	UCU, UCC, UCA, UCG, AGU, AGC
Glutamic Acid Glu E	GAA, GAG	Threonine Thr T	ACU, ACC, ACA, ACG
Glycine Gly G	GGU, GGC, GGA, GGG	Tryptophan Trp W	UGG
Histidine His H	CAU, CAC	Tyrosine Tyr Y	UAU, UAC
Isoleucine Ile I	AUU, AUC, AUA	Valine Val V	GUU, GUC, GUA, GUG
START	AUG, GUG	STOP	UAG, UGA, UAA

Figure 8: Genetic code

The *shape* of a protein is what largely determines its function, because proteins interact with each other, and with DNA and metabolites, through lego-like fitting of parts in lock and key fashion, transfer of small molecules, or enzymatic activation. Therefore, the elucidation of the three-dimensional *structure* of proteins is a central goal in biochemical research; this subject is studied in the fields of *proteomics* and *structural biology*. The *Protein Data Bank* (http://www.rcsb.org/index.html) based at Rutgers University, USA, serves as an online catalog of protein structures. Sometimes, protein structure can be gleaned through physical methods, such as X-ray crystallography or NMR spectroscopy. Very often, however, the structure of a protein P can only be estimated, based upon a comparison with an *homologous* protein Q whose structure has been already determined (as chemists say, "solved"). One says that P and Q are homologous if they are, in an appropriate sense, close in amino acid sequence, or equivalently, in the DNA sequences for the genes coding for P and Q. One measure of closeness is Hamming distance (by how many "letters" do P and Q differ?), but more sophisticated measures used in practice include allowance for deletions and insertions of letters in P and Q. The rationale behind homology-based protein shape determination is that homologous proteins probably share a common evolutionary or developmental ancestry, and hence perform similar functions. Mathematical methods of computational biology (bioinformatics) play a central role in homology approaches; the *critical assessment of structure prediction methods (CASP)* competition compares methods from different researchers. Yet another set of techniques for elucidating the shape of proteins from their description as a linear sequence of amino acids is that of *energy minimization methods*. One views the protein-folding process as a gradient dynamical system, of which steady states are the stable configurations. This method is very difficult to apply, because of the complexity of the energy function, but has been useful for comparatively small proteins.

After translation, proteins are typically subjected to *post-translational modifications*, such as the addition of phosphate or methyl groups, or, in eukaryotic cells, *ubiquitination*, the process by which a protein is inactivated by attaching ubiquitin to it. Ubiquitin is a protein whose function is to mark other proteins for *proteolysis* (degradation), a process which occurs at the *proteasome*.

One of the key properties of proteins is that their shape (conformation) can be modified in a predictable fashion, as the consequence of interactions with other molecules. One often says that the protein has been "activated" as a result of such an interaction. For instance, Figure 9 shows, in schematic form, two conformations of the recoverin protein, the second of which comes about when two calcium ions have been inserted at appropriate places (white balls). Notice how the insertion of these ions makes an "arm" swing out. Depending on the position (extended or not) of this arm, different interactions of this protein with other players in the cell will occur.

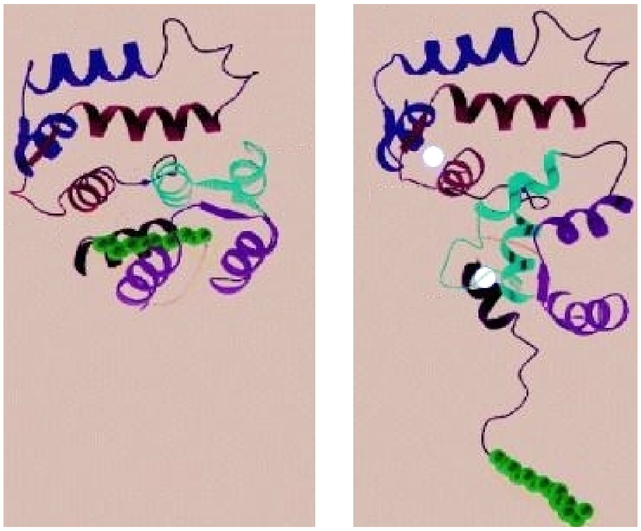

Figure 9: A protein in two conformations. Left one is Ca^{2+}-free. Right one is Ca^{2+}-bound

1.3 The Central Dogma revised

Recent work is forcing a rethinking of the roles of RNA and proteins.

For example, prions appear to take advantage of a direct mechanism for protein replication: when a prion infects an organism, it interacts with wild-type –that is to say, normal– proteins, causing them to change their shape.

For another example, until recently, RNA was not believed to be a direct player in cell control mechanisms, but now it is known that double-stranded RNA's (dsRNA's) can act, through RNA interference (RNAi) mechanism, to disrupt ("turn-off" or "silence") genes. RNA interference (RNAi) controls transcription by cleaving targeted mRNA's. Two examples are si("small interfering")RNA's and mi("micro")RNA's, which are similar, but are transcribed from the genome

We also remark the phenomenon, in eukaryotes, of "alternative splicing" that provides different proteins from the same gene, and the current interest in *epigenetic* changes such as histone modification and DNA methylation.

However, the central dogma remains the organizing principle: as usual in biology, the only general "theorem" is that every general fact has exceptions!

1.4 Proteins act as Sensors, Signal Relayers, and Actuators

Conformation changes in proteins typically happen in response to intracellular or extracellular ligand binding events, or because of binding with other proteins. (To *bind* means to reversibly join; *ligands* are small molecules that bind with larger molecules, typically proteins.) Two noteworthy instances of activation are provided by receptors and by phosphorylation reactions.

Receptors are proteins that act as the cell's sensors of outside conditions, relaying information to the inside of the cell. A receptor is typically made up of three parts. The *extracellular domain* ("domains" are parts of a protein) is exposed to the exterior of the cell. Extracellular ligands, such as growth factors and hormones, bind to receptors, most of which are designed to recognize a specific type of ligand. The *transmembrane domain* serves to "anchor" the receptor to the membrane. Finally, a *cytoplasmic domain* helps initiate reactions inside the cell in response to exterior signals, by interacting with other proteins. There is a special class of receptors which constitute a common target of pharmaceutical drugs: *G-protein-coupled receptors (GPCR's)* (Figure 10). The name of these receptors arises from the fact that, when their conformation changes in response to a ligand binding event, they activate G-proteins, so called because they employ *guanine triphosphate* and *diphosphate (GTP* and *GDP)* in their activity. GPCR's are made up of several subunits (G_α, G_β, G_γ) and are involved in the detection of metabolites, odorants, hormones, neurotransmitters, and even light (rhodopsin, a visual pigment).

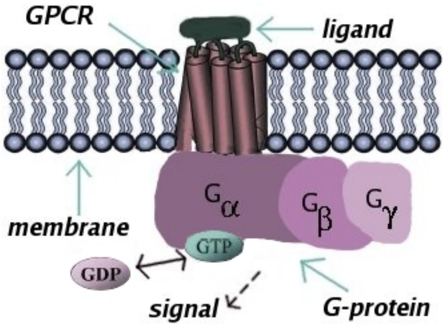

Figure 10: G-protein-coupled receptor and G-protein

Another example of activation is *phosphorylation*. *Adenosine triphosphate (ATP)* is a nucleotide that is the major "energy currency" of the cell. When one of its phosphate groups is used up, there remains ADP (*di*phosphate). ATP is manufactured by "recharging" ADP through the *citric acid cycle* or *Krebs cycle*, which is an enzyme-catalysed chemical reaction of central importance in living cells that use oxygen as part of cellular respiration (Figure 11). In eukaryotes, this process takes place in mitochondria (although in plants, this recharging also occurs in chloroplasts). In prokaryotes ATP is produced both in the cell wall and in the cytosol by the glycolysis process. Human cells contain approximately one billion ATP molecules, which are only sufficient for a few minutes of life.

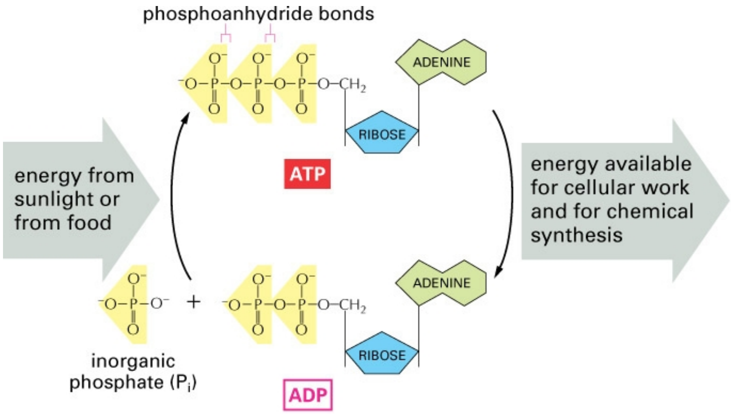
phosphoanhydride bonds
ADENINE
RIBOSE
ATP
energy from sunlight or from food
energy available for cellular work and for chemical synthesis
ADENINE
RIBOSE
inorganic phosphate (P_i)
ADP

Figure 11: ATP production

An *enzyme* is a protein that catalyzes a chemical reaction. Phosphorylation is a chemical reaction in which an enzyme X —called a *kinase* when playing this role— transfers a phosphate group (PO_4) from a "donor" molecule such as ATP to another protein Y, which becomes "activated" in the sense that its energy is increased. Once activated, protein Y may then influence other cellular components, including other proteins, itself acting as a kinase, or it may take an appropriate shape that allows it to to bind with yet another protein or to a segment of DNA so as to initiate, enhance, or repress expression of a gene. Normally, proteins do not stay activated forever; another type of enzyme, called a *phosphatase*, eventually takes away the phosphate group; see Figure 12. In this manner, signaling is "turned off" after a while, so that the system is ready to detect new signals.

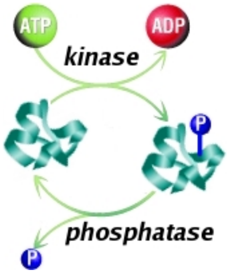

Figure 12: Phosphorylation and de-phosphorylation

Receptors and enzymatic cascades act in concert. Binding of extracellular ligands triggers signaling through a series of chemical reactions inside the cell, carried out by enzymes and often relayed by smaller molecules called *second messengers*. In this manner, regulatory pathways can be either turned "on" and "off" or modulated, and transcription of particular sets of genes may be started and stopped in response to environmental conditions. Figure 13 (from www.cellsignal.com) illustrates one such pathway, which involves GPCR activation as well as signaling through a MAPK cascade (more on MAPK cascades below).

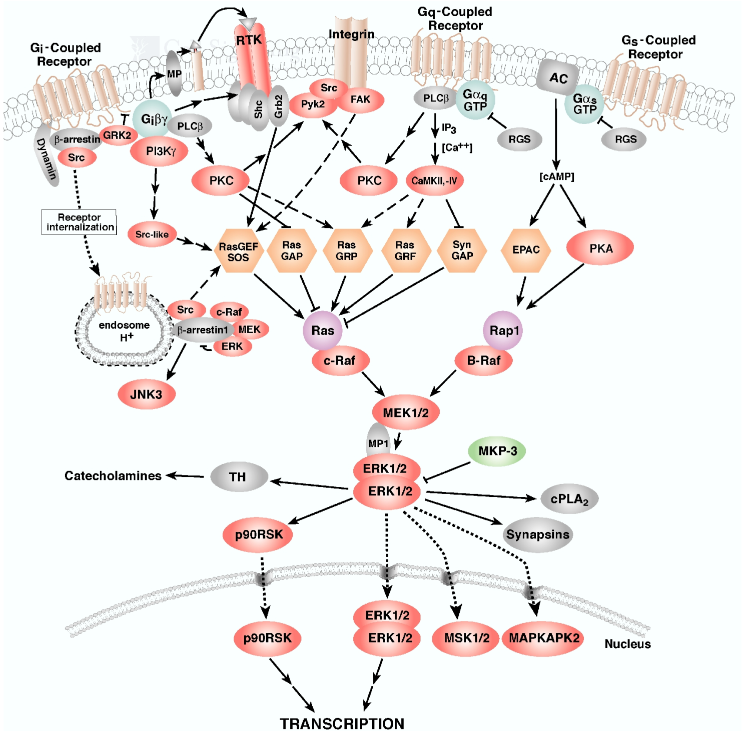

Figure 13: A GPCR pathway

The animation at *http://biocreations.com/pages/mapk.html* is strongly recommended as an illustration of signaling pathways.

As another illustration, consider the diagram shown in Figure 14, extracted from a well-known paper on cancer research [3] which describes the top-level schematics of a wiring diagram of signaling circuitry in the mammalian cell. The illustration shows the main signaling pathways for growth, differentiation, and apoptosis (commands which instruct the cell to die). Highlighted in red are some of the genes known to be functionally altered in cancer cells. Of course, such a figure, compared for example with the more detailed biochemical pathway shown in Figure 13, leaves out a lot of information, some known but omitted for simplicity, and some unknown. Much of the system has not been identified yet, and the functional forms of the interactions, much less parameters, are only very approximately known. However, data of this type are being collected at an amazing rate, and better and better models are being obtained constantly.

A particularly important role is played by ErbB receptor pathways (Figure 15 is from Yarden & Sliwkowski, Nature Rev Mol Cell Biol. 2001).

Inputs are hormones, neurotransmitters, lymphokines, and other signals EGF-family ligands

The recently marketed drug Herceptin is an antibody that blocks ErbB2 in breast cancer; many other drugs targeting this pathway are in advanced clinical testing.

The above examples were from eukaryotes. We now turn to one from a prokaryote. *Chemotaxis* is the term used to describe movement, in bacteria as well as other organisms, in response to chemoattractants or repellants, such as nutrients and poisons, respectively. *E. coli* bacteria (Figure 16) are single-celled organisms, about 2 μm long, which possess up to six flagella for movement. Chemotaxis in *E. coli* has been studied extensively. These bacteria can move in basically two modes: a “tumble” mode in which flagella turn clockwise and reorientation

[3] Hanahan, D., Weinberg, R.A., *The hallmarks of cancer*, Cell 100(2000): 57–70.

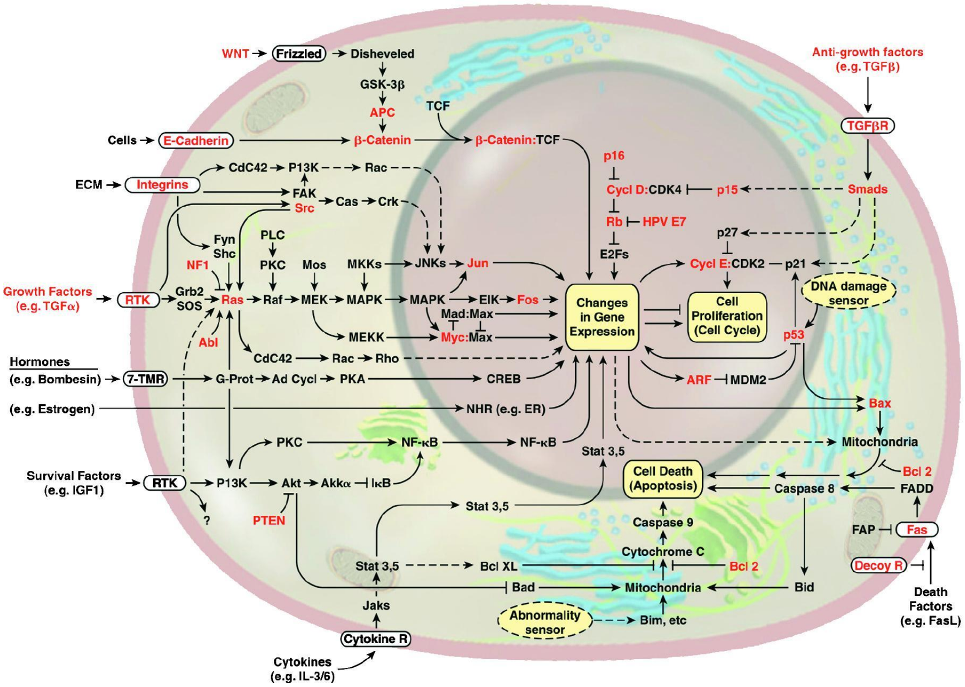

Figure 14: Signaling circuitry of the mammalian cell

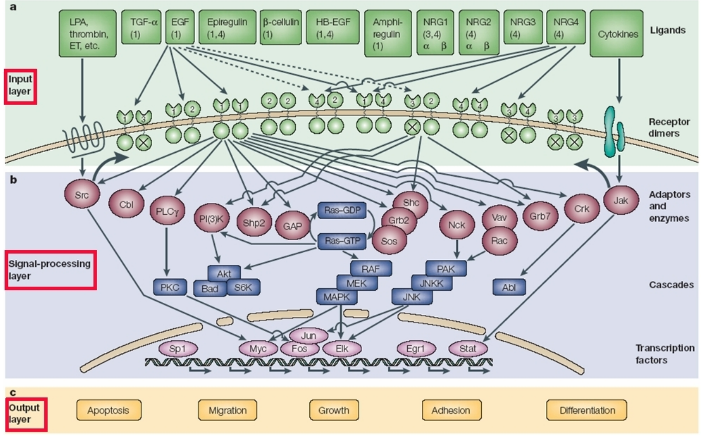

Figure 15: ErbB pathways

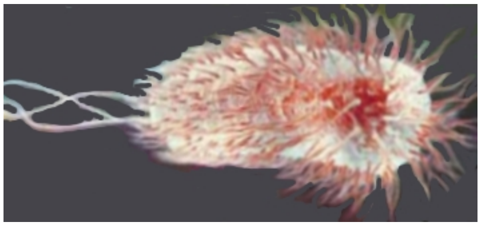

Figure 16: *E. coli* bacterium

occurs (Figure 17, left), or a "run" mode in which flagella turn counterclockwise, forming a bundle which helps propel them forward (Figure 17, right). The motors actuating the flagella are made up of several proteins. In the terms used by H. Berg[4], they constitute "a nanotechnologist's dream," consisting as they do of "engines, propellers, . . . , particle counters, rate meters, [and] gear boxes." Figure 18 shows an actual electron micrograph and a schematic diagram of a flagellar motor. The signaling pathways involved in *E. coli* chemotaxis are fairly well understood. Aspartate or other nutrients bind to receptors, reducing the rate at which a protein called CheA ("Che" for "chemotaxis") phosphorylates another protein called CheY transforming it into CheY-P. A third protein, called CheZ, continuously reverses this phosphorylation; thus, when ligand is present, there is less CheY-P and more CheY. Normally, CheY-P binds to the base of the motor, helping clockwise movement and hence tumbling, so the lower concentration of CheY-P has the effect of less tumbling and more running (presumably, in the direction of the nutrient). A separate feedback loop, which includes two other proteins, CheR and CheB, causes adaptation to constant nutrient concentrations, resulting in a resumption of tumbling and consequent re-orientation. In this manner, *E. coli* performs a stochastic gradient search in a nutrient-potential landscape. Figure 19 shows a schematic diagram of the system responsible for chemotaxis in *E. coli*.

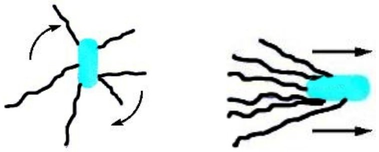

Figure 17: *E. coli* tumbling: flaggela apart. Running: flaggela in bundle

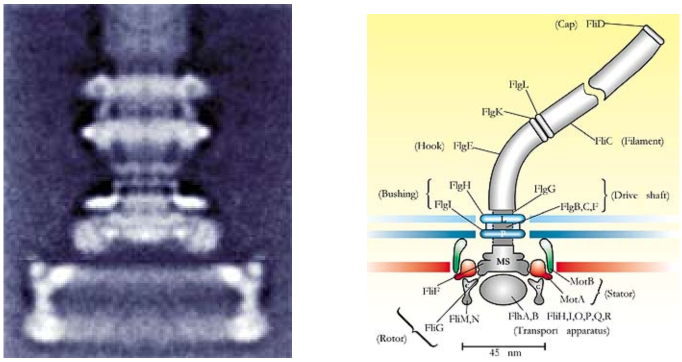

Figure 18: Electron micrograph and diagram of flagellar motor, from Berg's paper

1.5 Measurement Techniques

Massive amounts of data are being generated by genomics and proteomics projects, thanks to sophisticated genetic engineering tools (gene knock-outs and insertions, PCR) and measurement technologies (fluorescent proteins, microarrays, blotting, FRET). *Polymerase chain reaction (PCR)* is a technique that amplifies DNA (typically a gene or part of a gene). Creating multiple copies of a piece of DNA, which would otherwise be present in too small a quantity to detect, PCR enables the use of measurement techniques. Let us briefly discuss a couple of these measurement technologies, in order to provide an idea of their power as well as their severe limitations.

[4] H.C. Berg, *Motile behavior of bacteria*, Physics Today, January 2000

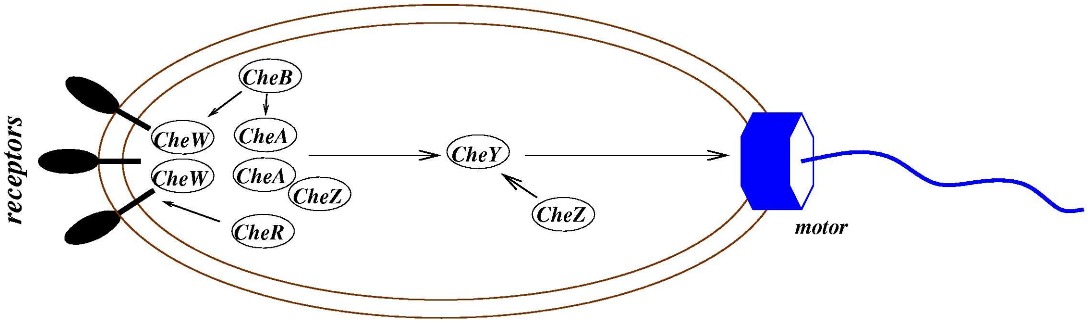

Figure 19: *E. coli* chemotactic circuit

Suppose that we wish to know at what rate a certain gene X is being transcribed under a particular set of conditions in which the cell finds itself. Fluorescent proteins may be used for that purpose. For instance, *green fluorescent protein (GFP)* is a protein with the property that it fluoresces in green when exposed to UV light. It is produced by the jellyfish *Aequoria victoria*, and its gene has been isolated so that it can be used as a *reporter gene*. The GFP gene is inserted (cloned) into the chromosome, adjacent to or very close to the location of gene X, so both are controlled by the same promoter region. Thus, gene X and GFP are transcribed simultaneously and then translated (Figure 20), and so by measuring the intensity of the GFP light emitted one can estimate how much of X is being expressed.

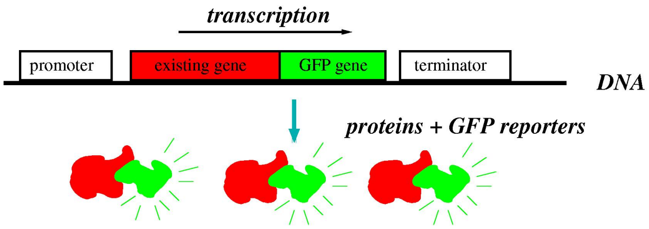

Figure 20: GFP

Fluorescent protein methods are particularly useful when combined with *flow cytometry* (Figure 21). Flow Cytometry devices can be used to sort individual cells into different groups, on the basis of characteristics such as cell size, shape, or amount of measured fluorescence, and at rates of up to thousands of cells per second. In this manner, it is possible, for instance, to count how many cells in a population express a particular gene under a specific set of conditions, and to study *stochasticity* of gene expression.

A set of technologies collectively referred to as *gene arrays* (DNA chips, DNA microarrays, Affymatrix gene chips) provide high-throughput methods for simultaneously monitoring the activity levels of thousands of genes, thus providing a snapshot of the current gene expression activity of a cell (Figure 22). An array is built using robotics and imaging equipment, very much as in electronic chip fabrication. The array has in each location (i, j) a detector "tuned" to a particular gene or small sequence of nucleotides X_{ij}. This detector (the usual name is a "target") is the complement $\overline{\mathrm{X}}_{ij}$ of X_{ij} or, more likely, of a subsequence of X_{ij}. (More precisely, one wants to find out how much of a specific X's mRNA is being transcribed. The first step is to reverse-transcribe RNA to DNA, which becomes complementary DNA (cDNA), and then PCR-amplify it. We omit details here, since we only want to explain the basic principle.) Because of hybridization, that is, the A-T and G-C base pairings for DNA, X_{ij} should "stick" to its complement $\overline{\mathrm{X}}_{ij}$. This allows one to estimate the presence and abundance of each X_{ij} in a sample. In order to be able to read the information in the different array positions, the sequences X_{ij} being tested for are first radioactively or fluorescently tagged, so that one can simply measure how much has accumulated at each position i, j. Pattern recognition, machine learning, and

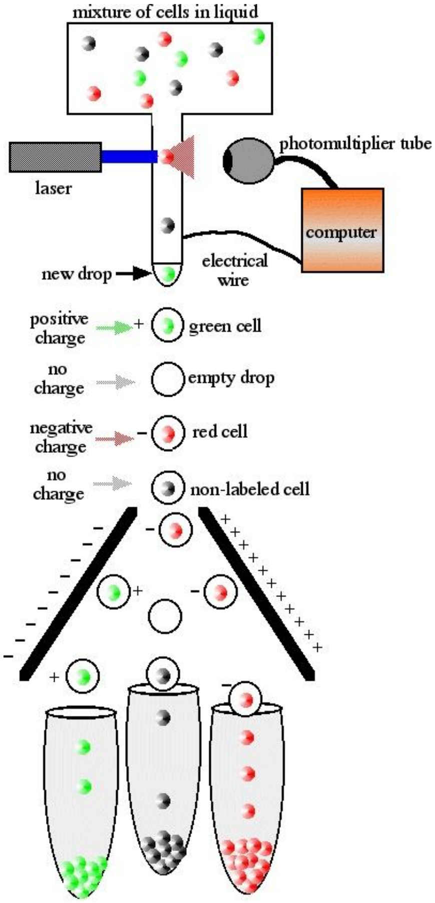

Figure 21: Fluorescence Activated Cell Sorting (FACS)

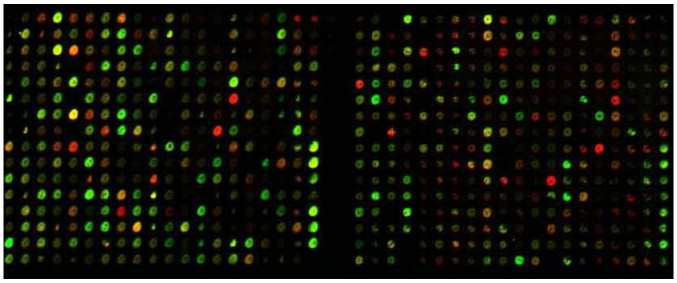

Figure 22: Gene array

control-theory tools such as clustering, Bayesian networks, and identification theory —especially when time-dependent data is available— can be and are used infer information about dynamic interactions among genes, and to sort out which particular sets of genes are triggered simultaneously or in a sequence (*co-expression* analysis) in response to different environmental factors or disease states. In control-theory language, we might think of gene arrays as giving a vector-valued output, in contrast to a technology such as GFP which provides merely a scalar value.

Actually, it is difficult to obtain absolute measurements with gene arrays, due to uncertainties in the PCR and hybridization processes. Rather, the method is often used in a comparative fashion. Gene array experiments can be done for different cell types in the same organism, for the same cell types under different experimental conditions, or even for comparing cells from two organisms, perhaps one of them having an engineered mutation of the original one. A fascinating application is the comparison of abnormal (e.g., cancerous) and normal cells, obtaining in that manner a gene expression "signature" that might be used for diagnosis.

A *Western blot* allows one to detect the presence of a specific protein, or a small number of them, in a sample taken from an experiment.[5] The proteins extracted from the sample, together with a small number of antibodies which recognize only specific proteins, are placed on membranes and allowed to interact. Different methods, for instance radioactive labeling of stains, are then used in order to visualize the results. As an example, Figure 23[6], shows Western blot data from an experiment in which three proteins (Cdc25, Wee1, and MAPK) have been observed under different conditions (concentrations 0, 25nM, etc.) of another protein named Δ65-cyclin B1, during two experiments (labeled "going up" and "coming down" in the figure). The higher placements on the blot correspond in this case to the relative abundance of the phosphorylated form of the protein; for example, phosphorylated Cdc25 is more abundant in the "100" than in the "0" lanes.

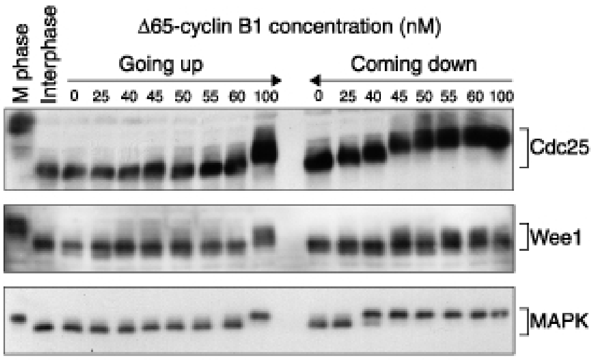

Figure 23: Western blots

1.6 Limitations

Notwithstanding the power of the techniques just described, GFP, arrays, and blots, they are intrinsically noisy, because of chemical interactions in blots, production errors in arrays, or other sources of interference. In addition, the resulting measurements have low precision: very few bits of information can be extracted from data such as that shown in Figures 22 or 23. These limitations of imprecision and noise are sometimes ignored in systems biology modeling, but it is obviously pointless to try to tightly fit model parameters to such data. On the other hand, for certain types of quantities, such as the amount of calcium in a cell, currents through channels, or certain enzyme concentrations, there are other techniques that may result in higher precision measurements. In such cases, parameter fitting is more reasonable.

The field suffers from what has been called a *data-rich/data-poor* paradox: while on the one hand a huge amount of *qualitative* network (schematic modeling) knowledge is available, as evidenced by figures such as 13 and 14, on the other hand little of this knowledge is *quantitative*, at least at the level of precision demanded by most mathematical tools of analysis. The problem of exploiting this qualitative knowledge, and effectively integrating relatively sparse quantitative data, is among the most challenging issues confronting systems biology.

[5] "Southern" blots are techniques for detecting DNA, and "Northern" blots for detecting RNA. The names originated with the first of these, which was developed by a UK biologist named Southern.

[6] taken from Pomerening, J.R., Sontag, E.D., Ferrell Jr., J.E., *Building a cell cycle oscillator: hysteresis and bistability in the activation of Cdc2* Nature Cell Biology, 5(2003): 346–351.

1.7 Model Organisms

Since many organisms follow the same basic principles, biologists have concentrated on a small number of *model systems*. This allows them to focus on specific systems, easing comparisons and facilitating sharing of research results. Different aspects may be easier to study in different model organisms (embryonic cycles in frog eggs, differentiation and development in flies, aging in worms), by taking advantage of fast breeding or speed of maturation.

As cataloged in the US National Institutes of Health website (http://www.nih.gov/science/models), there are the bacterium *E. coli* (Figure 24), the mammalian models mouse (Figure 27) and rat, and the main non-mammalian models *S. cerevisiae* (Figure 29) (budding yeast), *Neurospora* (filamentous fungus), *D. discoideum* (social amoebae), *C. elegans* (Figure 25) (round worm), *D. melanogaster* (Figure 26) (fruit fly), *D. rerio* (zebrafish), and *Xenopus* (Figure 28) (frog). In addition, a popular plant model is *Arabidopsis* (Figure 30) (a small flowering plant, member of the mustard family).

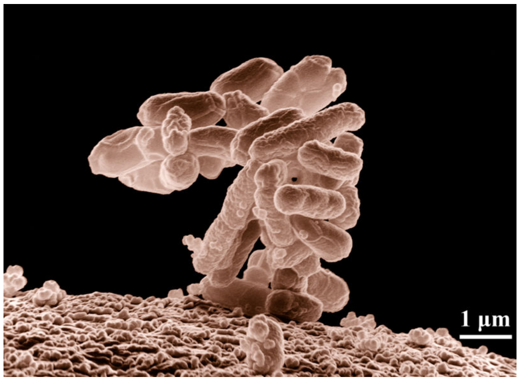

Figure 24: Ecoli

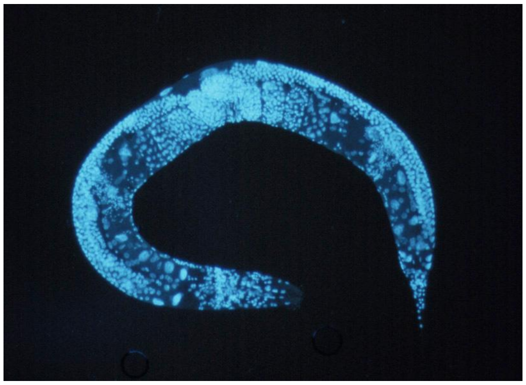

Figure 25: Caenorhabditis Elegans

2 Synthetic Biology

Section being written.

Figure 26: Drosophila Melanogaster

Figure 27: Mice

Figure 28: Xenopus (frog)

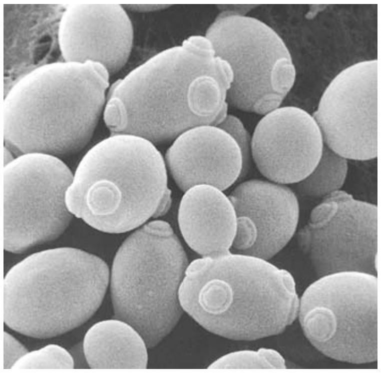

Figure 29: (Budding = Bakers's) Yeast

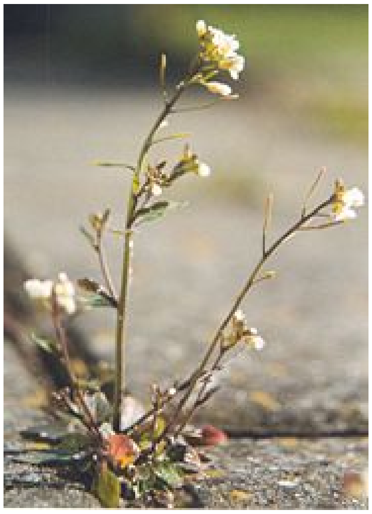

Figure 30: Arabidopsis Thaliana