2 Molecular biology of gene regulation

Entering the post-genomic era where sequencing of whole genomes has become routine work, we have the opportunity to unravel a plethora of functional elements, which ultimately give rise to the complexity of life. We might one day understand the function and interactions of all constituents of a given genome by studying subsets of these constituents. This thesis is about computational approaches to mine for sequence elements that are vital components of gene regulation in higher organisms like mammals. Other issues that are addressed include building up a repository for the genome-wide annotation of regulatory elements and the exploration of associations between predicted patterns of regulatory elements and measured gene expression levels. We firstly give some biological background on the elements of a eukaryotic genome (Section 2.1). Then, we cover in greater detail the DNA and protein components of gene regulation in proximal promoter regions (Section 2.2). Subsequently, these components will be briefly reviewed from an evolutionary viewpoint (Section 2.3). This chapter closes with a summary of selected large-scale methods to study gene regulation experimentally (Section 2.4). We also encourage the reader to consult textbooks on molecular biology of eukaryotes for further details, which we cannot give here. Alberts et al. (1994) is a remarkably well written volume for broadening one's horizon in molecular biology.

2.1 Genome biology

Since our research focuses on eukaryotic organisms, we present a brief overview on the key features of eukaryotic genomes. Unlike bacterial cells, eukaryotic cells contain a nucleus that accommodates the chromosomes. The genome of an eukaryotic organism is distributed across chromosomes. Chromosomes are composed of linear double-stranded DNA, which is the carrier of genetic information, and accessory structural proteins (Kornberg, 1974). Different kinds of organisms have different numbers of chromosomes. Humans have 23 pairs of chromosomes, 46 in all: 44 autosomes and two sex chromosomes. Each parent contributes one chromosome to each pair, so children get half of their chromosomes from their mothers and half from their fathers. The different levels of structural organization of the genome are depicted in Figure 2.1

where a replicated chromosome with two sister chromatids is shown. The genetic material of cells only fully condenses upon entering the process of cell division. Resting cells show a varying pattern of "active" relaxed chromatin and "silent" condensed chromatin. The terms "active" and "silent" refer to the process of gene transcription, which will be explained later on.

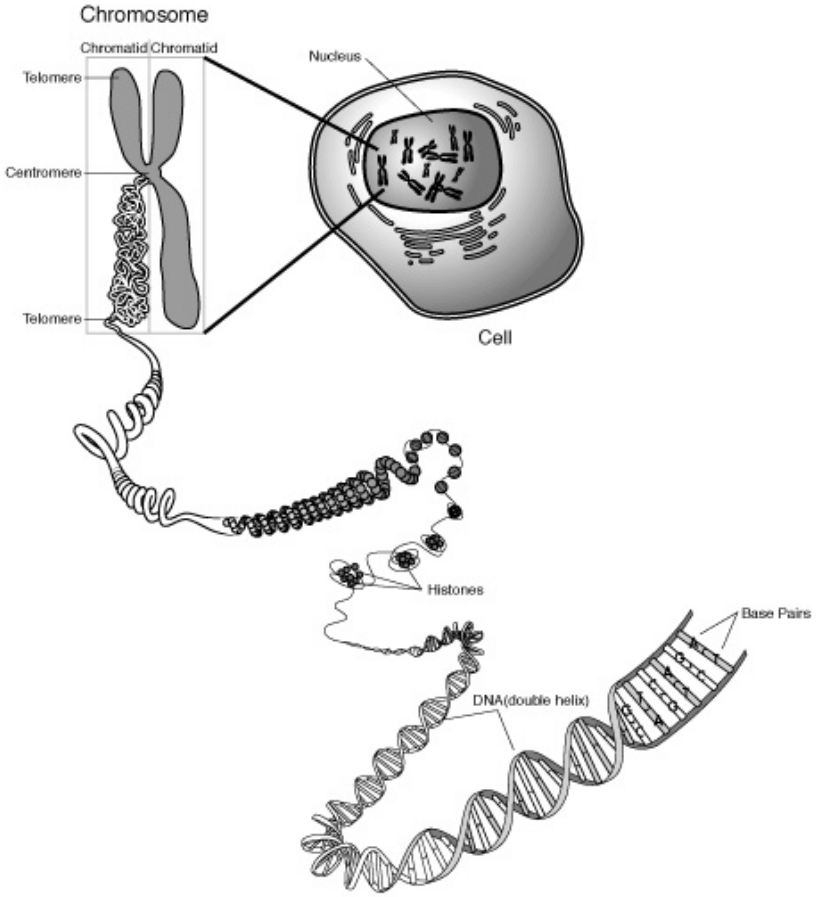

Figure 2.1: Structural organization of the genome. The different levels of DNA packing are presented for a replicated chromosome consisting of two identical copies (sister chromatids). Image reproduced with permission from http://www.accessexcellence.org, The National Health Museum, USA.

2.1.1 DNA - the carrier of genetic information

The two key components of chromosomes were identified decades ago. Deoxyribonucleic acid (DNA) carries genetic information, which constitute the blueprint of life.

DNA is wrapped around histone proteins. Both taken together from complexes called nucleosomes. Gene regulation acts on both protein and DNA components by myriad interactions and modifications. In 1944, Avery et al. (1944) published their groundbreaking work on DNA being the carrier of genetic information. They showed that harmless bacteria can be turned into deadly ones by the transfection of DNA. Later, Watson and Crick (1953) succeeded in elucidating the molecular structure of DNA.

Genomic DNA forms a double-helix and consists of two antiparallel complementary strands. Each strand is a directional linear polymer of four types of nucleotides or bases (adenine `A`, cytosine `C`, guanine `G`, and thymine `T`), held together by a sugar-phosphate backbone. `A` and `G` belong to the class of purines (double heterocycle), whereas `C` and `T` are pyrimidines (single heterocycle, see Figure 2.2). The sugar-phosphate bonds of the DNA backbone are phosphodiester bonds linking the 5' carbon of a deoxyribose to the 3' carbon of the subsequent deoxyribose. These bonds impose a directionality on both DNA strands. DNA is only synthesized and processed from its 5' end to its 3' end. The two DNA strands are kept together by hydrogen bonds. Bases `A` and `T` pair by forming two hydrogen bonds. Three hydrogen bonds are formed between bases `C` and `G`. This fact results in a higher stability of `GC`-rich regions.

Although different species have uniquely different ratios of pyrimidines or purines, the relative concentrations of `A` always equal that of `T`, and `G` equal that of `C` (Chargaff's Law, Chargaff (1950)).

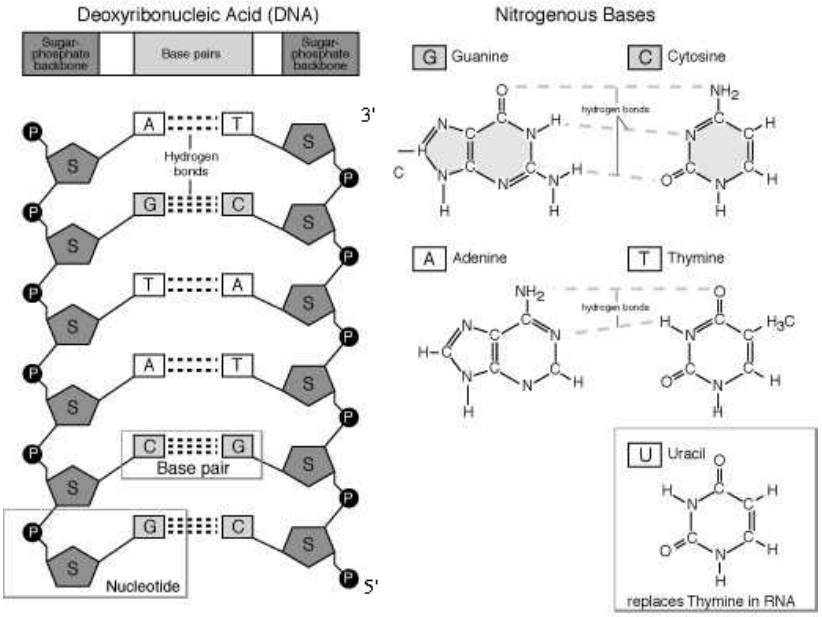

Figure 2.2: Schematic overview view of DNA structure and building blocks. Image reproduced from `http://www.accessexcellence.org`, The National Health Museum, USA.

2.1.2 Genes - entities of genetic information

We only begin to understand the information that is encoded in genomic DNA. Going back to the beginning of genetics, the well known concept of a "gene" has been established long before its molecular nature became apparent. In 1865, Gregor Mendel introduced the notion of "factors" that confer individual traits. In 1909, Wilhelm Ludvig Johannsen (1857-1927), a Danish botanist, coined the word gene (using Greek "to give birth to") with reference to Mendel's "factors".

Today, we think of a gene as being the functional and physical unit of heredity passed from parent to offspring. Intriguingly, a surprisingly small proportion of the human genome is made up of genes (only up to 1.5 % are covered by protein-coding genes). Nevertheless, it is the set of genes and their interactions that define all living beings. Eucaryotic genes map to individual sequence ranges on genomic DNA. Their structure is given by the following elements (listed from 5' to 3'):

Promoter The part of a gene that contains the information to turn the gene on or off. The process of transcription is initiated at the promoter. The extent of a promoter is often difficult to determine. Proximal and distal promoter elements play a role in controlling the expression level of a gene.

Exons Regions downstream of the promoter of a gene that are transcribed and exported from the nucleus as part of the **messenger RNA (mRNA)**. mRNA contains all information ("the message") for the formation of the final protein product of a gene.

Introns Regions downstream of the promoter of a gene that are also transcribed into RNA but are excised (spliced) from the maturating RNA. Thus, these regions are absent from the **messenger RNA**.

The readout of a gene is split into three steps (see Figure 2.3). These three steps, taken together with the process of DNA replication, are often referred to as "central dogma" of molecular biology:

1. **Transcription** The process of transcription starts in the promoter region of a gene. Promoter elements support the buildup of the RNA polymerase machinery. A full-length RNA copy of the genomic DNA including exons and introns is generated. Transcription termination is linked to 3' polyadenylation of the transcript and involves transitions at the 3' end of genes that may include an exchange of elongation and polyadenylation/termination factors.

2. **RNA processing** The transcribed product of a gene, nuclear RNA, is subject to further modifications.

   A capping component is added to the 5' end and a Poly-`A` tail to the 3' end. The nuclear RNA is additionally shortened by the excision of introns and occasionally

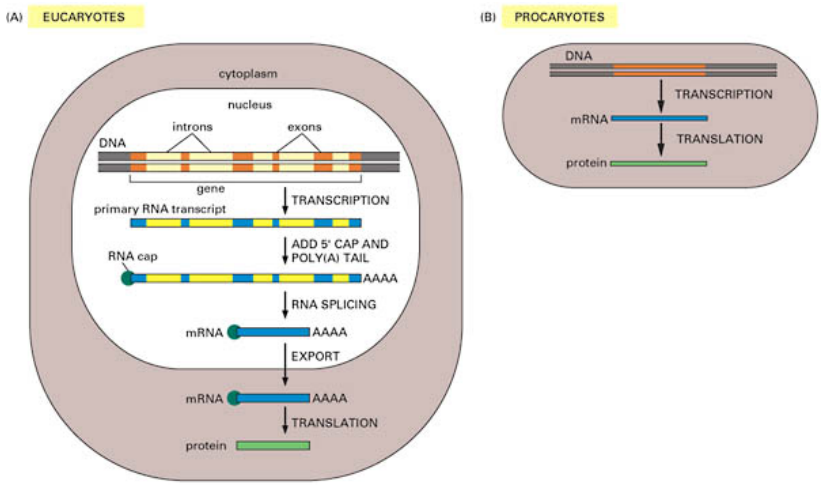

Figure 2.3: (A) In eucaryotic cells, the initial RNA molecule produced by transcription (the primary transcript) contains both intron and exon sequences. Its two ends are modified, and the introns are removed by an enzymatically catalyzed RNA splicing reaction. The resulting mRNA is then transported from the nucleus to the cytoplasm, where it is translated into protein. (B) In procaryotes, the production of mRNA molecules is simpler. The 5' end of an mRNA molecule is produced by the initiation of transcription by RNA polymerase, and the 3' end is produced by the termination of transcription. Since procaryotic cells lack a nucleus, transcription and translation take place in a common compartment. Reproduced with permission from http://www.accessexcellence.org

further exons. This step is known as "splicing" and leads to the diversity of transcripts by facilitating various exon combinations. Recent evidence suggest a coupling of RNA processing to transcription (Moteki and Price, 2002).

3. **Translation** The process of translation takes place after the export of the mature mRNA from the nucleus. Translation means to transfer protein-coding information on mRNAs into actual proteins by another synthesis step. Ribosomes, which are large complexes of RNA and protein, are the factories of protein biosynthesis that utilize mRNA as a template.

   There are sections of the mRNA before and after its start and stop sequences that are not translated. These come from the template DNA strand that the RNA was transcribed from. These regions, known as the 5'UTR and 3'UTR (five-prime and three-prime untranslated regions, respectively) code for no protein sequences. Their importance lies in the belief that the sequence of the 5' UTR and 3' UTR may, by their varying affinity for certain RNase enzymes, promote or inhibit the relative stability of the RNA molecule. Certain UTRs may allow the RNA to survive longer in the cytoplasm before being degraded, thus allowing them to produce more protein, while others may be degraded sooner, thus lasting

a shorter time and producing a smaller relative amount of protein. See Alberts et al. (1994), chapter 9 for further details.

Also, there is evidence that certain complexes within the UTRs may not only affect the stability of the molecule, but that they may promote translational efficiency or even cause inhibition of translation altogether, depending on the sequences in the UTRs.

Not all genes that are transcribed are also translated. Evidently, the most prominent example are the RNA components of translation (i.e. Transfer RNA). Transfer RNA (abbreviated tRNA) is a small RNA chain (74-93 nucleotides) that transfers a specific amino acid to a growing polypeptide chain at the ribosomal site of protein synthesis during translation. There is an increasing body of evidence that transcription occurs more frequently than anticipated for a long time (Cawley et al., 2004).

2.1.3 Strategies for gene detection

The information in mRNA can be transferred into **complementary DNA (cDNA)** by the process of reverse transcription. cDNA may be used for gene cloning or as a gene probe. Due to the instability of RNA, it is usually difficult to obtain a full-length clone of the underlying mRNA. Adams et al. (1991) entered high-throughput sequencing of cDNA clones as an affordable source of gene probes. The term **Expressed Sequence Tag (EST)** was introduced to refer to this new class of sequence, which is characterized by being short (typically about 400 - 600 bases) and relatively inaccurate (around 2% error). The use of single-pass sequencing was an important aspect of making the approach cost effective. In most cases, there is no initial attempt to identify or characterize the clones. Instead, they are identified using only the small bit of sequence data obtained, comparing it to the sequences of known genes and other ESTs. Once a genome assembly becomes available, ESTs facilitate the detection of transcribed genome parts *in silico* by similarity searches and *in vivo* by hybridization.

2.2 Gene regulation at the promotor level

Turning genes into products is extensively controlled at all processing stages. In the 1960s, two French pioneers (Jacob et al., 1960) demonstrated that genes do not continuously form their products. They showed that genes relevant to the metabolism of the bacterium *Escherichia coli* could be induced by an external stimulus. A recurring motif of gene regulation in this context are interactions of proteins and DNA.

The subsequent introduction will focus on promoter-level events since that is the major aspect of this thesis. A general introduction would go well beyond the scope of this

thesis. Figure 2.4 depicts the basic blueprint of a gene promotor. The absence or presence of DNA sequence elements has been shown to influence the initiation rate of transcription. Usually, one distinguishes between elements that act over long distances (enhancers) and proximal elements around the start site of transcription. Interactions of sequence elements as far apart as several kilobases occur in the cell due to the looping of long DNA stretches. Since long-distance interactions are difficult to study in experiment, little is known about the underlying principles. The situation is better described for proximal elements. Proximal sequence elements exert control on gene transcription by being the targets of proteins called transcription factors. However, these DNA sequence elements or binding sites are not accessible to transcription factors in tightly packed chromatin. An "open" conformation of chromatin is seen in promotor regions of actively transcribed genes. Here, the chromosomal structure is loosened by enzymes that act on chromosomal proteins. Initially, this phenomenon was observed in lampbrush chromosomes where the DNA loops out brushlike on many spots along the main axis of the chromosome.

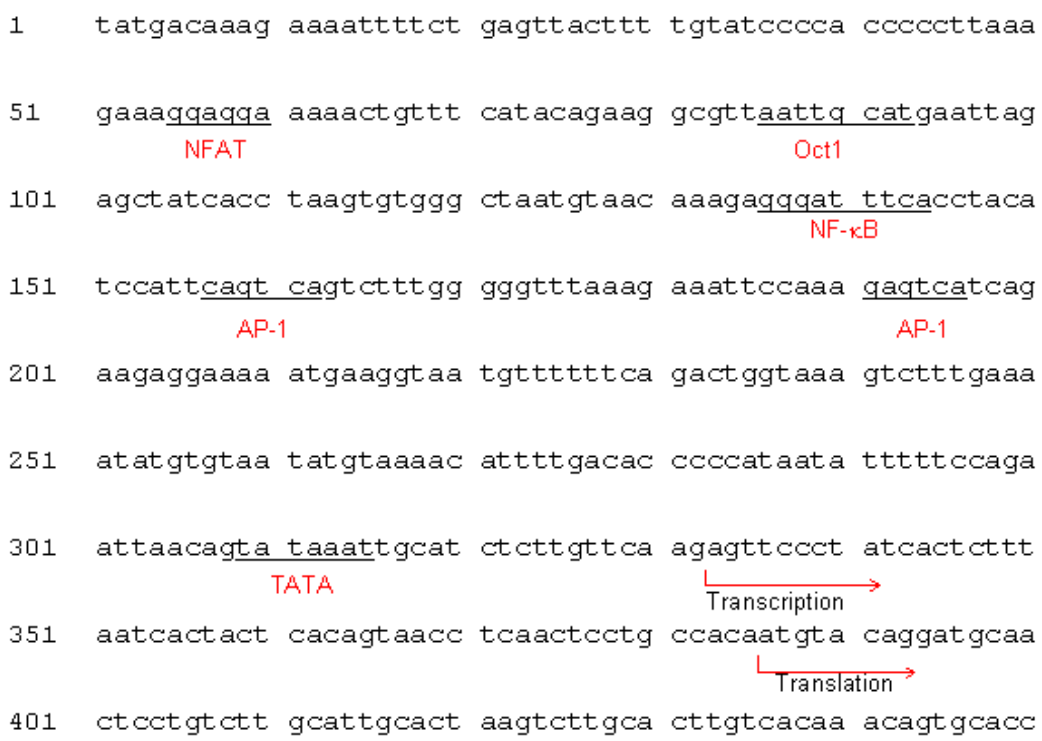

Figure 2.4: Illustration of the binding site arrangement of the human IL2 promotor (GenBank accession AF031845). Binding sites are highlighted in red. The different binding sites are recognized by distinct proteins. Starts of transcription and translation are represented by red arrows.

Transcription factors are the end points of cellular signaling pathways and often convey a stimulus from the cell's surface to the genome. In the case of the human IL2 gene (Figure 2.4), many interacting DNA/protein pairs have been elucidated. Again, binding sites can be divided into general and specific ones. A prerequisite to start transcription is satisfied if a set of binding sites is occupied and/or released by the corresponding transcription factors. The RNA polymerase II complex is then recruited to the promoter and transcription starts upon phosphorylation of the C-terminal domain of Polymerase II. As a consequence, promotors are key components of combinatorial control. A good introductory textbook on eukaryotic gene regulation is Latchman (2002).

2.2.1 Transcriptional control – DNA sequence elements

Cells have to react and adapt to various external constraints like temperature, oxygen supply or mechanical stress. Evidently, external conditions require complex responses involving the coordinate expression of many genes. Unlike in bacteria, coordinately expressed genes are not spatially linked. Nevertheless, it should be possible to activate genes from disjunct parts of the genome, simultaneously. Britten and Davidson (1969) proposed a model for coordinate expression in unlinked genes. Genes regulated in parallel with one another would contain common control elements. As specific signals have to be met by individual responses, cells require a set of freely combinable control elements. Each control element would be recognized by the product of an integrator gene. This product would then activate all genes containing one particular control element.

Britten and Davidson speculated on the nature of such control elements. Today, we know that DNA sequence elements about 6-30 base pairs in length serve this purpose. Such elements can be divide in two classes: general and specific ones. General elements can be found in most of the genes whereas specific elements occur only in one or few genes.

General sequence elements like the TATA box support the buildup of the basal transcription complex. This complex can only produce a low rate of transcription.This rate increases if additional sequence elements are freed or occupied by specific protein factors. The role of individual sequence elements can be assessed either by destroying them by deletion or mutation, or by placing them upstream of reporter genes in an attempt to confer the specific pattern of regulation to the reporter gene.

Table 2.1: Examples of general and specific control elements are given here.

General elements as in Bucher (1990)		
Name	Consensus	Protein factor
TATA Box	TWTWWAW	TFIID(TBP), Histones
Initiator	CWBHY	TFIID(TBP)
CCAAT Box	RRCCAWWSR	CBF, NF-1, C/EBP
GC Box	RGGHK	Sp1
Specific elements		
Name	Consensus	Protein factor
NF-κB	GGGAMTTYCC	p50/p65
CArG box	CC(A/T)$_6$GG	SRF

2.2.2 Transcription factors - protein components of gene regulation

The previously discussed DNA sequence elements are recognized by proteins called transcription factors. Transcription factors bind either directly to DNA or as part of a larger protein complex. Taking a classical view, transcription factors are composed of protein domains that are responsible for individual tasks. In the case of an activating transcription factor, two domains would be absolutely required: one domain that binds to DNA (DNA binding domain) and one that activates the transcription machinery (Activating domain). Evidently, both domains are not necessarily part of the same protein. Specific protein interactions often support the teaming up of the appropriate partners. Oct-1, a cellular factor bearing a DNA binding domain, and VP16, a viral protein with an activating domain, are well studied examples of the latter case.

Many different structural motifs with DNA binding capabilities are known and we do not attempt to give a comprehensive coverage. Figure 2.5 shows a sketch of a homeodomain, which is the DNA binding domain of several transcription factors (i.e. HOX transcription factors). Homeotic genes are crucial for the development of an organism. Experimental studies in the fruit fly (*Drosophila melanogaster*) accumulated evidence for a broad regulatory role of these homeotic genes. Structural analysis of the gene products unraveled a recurring motif called "the helix-turn-helix motif".

2.2.3 Activating transcription - mechanistic insights

Promoter regions encompass a number of DNA binding sites and thus signals from many different sources can be integrated at this level. Activating factors either ease complex assembly of the transcription machinery or stimulate the activity of the already assembled complex. Activators have a multitude of targets to exert their function and multiple interactions are the reason for strong synergistic activation. Fur-

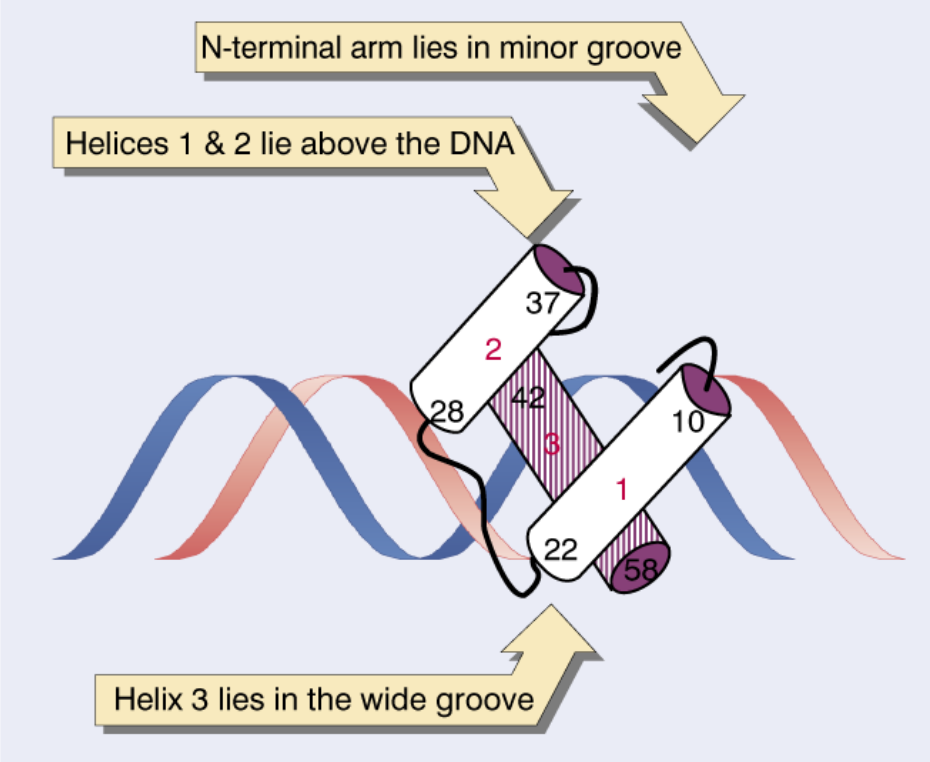

Figure 2.5: The conserved homeodomain DNA binding motif. Helix 3 of the homeodomain binds in the major groove of DNA, with helices 1 and 2 lying outside the double helix. Helix 3 contacts both the phosphate backbone and specific bases. Reproduced with permission from http://www.oup.co.uk/best.textbooks/biochemistry/genesvii/illustrations/

thermore, complexity increases due to additional protein players known as mediators or co-activators. The mediator complex partially envelops the RNA Polymerase II complex and conveys signals to its components.

2.3 Conservation of genes and their regulation

Gene configurations of species often resemble one another. This is rather intuitive as many biological processes like the cell cycle are shared among species. This idea is readily carried over to single gene pairs. Gene products from different species like proteins or microRNAs are far more similar than what would be expected by chance. The theory of Evolution explains this remarkable observation by postulating common ancestor sequences or species. Thus, the evolutionary history of genes or species can

be represented by tree structures. However, gene and species trees may differ for individual gene groups.

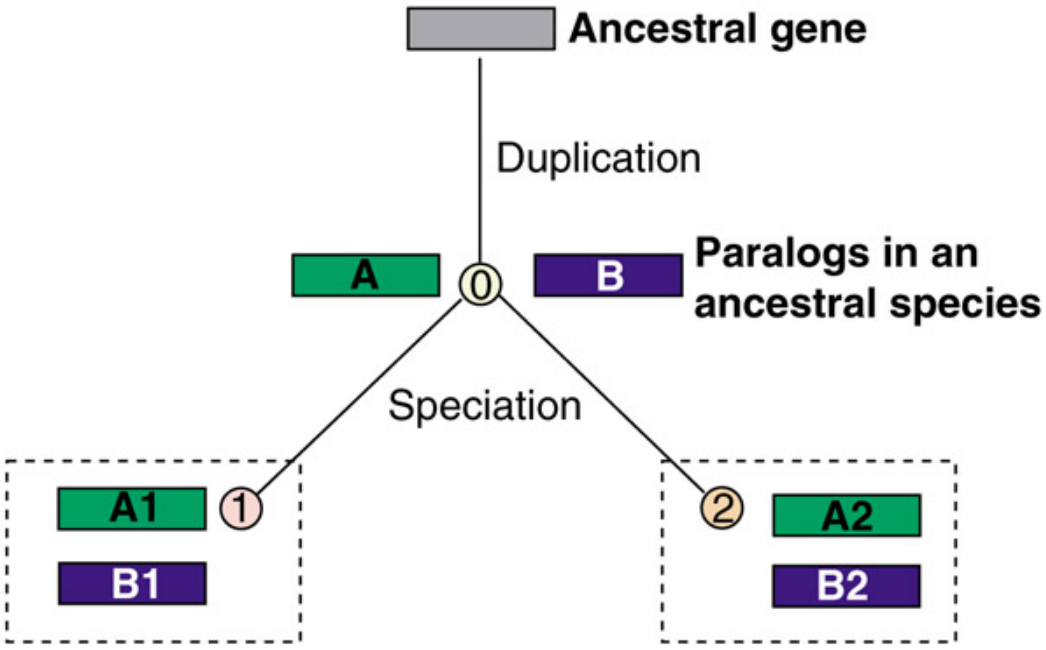

Figure 2.6: Definition of gene orthologs and paralogs. Genes are depicted as colored boxes. Genes A and B originate from a duplication event in an ancestral species. A speciation event isolates copies of the duplicated gene, which now evolve independently. Gene pairs in species 1 and 2 are called orthologous if the share the same color, and paralogous if they differ in color. Adapted from Koonin (2001).

Another problem may occur when the gene studied belongs to a multi-gene family: Suppose that two related species, 1 and 2, have two duplicated genes (A_1,B_1) and (A_2,B_2), respectively. Gene duplication occurred before the speciation event. In this case, genes A_1 and A_2 as well as B_1 and B_2 are called **orthologous genes**. All other pairings are called **paralogous**. In other words, "Two genes are said to be paralogous if they are derived from a duplication event, but orthologous if they are derived from a speciation event." (W-H Li, see Figure 2.6). Taken together, orthologous and paralogous genes form a group of **homologous genes** because of their shared ancestry. Bioinformatics offers solutions for the search of related genes based on sequence similarity. Relevant approaches and pitfalls will be briefly presented in Section 4.2.1.

The term **'synteny'** (or syntenic) refers to gene loci on the same stretch of DNA regardless of whether or not they are genetically linked by classic linkage analysis. Contiguous DNA regions that encompass two or more related genes in the same order

in different species (i.e. man and mouse) are an example of **conserved synteny**. Conserved synteny generalizes the concept of homology to large chromosomal regions. That is why, one would expect functional sequence conservation in non-exonic DNA regions.

Speaking of gene regulation, it has been known for a long time that there is considerable sequence conservation between species in non-coding regions of the genome. A comprehensive explanation of the observed sequence conservation patterns cannot be given, as research on this issue is far from completion. However, sequence conservation within promoter regions of genes often stems from transcription factor binding sites that are under selective pressure (see Hardison (2000) for a review and Liu et al. (2004) for a systematic assessment of binding site conservation in man and mouse comparisons).

2.3.1 Phylogenetic footprinting

Duret and Bucher (1997) give an overview on exploiting sequence conservation across species for the detection of regulatory elements. This concept is called **"phylogenetic footprinting"** and related computational approaches will be discussed in Section 3.

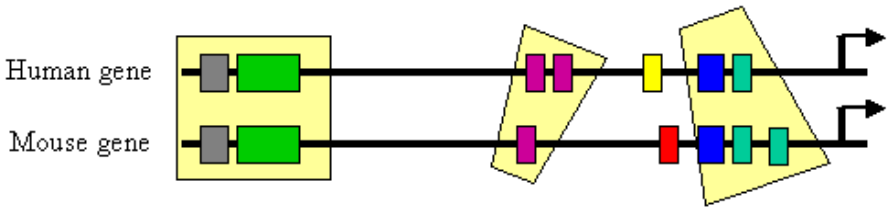

Figure 2.7: The concept of phylogenetic footprinting. Local sequence similarities in orthologous promoter regions of genes (light yellow regions) occur often due to selective pressure on transcription factor binding sites (colored boxes).

Phylogenetic footprinting in a strict sense is carried out on orthologous promoter regions. Local sequence similarities can then be directly interpreted as related regions harboring conserved functional binding sites. Selecting a suitable pair or set of sequences is crucial to the footprinting approach since evolutionary parameters vary for different species and loci. Section 4.2 sums up these issues in the context of the CORG database.

2.4 Selected experimental approaches

Tremendous technical progress has opened up the field of molecular biology to computational analysis. It is now possible to get a more holistic view of gene regulation on the system level. For instance, microarray technology facilitates the measurement of the expression levels of thousands of genes simultaneously. Thousands of interactions of proteins with proteins, or DNA, are now accessible by large-scale immuno-precipitation experiments. mRNA levels can be measured at an unprecedented level of sensitivity and accuracy. This section introduces the corresponding technologies that are relevant in the context of this thesis.

2.4.1 Sensitive and accurate detection of gene expression – RT-PCR

RT-PCR (reverse transcription-polymerase chain reaction) is the most sensitive technique for mRNA detection and quantitation currently available. Compared to the two other commonly used techniques for quantifying mRNA levels, Northern blot analysis and RNase protection assay, RT-PCR can be used to quantify mRNA levels from much smaller samples. In fact, this technique is sensitive enough to enable quantitation of RNA from a single cell.

Over the past several years, the development of novel chemistries and instrumentation platforms enabling detection of PCR products on a real-time basis has led to widespread adoption of real-time RT-PCR as the method of choice for quantitating changes in gene expression. Furthermore, real-time RT-PCR has become the preferred method for validating results obtained from array analyses and other techniques that evaluate gene expression changes on a global scale.

To truly appreciate the benefits of real-time RT-PCR, a review of PCR fundamentals is necessary (see Figure 2.8). At the start of a PCR reaction, reagents are in excess, template and product are at low enough concentrations that product renaturation does not compete with primer binding, and amplification proceeds at a constant, exponential rate. The point at which the reaction rate ceases to be exponential and enters a linear phase of amplification is extremely variable, even among replicate samples, but it appears to be primarily due to product renaturation competing with primer binding (since adding more reagents or enzyme has little effect). At some later cycle the amplification rate drops to near zero (plateaus), and little more product is made.

For the sake of accuracy and precision, it is necessary to collect quantitative data at a point in which every sample is in the exponential phase of amplification (since it is only in this phase that amplification is extremely reproducible). Analysis of reactions during exponential phase at a given cycle number should theoretically provide several orders of magnitude of dynamic range. Rare targets will probably be below the limit

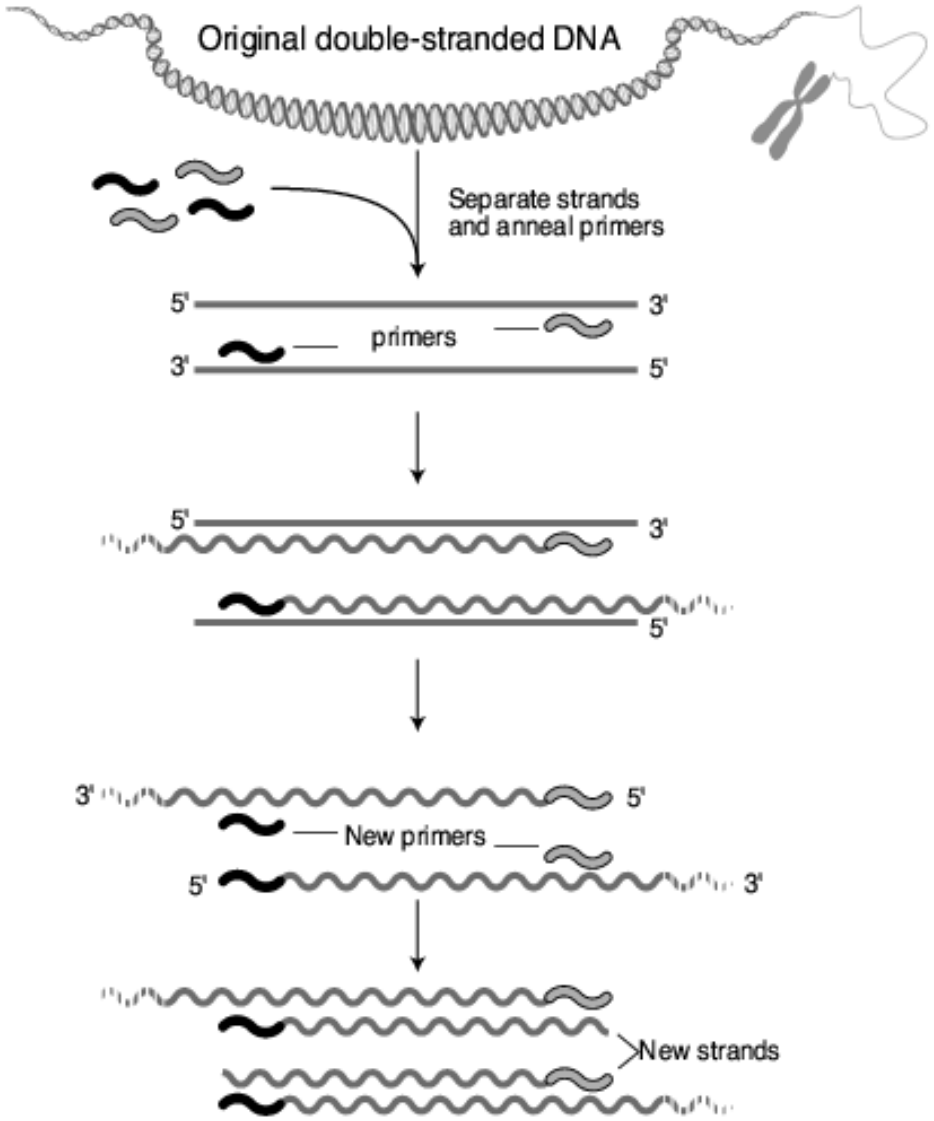

Figure 2.8: Schematic overview of the Polymerase Chain Reaction steps. PCR technology is a versatile tool to rapidly clone small segments of DNA ($\approx$ 1Kb). Original double-stranded DNA serves as template. The single DNA strands are separated by heating up the reaction vial to 95°C (Step 1). Specific probes (primers) that flank the region of interest are annealed by lowering the temperature to about 55°C (Step2). New strands are then synthesized with an enzyme called Taq polymerase at a slightly elevated temperature (Step 3). These new DNA strands serve then themselves as templates. If these steps are repeated several times, one easily gets 1 billion copies of a single template. Reproduced with permission from http://www.accessexellence.org

of detection, while abundant targets will be past the exponential phase. In practice, a dynamic range of 100 to 1000-fold product increase can be quantitated during end-point relative RT-PCR. In order to extend this range, replicate reactions may be performed for a greater or lesser number of cycles, so that all of the samples can be analyzed in the exponential phase.

Real-time PCR automates this otherwise laborious process by quantitating reaction products for each sample in every cycle. The result is an amazingly broad 107-fold dynamic range, with no user intervention or replicates required. Data analysis, including standard curve generation and copy number calculation, is performed automatically. With increasing numbers of labs and core facilities acquiring the instrumentation required for real-time analysis, this technique is becoming the dominant RT-PCR-based quantitation technique.

2.4.2 Large-scale measuring of gene expression levels with microarray technology

An eukaryotic cell can be regarded as a sophisticated machine. A common approach to collect information on this particular machine would be to perturb or stimulate the system. This is exactly what researchers do when challenging a cell with an external stimulus i.e. a chemical compound in a toxicity test. It is now desirable to measure the overall impact of this stimulus on the system. DNA microarray technology is the right choice for that purpose with respect to gene regulation.

DNA microarrays are arrays of many DNA molecules on a quartz, glass, or nylon substrate. Because of their resemblance to microchips used in computers, they are also called *DNA chips*. The foundation of microarray technology lies in the Watson-Crick complementarity of double-stranded DNA or RNA-DNA-hybrids. DNA molecules with the known sequence of genes (or parts of it) are printed on the chips as *probes* at regularly spaced and well defined locations called *spots* or *features*. The mRNA molecules, called *targets*, which are extracted from a tissue or blood sample, are prepared and labeled with a fluorescent or radioactive dye. The details of this process vary with the technology platform. The labeled targets are then allowed to hybridize to the probes on the array. Whenever the Watson-Crick complementary sequence of a probe is present in a target sequence, that target will hybridize to the probe. Unhybridized target molecules are washed off the chip, and the amount of hybridized target at each spot can be measured by the intensity of the dye or radioactivity. The idea behind this procedure is that each spot represents one gene and that the amount of hybridized target at a spot is a quantitative measure of the gene's transcript abundance in the cell sample, which is often interpreted as the gene expression level or its "activity" in the cell sample.

Naturally, the given explanation is generic in the sense that different DNA chip technologies exist. Differences between technologies occur in steps like target preparation and labeling, the nature of probes and the manufacturing of DNA chips.

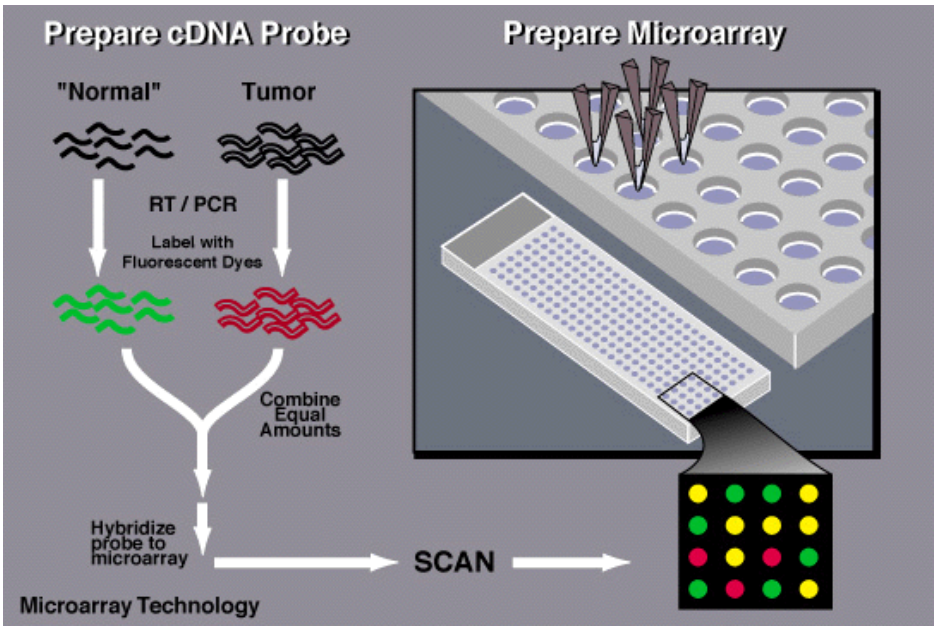

Figure 2.9: Overview of the cDNA microarray technology. mRNA is extracted from a reference sample ("normal" condition) and a test sample that deviates somehow from the norm. Here, the test sample originates from tumor tissue, which classifies as diseased tissue. The two RNA pools are reverse transcribed into cDNA and each pool is distinctly labeled with either a green or red dye. As a next step, equal amounts of the two fractions are pooled and hybridized to a DNA chip. The relative abundance of a gene's mRNA can be measured by scanning the intensities of the corresponding spots. Green spots indicate an overrepresentation of mRNA in the "normal" condition wheres red spots refer to an enrichment of mRNA in the "diseased" condition. Reproduced with permission from http://www.accessexcellence.org.

2.4.3 Chromatin immuno-precipitation

In vivo studies of protein-DNA binding have recently emerged as an active area of research. The importance of such studies cannot be overestimated since they deliver a snapshot of binding site occupancy in a genome. Large-scale mapping of DNA-binding sites for a particular transcription factor is possible by combining microarray technology with immuno-precipitation. A key component of the method is a specific antibody. An antibody is a blood protein that is produced as part of the immune response. An antibody can be raised to a transcription factor protein target or the protein itself is "tagged" such that an existing antibody recognizes it. "Tagging" introduces a new known protein domain (**tag**) at either the C-terminus or N-terminus of the protein. The latter case is preferred in species that are easily accessible to genetic modification like the baker's yeast (*Saccharomyces cerivisiae*). Once an antibody is available for the desired target, it can be used to select for, or enrich the target protein from a solution of proteins.

The following list highlights all principal steps of the experimental protocol as outlined in Wyrick and Young (2002).

1. **Sample collection** Depending on the model organism or question of interest, either cells are grown under well-defined conditions or samples are taken from the living animal. For example, Simon et al. (2001) studied the serial regulation of the yeast cell cycle in a population of asynchronous dividing cells. Here, cells were harvested from an exponentially growing batch.

2. **Crosslinking and DNA shearing** DNA-protein interactions are covalently linked using formaldehyde treatment of the living cells. These links are reversible. Genomic DNA is subsequently torn into small pieces of $\approx$ 1Kb size.

3. **Immuno-precipitation** Antibodies are now employed to select for DNA target sequences of a given transcription factor. In short, a test fraction (treated with antibody) and a reference fraction (pool of genomic DNA fragments) are compared with one another. Antibody-protein-DNA complexes are removed from the test fraction by addition of an insoluble form of an antibody binding protein such as Protein A, Protein G or a second antibody (precipitation). Cross-linking is reversed and short linker regions are ligated to the DNA pieces.

4. **Amplification and Hybridization** A linker mediated PCR reaction amplifies the amount of DNA in both fractions and labels both fractions with a different dye. The last step is the hybridization to an array of selected probes. For instance, Simon et al. (2001) used a DNA chip of yeast intergenic regions.