CHAPTER 2

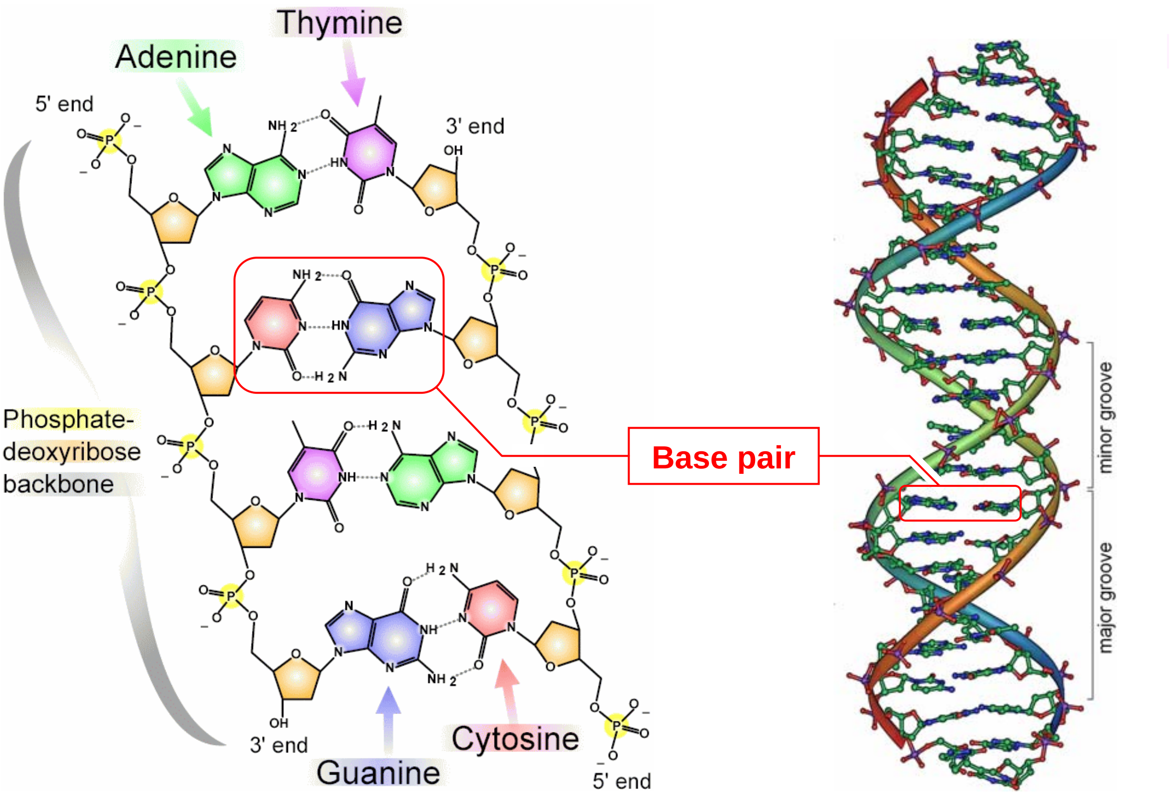
Organisation of eukaryotic genomes

Since the focus of this thesis lies on improving our understanding about transcription factor dependent gene regulation in eukaryotes, in this chapter I will briefly outline the structure of eukaryotic genes and the basic concepts of transcriptional regulation. In addition, several experimental techniques including microarrays, EST sequencing and ChIP on chip will be introduced as they play an important role for the analyses presented in later chapters. Comprehensive information about most of the presented material can be found in e.g. Alberts et al., 2007 and Lodish et al., 1995.

2.1 DNA as carrier of the genetic code

Despite the enormous differences in the size and complexity of eukaryotes the way their genomes are organized is surprisingly similar. As for prokaryotes the genetic information, which encodes the building plan of each eukaryotic organism is stored in its deoxyribonucleic acid sequence (DNA). DNA is a linear biopolymer composed of four different subunits or bases called adenine (A), cytosine (C), guanine (G) and thymine (T). Typically millions of these bases are connected via a sugar phosphate backbone to form individual eukaryotic chromosomes (Lodish et al., 1995). The order of the basepairs along the chromosomes from its 5’ end to its 3’ end hereby encodes the genetic information of the cell. An important property of DNA is that in the cell it does not exist as a single long molecule but that each DNA strand is paired with a complementary sister molecule (see Figure 2.1). The two sister strands are thereby held together via hydrogen bonds that form between their bases. The base pairing is specific with an adenine always forming two hydrogen bonds with a thymine and a guanine always forming three hydrogen bonds with a cytosine. The created base pairs (A=T and G≡C) are stacked upon each other like the steps of a rope ladder. Steric constraints together with Van der Waals interactions, which attract neighbouring base pairs, force the DNA into its predominant form in the cell, the typical B-DNA double helix structure (Lodish et al., 1995).

Figure 2.1 – Structure of the double stranded DNA.



The left panel shows a schematic image of chemical structure of a DNA molecule bound to its complementary sister strand. The interaction between the two strands is achieved by the formation base pairs via hydrogen bonds between the bases cytosine and guanine respectively adenine and thymine. The order of the four bases along the DNA molecule (read from the 5’ end to the 3’ end) encodes the genetic information of an organism. Attractive forces between the base pairs cause the DNA to adopt its famous double helix structure shown on the right (images obtained from http://en.wikipedia.org/wiki/DNA). One turn of the helix thereby corresponds to 10 base pairs or 3.4 nm in length. The diameter of the helix is about 2 nm.

A typical genome of a higher eukaryote is between several hundred million and several billion bases long (Lodish et al., 1995). For instance the human genome has a length of around 3 billion base pairs (separated into 22 autosomes and 2 sex chromosomes). Over the last decade a large number of genomes have been successfully sequenced, that is, the order of the bases along their chromosomes has been experimentally determined (e.g. Ruvolo 2004). Due to the development of new high throughput sequencing techniques such as pyrosequencing (Ronaghi et al., 1998) the number of available DNA sequences is ever more quickly increasing. While sequence data is thus becoming overabundant deciphering the genetic code hidden in it remains a challenging task.

2.2 The structure of eukaryotic genes

An important first step in unravelling the genetic code of an organism is the discovery of its genes. Under the classical definition, genes are regions along the chromosomes that encode functional cellular components such as proteins. In a process called transcription the sequence of a gene is used as template to synthesize ribonucleic acid (RNA) molecules. This is done by an enzyme called RNA polymerase (Lodish et al., 1995). RNA is a linear biopolymer chemically similar to DNA, which serves two important roles. Firstly, many short RNA species either directly catalyse vital metabolic reactions or serve as important structural components of larger RNA – protein complexes (Valadkhan 2007, Komatsu 2004, Szymański et al., 2003). Secondly, RNA is used as messenger molecule (mRNA) for the production of proteins (Lodish et al., 1995). The part of an mRNA molecule that is later translated into a chain of amino acids is referred to as the open reading frame (ORF) of the gene. The ribosome (a large protein - RNA complex that catalyzes the assembly of the amino acid chains, Szymański et al., 2003) hereby translates always three consecutive bases of the ORF into a particular amino acid. In prokaryotes genes are composed of continues ORFs, which when transcribed into RNA, can serve directly as a template for protein synthesis (Lodish et al., 1995). In eukaryotes the gene structure is more complex with parts of the ORF being separated by none coding DNA segments (Lodish et al., 1995). As shown in Figure 2.2 eukaryotic genes can be separated into the following elements:

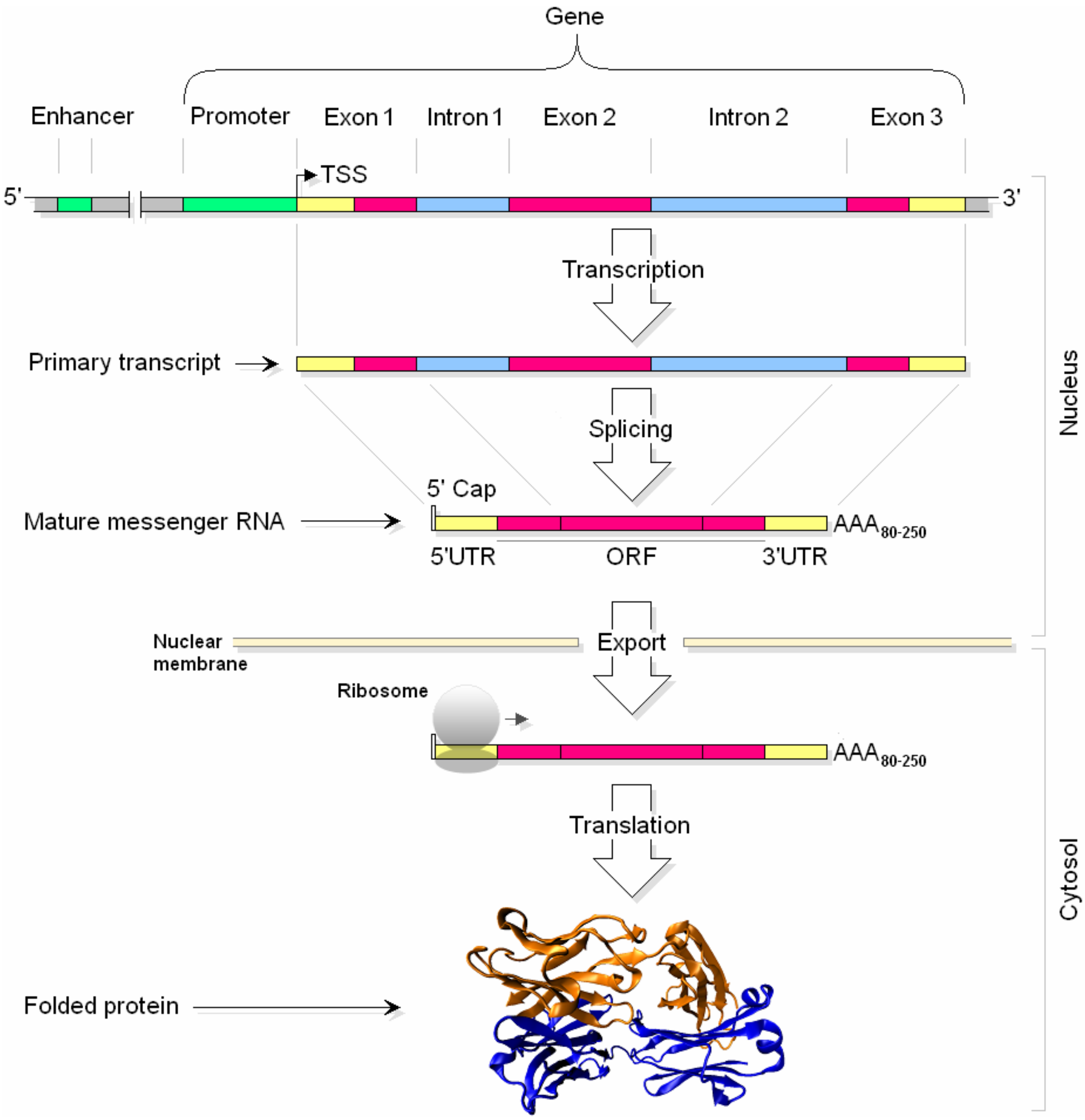
Promoter

The promoter of a gene is a stretch of DNA that is recognized by proteins called transcription factors which promote or block the binding and activity of RNA polymerase and subsequently the transcription of the gene. Different transcription factors thereby recognize and bind specifically to different short DNA base pair patterns. Without the binding of such transcription factors most eukaryotic genes are not actively transcribed. Regulatory DNA sequences further upstream of the gene, which come in contact with the promoter via looping of the DNA, are referred to as enhancer elements. These can lie mega-bases away downstream or upstream of the gene (Kininis et al., 2008, Dostie et al., 2007).

Transcription start site

The transcription start site (TSS) corresponds to the position of the first base translated into RNA by RNA polymerase. The RNA polymerase initiation complex binds from position -43 to +24 around the TSS (Hahn 2004).

Figure 2.2 – From eukaryotic genes to proteins



The top line shows the structure of a typical eukaryotic protein coding gene along a chromosome. Exons, Introns, promoter region and enhancer elements are indicated. The first and last exon can further be subdivided into an untranslated region (UTR) and a region that belongs to the open reading frame (ORF) of the gene. Starting at the transcription start site (TSS) the gene is first transcribed by RNA polymerase which synthesises the primary transcript (middle line). During a second step introns are removed from the transcript via splicing, a poly-A tail is added to the 3' end and a cap structure is added to the 5' end thus giving rise to a mature messenger RNA (bottom line). The mRNA is then exported from the nucleus to the cytosol where its ORF gets translated by a ribosome into an amino acid chain which subsequently folds into a protein (3D protein structure obtained from http://en.wikipedia.org/wiki/Protein).

Exons

Exons are the sequence blocks which are joined together via splicing to form the mRNA. Once a poly-A tail is added to the 3’ end of the mRNA and a cap structure is added to its 5’ end the mature mRNA is transported from the nucleus to the cytosol and its open reading frame is translated into a protein (Lodish et al., 1995). Usually the 5’ and 3’ most exons can further be divided into a translated part (which becomes part of the ORF) and an untranslated part (UTR). Particularly the 5’UTR serves as regulatory region with potential binding sites for transcription factors in the nucleus and later as recognition site for the ribosome (Lodish et al., 1995, Pickering et al., 2005).

Introns

Introns are stretches of DNA which get excised from the primary transcript and thus do not become part of the mature mRNA (Lodish et al., 1995). The presence of introns opens the possibility for alternative splicing. Alternative splicing permits different exons to be joined together to form different mature messenger RNAs. For instance, for the case shown in Figure 2.2, two possible mRNAs could be formed, one by joining exons 1, 2 and 3 and one by splicing together exons 1 and 3 thereby cutting out exon 2 together with introns 1 and 2. In many cases alternative splicing also allows for the presence of different transcriptional start sites by varying the choice of the first exon. Splicing therefore greatly increase the combinatorics of eukaryotic gene products. In addition to this vital role introns also often contain regulatory elements.

While coding regions make up most of the genomic sequence of a prokaryotic cell in higher eukaryotes genes are interspersed among vast regions of non-coding DNA. In humans for instance only 1-2% of the DNA encodes for exons while most of the genome has no yet discovered function (Alberts et al., 2007). Identifying regulatory regions and in particular transcription factor binding sites located many kilobases away from their corresponding genes therefore remains a major challenge today.

2.3 Regulation of gene expression

Evolution has provided cells with a variety of different mechanisms to control the production and activity of its final gene products. Next to controlling transcription these include RNA splicing, editing and translation, post-translational modifications of proteins and subsequently degradation of the gene products. Together these mechanisms permit prokaryotes such as *E. coli* or simple eukaryotes such as yeast to adapt to different environmental conditions and allow the fertilized egg of a higher eukaryote to give rise to hundreds of different cell types during embryonic development. In the scope of this manuscript we are concerned only with the first step, the regulation of transcription. The following section will therefore outline how the cell establishes its transcriptome (the set of all RNA molecules, produced in a cell) by reviewing the crucial role of transcription factors and chromatin modifications in modulating gene expression.

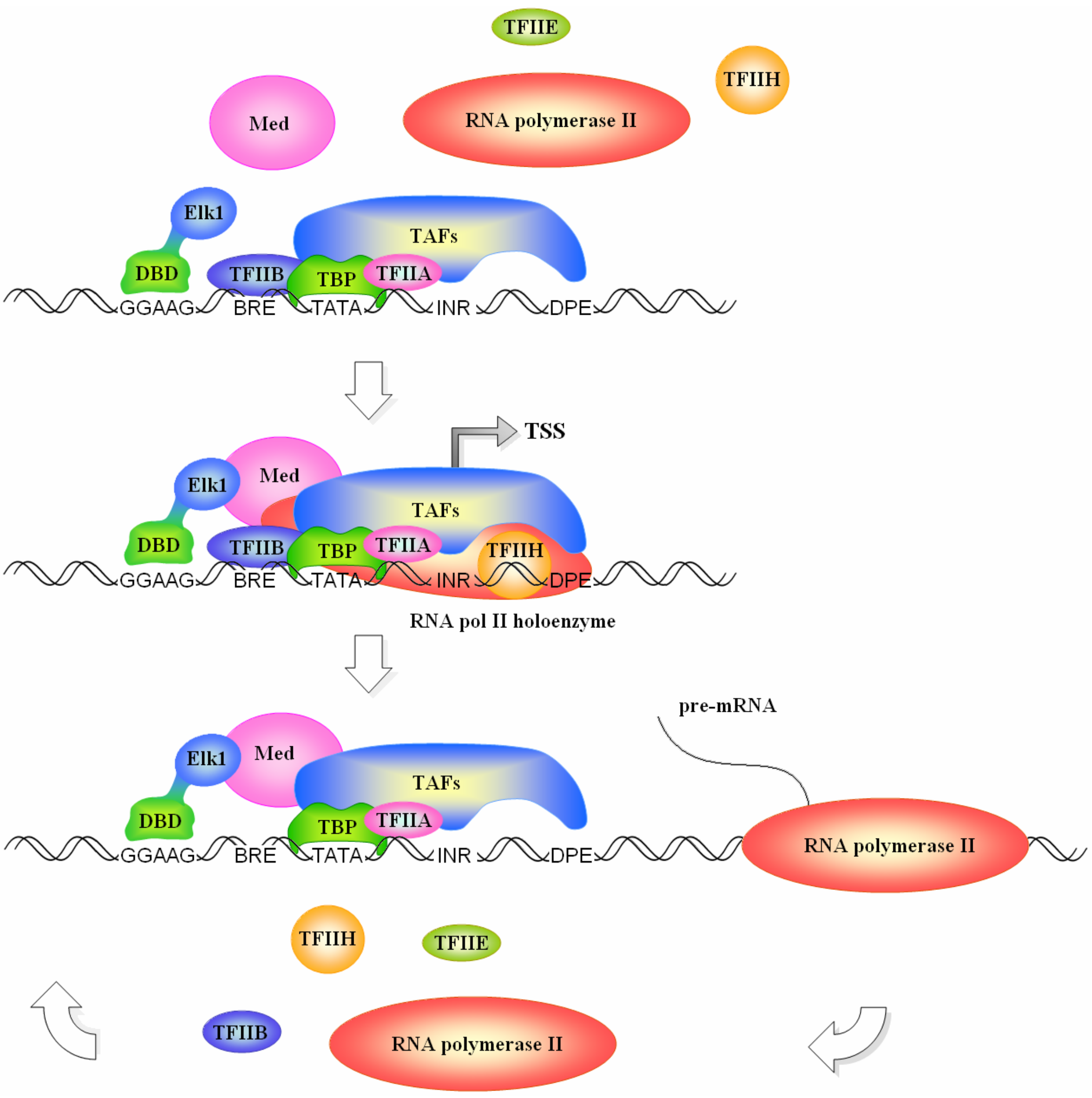
2.3.1 Transcription factors

Transcription factors (TFs) are a main component of the transcriptional machinery that cells can modify in order to control the levels of gene expression at a given gene locus (Alberts et al., 2007, Lodish et al., 1995). Based on their function transcription factors (TFs) can be divided into several groups. The first group are general TFs which are required to initiate transcription but do not recognize gene specific DNA sequences (Alberts et al., 2007). In fact not all of these factors actually bind DNA but many rather mediate between DNA binding proteins and the RNA polymerase. The second group are sequence or gene specific activators. Without their presence most eukaryotic genes are silent (Alberts et al., 2007). Activators recognize specific DNA sequences in the vicinity of genes and up-regulate their expression level. Many of their corresponding DNA motifs are located in the proximal promoters (within several hundred base pairs around the TSS) but can also lie in enhancer elements far away from a given transcription start site (Alberts et al., 2007, Kininis et al., 2008, Dostie et al., 2007). A last group of TFs are transcriptional repressors. These TFs decrease rather than increase the expression level at a given gene locus. Repressors function either by making the DNA inaccessible for the transcriptional machinery, by blocking the binding of activating TFs or by directly modifying the RNA polymerase so that transcription cannot be initiated (Alberts et al., 2007). In the following I will briefly review the structure of the RNA initiation complex and then discuss general principles of how transcriptional activators function.

General transcription factors and the RNA polymerase II initiation complex

Eukaryotic cells possess three different RNA polymerases. RNA polymerase I is responsible for transcribing ribosomal genes (Russell et al., 2005). RNA polymerase II is responsible for transcribing protein coding genes, which is the focus of our attention in this manuscript. Lastly, RNA polymerase III transcribes transfer RNAs and other small RNA species (Haeusler et al., 2006).

Figure 2.3 – Binding of RNA polymerase II holoenzyme to the core promoter



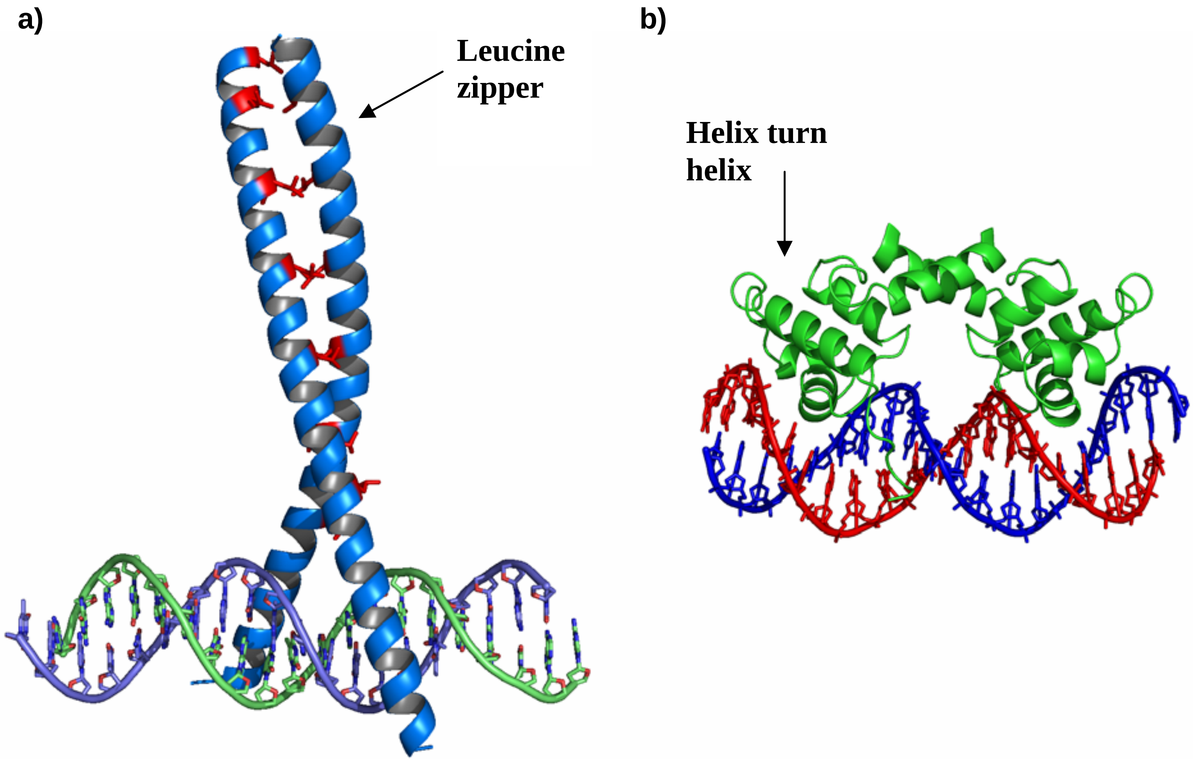
First TFIID (TBP + TAFs) binds together with TFIIB and TFIIA to the core promoter elements (TATA box, BRE, Inr and DPE) either independently or with the help of gene specific TFs. Then RNA pol. II and other basal TFs bind to the platform established by TFIID. Particularly this step can be facilitated by the presence of gene specific TFs such as Elk1. These specific TFs bind to RNA pol. II complex via a large mediator protein complex. Finally, TFIIH binds near the initiator site and helps to melt the DNA helix. After transcription is initiated RNA pol. II leaves the core promoter together with several subunits of the holoenzyme while the TFIID and other factors remain at the core promoter. These subunits help to readily reassemble the holoenzyme.

The RNA polymerase II holoenzyme is a multi protein complex that consists of the actual enzyme and several associated basal TFs which are required for proper binding to core promoters and subsequently for initiating transcription (Hahn 2004, Butler et al., 2002). Several sequence motifs found in typical core promoters play an important role for the assembly of the holoenzyme. The most famous of these elements is the TATA box whose sequence reads TATAAA and which is usually present in tightly controlled promoters (Hahn 2004, Butler et al., 2002). The TATA box is located about 34 base pairs upstream of the transcription start site and is specifically recognized by the general transcription factor TBP (TATA box binding protein). TBP plays an important role as a platform on which the transcription machinery is assembled (Alberts et al., 2007, Hahn 2004, Butler et al., 2002). In addition it is responsible for correctly positioning RNA polymerase at the TSS. Directly upstream of the TATA box lies the BRE motif with its sequence G/C-G/C-C-G-C-C (Hahn 2004, Butler et al., 2002). If present, the BRE motif is bound by the general transcription factor TFIIB. Two additional sequence motifs are the short initiator element (Inr) which is located directly at the TSS, and the downstream promoter element (DPE) which is located 30 bps downstream of the TSS and whose sequence reads A/G-G-A/T-C/T (Hahn 2004, Butler et al., 2002). Both of these elements are bound by TBP associated factors (TAFs). TAFs together with TBP form the general factor TFIID (Lodish et al., 1995, Hahn 2004). Most promoters contain at least one of the above elements, although none of the elements is strictly required for transcription to occur. Figure 2.3 shows the events at the core promoter that lead to the binding of RNA polymerase II. TFIID, TFIIA and TFIIB are believed to associate with the core promoter before the rest of the holoenzyme can bind (Hahn 2004). TFIID is thereby required for promoter assembly even in the absence of a TATA box. RNA polymerase II as well as most other basal TFs can either bind in a sequential fashion to prebound TFIID or enter as a preassemble complex. As will be discussed below, the assembly of the RNA polymerase II holoenzyme as well as the binding of TBP can be guided by gene specific TFs.

Gene specific transcription factors

While TBP and the other basal DNA binding TFs play an important role in positioning RNA polymerase at the core promoter their DNA binding specificity is not sufficient to detect specific promoters among the vast genomic regions present in eukaryotic cells. Therefore, the gene specific binding of additional activating transcription factors is required. Activators exert their function via three main mechanisms. First, some factors modify the chromatin structure (see below) in order to give the basal transcription machinery access to the DNA (Alberts et al., 2007, Kininis et al., 2008). Second, as illustrated in Figure 2.3 they may help in assembling the transcription complex at the promoter by linking RNA polymerase to the promoter via a mediating protein complex (Alberts et al., 2007, Cantin et al., 2003, Kininis et al., 2008). The interaction with the activator thereby lowers the binding energy of the holoenzyme which enhances the assembly process. Lastly, activators can modify the holoenzyme thereby facilitating the initiation of transcription (Kininis et al., 2008). Repressors on the other hand are designed to counteract any of these mechanisms.

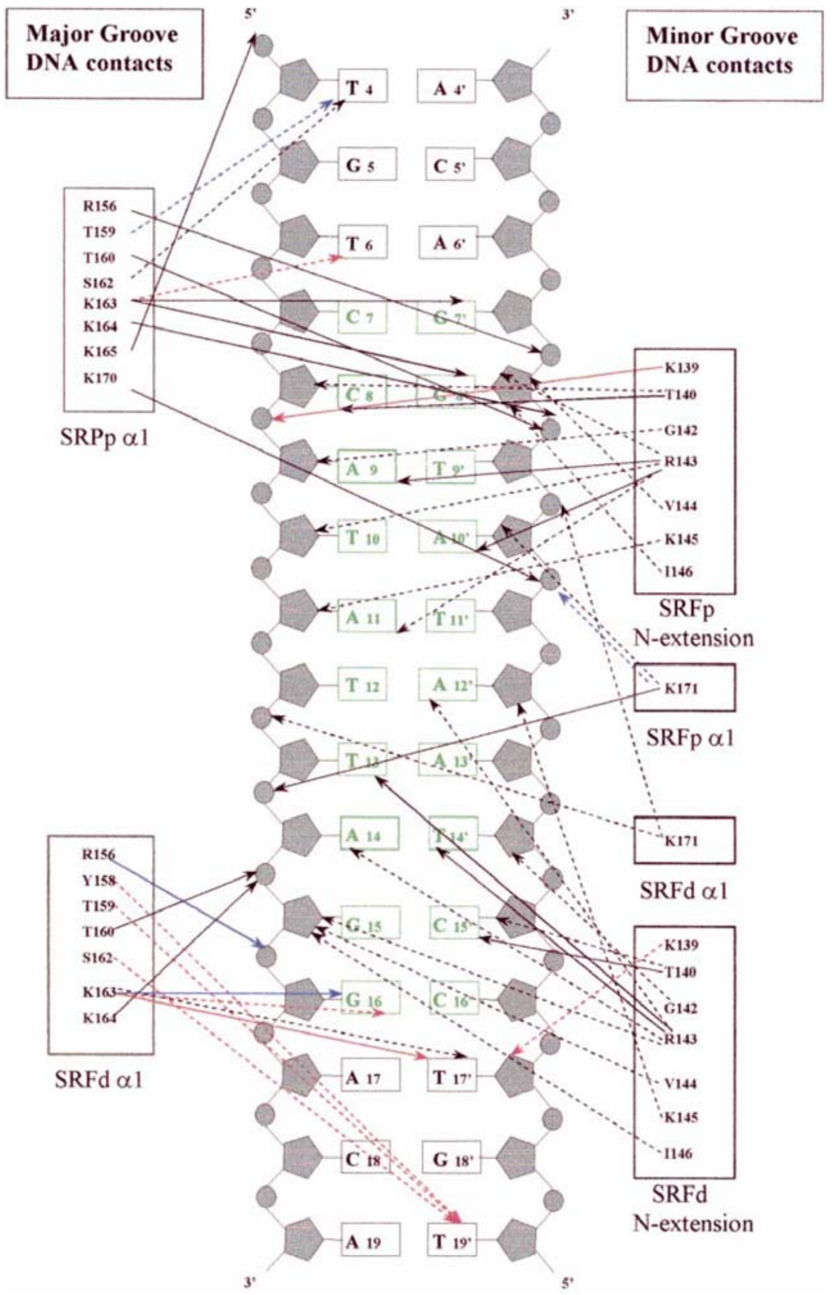
Figure 2.4 – Binding of a leucine zipper and a helix turn helix TF to DNA



a) Structure of a leucine zipper bound to DNA. These transcription factors consist of two long alpha-helical proteins which are held together by the interaction of leucine residues (shown in red) found at every 7th position (image obtained from http://en.wikipedia.org/wiki/Leucine_zipper). b) Structure of a helix turn helix transcription factor bound to DNA. The DNA binding domain is composed of two short alpha helices connected via a short amino acid stretch. One of these helices stabilizes the interaction between the TF and the DNA in an unspecific fashion while the other helix (recognition helix) forms sequence specific interaction with the base pairs (image obtained from http://en.wikipedia.org/wiki/Helix-turn-helix). Both factors bind as homodimers to DNA.

Characteristically, activators as well as repressors are composed of at least two functional subunits, a DNA-binding domain and a transactivation domain. The different DNA binding domains such as Helix-turn-helix, Zinc finger and Leucine zipper are used to categorize TFs (Wilson et al., 2008). Figure 2.4 shows as an example the crystal structure of a leucine

Figure 2.5 – Interactions between the DNA-binding domain of SRF and its DNA motif



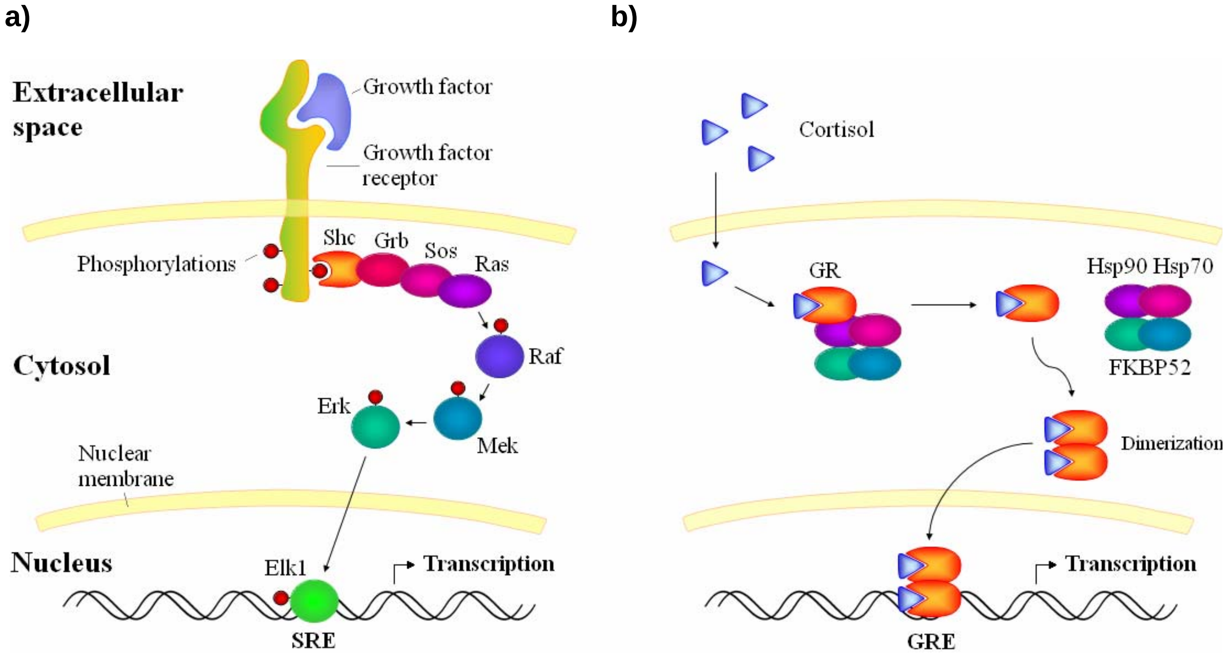
Letters and numbers in the boxes indicate the type and position of the amino acids in the SRF protein that stay in contact with the DNA. Van der Waals interactions between the DNA and amino acid are shown as dotted lines while hydrogen bonds are indicated as solid arrows. The bases of the DNA motif recognized by SRF are indicated in green (Mo et al., 2001). It should be noted that the TF does not specifically interact with all bases in the binding site as can be seen in case of the second to top GC base pair. Correspondingly, such bases are not restricted and tend to change randomly between binding sites.

zipper and a helix turn helix transcription factor bound to DNA. As shown, both TFs are complexes made up of two identical proteins (homodimers). This is in fact a general feature with most TFs working as homo or heterodimers. However, some factors can bind also as monomers or form higher order complexes such as P53 which binds to DNA as tetramer (Ma et al., 2007). Figure 2.5 illustrates a detailed analysis of the interaction between the amino acid side chains of a TF and its preferentially bound DNA binding site (Mo et al., 2001). The affinity of a TF to a site is provided by the amino acid residues that interact with the DNA backbone and/or with specific base pairs. Electrostatic interactions between positively charged amino acids and the negatively charged sugar phosphate backbone of the DNA are primarily unspecific and give TFs a general affinity for DNA. In contrast, the sequence specific binding character of a TF is provided largely by the hydrogen bonds and van der Waals interaction that form between the amino acids and the base pairs (Pabo et al., 1984, Vázquez et al., 2003). Importantly, bases that form strong interactions with the TF tend to be retained between different binding sites while bases that do not interact with the TF are not restricted and thus may vary randomly from site to site.

Cellular control over TF activity

In order to adapt to signals from the environment cells can modulate the activity of their TFs in multiple ways. A first mechanism is to change the expression of a TF itself. This can be done on the transcriptional as well as on the translational level. For instance, for certain factors proteins exist that can bind to the mRNA of the factor and block its translation (Abaza et al., 2008). Second, the TF can be retained in the cytosol by the binding of proteins that block its transfer to the nucleus (Lodish et al., 1995, Bruner et al., 1997). Alternatively, the factor might have to dimerize or undergo posttranslational modification before it can bind to DNA (Lodish et al., 1995, Drouin et al., 1992, Herdegen et al., 1998, Davis 1995). Both events can be triggered by chemical modifications such as the phosphorylation of tyrosin or serine residues or the binding of ligand molecules. Similarly, the availability of cofactors needed for relocalization to the nucleus or for the interaction with the basal transcription machinery can be restricted. Figure 2.6 illustrates some of these mechanisms based on two exemplary signal transduction pathways that lead to the expression of specific target genes (Bruner et al., 1997, Drouin et al., 1992, Herdegen et al., 1998). Lastly, the accessibility of a DNA binding site for a given activator can be restricted either by competitive binding of a repressor to an overlapping DNA motif or by modifications made to the DNA and the chromatin structure. These mechanisms will be reviewed below.

Figure 2.6 – Cell signaling leading to transcriptional activation

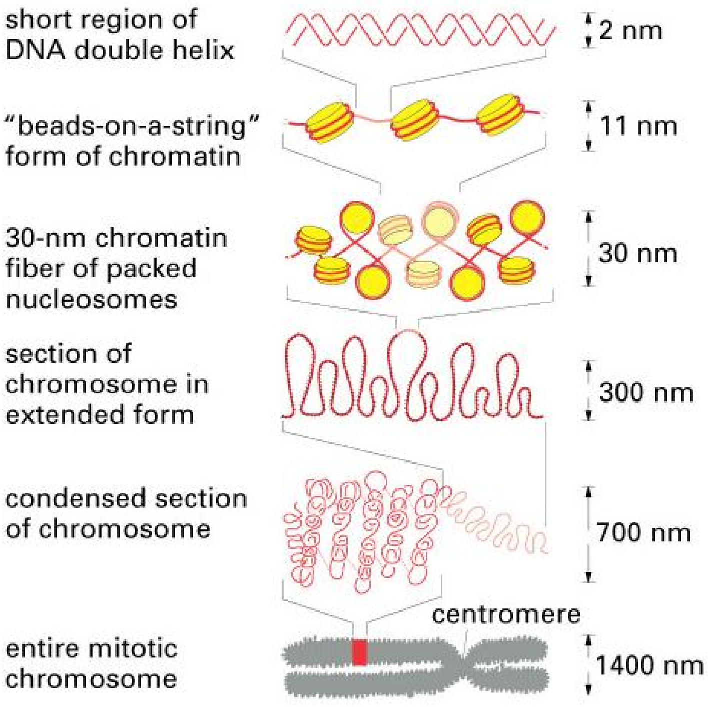


a) Signal transduction pathway leading to the activation of the transcription factor Elk1 (Herdegen et al., 1998). The cascade is initiated by the binding of growth hormone to its cell surface receptor which leads to the phosphorylation of tyrosine residues in the intracellular domain of the receptor and in turn triggers the activation of the Ras/Raf Map kinase pathway. Eventually the signal is transmitted to Elk1 whose phosphorylation leads to its relocalization to the nucleus. There Elk1 binds to the serum response element (SRE) and helps in the assembly of the basal transcription machinery at the TSS. b) Activation of glucocorticoid nuclear receptor by the binding of Cortisol (Bruner et al., 1997, Drouin et al., 1992). Like other steroid hormones Cortisol can diffuse directly through the cell membrane into the cytosol where it binds to the GR receptor, which functions directly as transcription factor. GR needs to dissociate from a retaining protein complex composed mainly of heat shock proteins before it can dimerize and relocalize to the nucleus where it binds to the glucocorticoid responsive elements (GRE) and initiates transcription.

2.3.2 Chromatin modifications

Aside from directly modifying the activity of transcription factors cells can also modulate the accessibility of the DNA for the transcriptional machinery. This is achieved by modifying histones, large ubiquitous protein complexes that are required for DNA packing, or via DNA methylation. In the following both mechanisms will be briefly reviewed.

Figure 2.7 – Nucleosome and chromatin structure


short region of DNA double helix
2 nm
"beads-on-a-string" form of chromatin
11 nm
30-nm chromatin fiber of packed nucleosomes
30 nm
section of chromosome in extended form
300 nm
condensed section of chromosome
700 nm
centromere
entire mitotic chromosome
1400 nm

The binding between histones and DNA can be strengthened or weakened by chemically modifying the amino acid side chains of the histones via acetylation and methylation thereby modifying the electrostatic charge of the histone surfaces. In addition, adding another histone subunit called H1 causes individual histones to bind to each other thereby causing the formation of a structure called 30-nm chromatin fibre. By an unknown mechanism these fibres can form even larger super structures of densely compacted chromatin called heterochromatin. Most heterochromatin in the cell is transcriptionally inactive and corresponds to a good degree with the centromeres and telomeres of the chromosomes. (Image obtained from: http://www.accessexcellence.org/RC/VL/GG/ecb/chromatin_packing.html).

Histone modifications

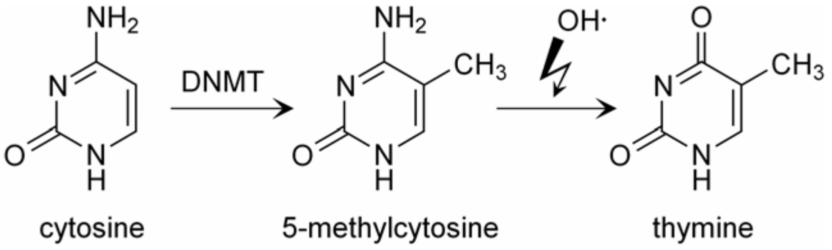
As mentioned before, the genomic DNA of a human cell is about 3 billion base pairs long. This corresponds to a molecule length of nearly two meters. In order to fit this amount of DNA into the nucleus of a eukaryotic cell with a diameter of less than a micrometer the DNA needs to be heavily compacted. At the same time DNA must be easily accessible in order to permit the interaction with the proteins that regulate and perform transcription and replication. This is achieved by the binding of DNA to globular protein complexes called histones (Alberts et al., 2007). Histones are composed of eight subunits. Their positively charged surface readily causes the negatively charged DNA backbone to associate with them. A single DNA - histone complex is thereby referred to as nucleosome (Alberts et al., 2007). On a larger scale the structure of subsequent nucleosomes is reminiscent of a pearl string with always 157 base pair long stretches of DNA wrap around one histone and roughly 30 base pair long stretches connecting to the next nucleosome (see Figure 2.7). The entirety of DNA - histone complexes is referred to as chromatin.

Cells can modify the binding strength between histones and DNA in order to increase or decrease the accessibility of promoters for the transcriptional machinery. This is done by transferring chemical residues such as acetyl or methyl groups to the side chains of the histones. For instance, transferring acetyl groups to lysine residues at the histone tails causes histones to loose positive charges. This weakens the histone - DNA interaction and in turn tends to increases transcription (Fukuda et al., 2006, Alberts et al., 2007). Acetylation and deacetylation are controlled by histone acetyltransferases and histone deacetylases (Fukuda et al., 2006, Alberts et al., 2007). Some of the acetyltransferases such as p300 can thereby interact with various basal and sequence specific transcription factors. This allows the sequence specific opening of the chromatin structure and local access of RNA polymerase to the DNA.

DNA methylation

Aside from modifying transcription factors and histones cells can also modify the bases of the DNA in order to change transcription levels. The most common of these modifications is the methylation of cytosines in CG dinucleotides. CG dinucleotides are rare in the human genome due to an increased rate in transitions from methylated cytosine to thymine (Figure 2.8). However, in promoter regions, likely due to their regulatory function, CG dinucleotides tend to be retained and are found in clusters referred to as CpG islands. DNA methylation is carried out by a class of enzymes called DNA methyltransferases or DNMTs (Turek-Plewa et al., 2005, Alberts et al., 2007). In humans these enzymes keep three quaters of all CG dinucleotides methylated. Their action can be reversed by specific demethylases. In general DNA methylation in the promoter region leads to transcriptional repression. Therefore CpG islands in promoters of houskeeping genes has to stay largly unmethylated. Several mechanisms exist how methylation influences transcription. First, the binding ability of transcription factors which bind to CpG containing motifs can be disrupted by the presence of a methyl group. Second, some repressors specifically require a methyl group in order to effectively bind to DNA. Most of these factors bind to DNA via a methyl CpG binding domain (MBD). Lastly, the DNA methylation status frequently goes in hand with the acetylation status of histones and changes in chromatin structure (Turek-Plewa et al., 2005).

Figure 2.8 – Modification of cytosine and transition to thymine



Cytosine bases within CG dinucleoties are frequently methylated at their C5 position giving rise to 5-methylcytosine. This reaction is specifically catalyzed by DNA methyltransferases (DNMTs). In a reaction with hydroxy radicals the amino group at the C4 position can be replaced with a oxygen atom thereby mutating the cytosine base into a thymine base. This mutation cannot be unambiguously detected by the DNA repair machinery leading to a conderably higher transition rate from C to T than vice versa (Lodish et al., 1995).

This is likely caused by MBD posessing repressor which are part of histone deacetylation complexes. The deacetylation induced by these complexes causes histone side chains to acquire additional positive charges thereby streengthening the binding to DNA and thus leading to denser chromatin packing and blocking of the transcriptional machinery (Vaillant et al., 2007, Alberts et al., 2007).

The methyl status of a mother cell is inherited by the daughter cells. Therefore, methylation plays an important role in embryonic development. Large scale experiments are currently on the way to decipher the methylation status of promoters in various different cell types and to further resolve the role of methylation in chromatin remodelling and influencing expression levels (Beck et al., 2008, Siegmund et al., 2007, Das et al., 2004).

2.4 Measuring gene expression

Starting from small scale experiments several high throughput methods have been developed that allow to simultaneously assess the transcriptional activity of a large number of genes. Since in the course of this thesis I make heavy use of microarray and EST data these two methodologies will be the focus of this section.

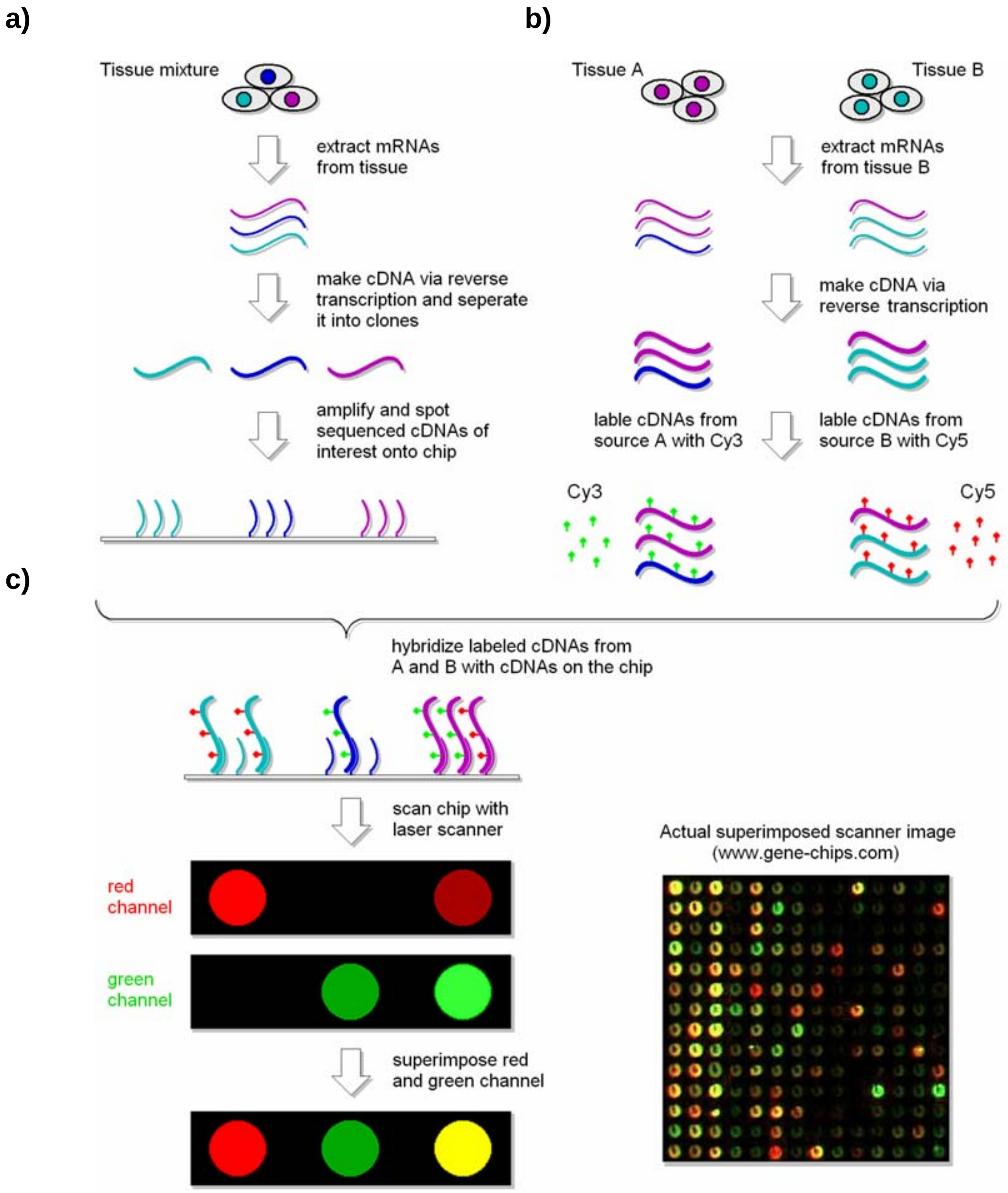
2.4.1 Microarray analysis

Microarray analysis is a widely used technique to simultaneously measure the transcript levels of a set of genes whose sequences are known *a priori* (Albters et al., 2007). Several methodologies exists for producing the microarrays. For instance, as shown in Figure 2.9, PCR amplified DNA sequences can be spotted onto a carrier matrix such as an microscopy slide (spotted cDNA microarrays, Hager 2006). Alternatively, short DNA sequences can directly be synthesized on the matrix (e.g. Affymatrix chips, Dalma-Weiszhausz et al., 2006). Both approaches have in common that each spot on the array consists of a large number of copies of the same sequence. Which sequence corresponds to which spot must thereby be known. Having generated a microarray it can subsequently be used to assess differences in the transcriptome between different cell types or cellular conditions. To this end, mRNAs are extracted separately from two different cell types under investigation and reverse transcribed into cDNA. The cDNAs from either cell type are then differently labelled using a red and green fluorescent dye, respectively. Next, the differently labelled cDNAs are pooled and incubated with the microarray (Figure 2.9). During incubation, due to the formation of specific base pairings, cDNAs from either sample will hybridize to the spots on the microarray with complementary sequence. Subsequently, depending on whether a gene is more strongly expressed in one or the other cell type more red or green light will be emitted from the corresponding spot. For the example shown in Figure 2.9 a green spot would correspond to a gene higher expressed in tissue A and a red spot would correspond to a gene higher expressed in tissue B. A yellow spot would indicate genes that are equally strong expressed in both tissues. One major advantage of microarrays is that once the design of a chip is set up, it can be mass produced and subsequently applied to readily assess the transcriptome of any cell type (Hager 2006, Dalma-Weiszhausz et al., 2006).

2.4.2 EST sequencing

While comparably cheap when mass produced, microarrays have the disadvantage that one can measure only the expression level of genes whose sequence was previously spotted onto the array. An additional problem in microarray technology is the high level of noise introduce by unspecific hybridization of the labelled cDNAs to the spots on the microarray or fluctuations in the amplification of the sample cDNAs that sometimes precede the hybridization step (Sundaresh et al., 2005). Alternatively, instead of hybridizing the cDNAs with a microarray, they can directly be cloned into vectors and sequenced. While it would be desirable to obtain the full sequences of the mRNAs from the produced cDNAs this is not readily possible due to the following technical problems. First, standard dideoxy sequencing methods can resolve only sequences of up to roughly 500 base pairs while an average

Figure 2.9 – Production and usage of spotted microarrays


a)
b)
c)
Tissue mixture
extract mRNAs from tissue
make cDNA via reverse transcription and seperate it into clones
amplify and spot sequenced cDNAs of interest onto chip
Tissue A
Tissue B
extract mRNAs from tissue B
make cDNA via reverse transcription
lable cDNAs from source A with Cy3
lable cDNAs from source B with Cy5
Cy3
Cy5
hybridize labeled cDNAs from A and B with cDNAs on the chip
scan chip with laser scanner
red channel
green channel
superimpose red and green channel
Actual superimposed scanner image (www.gene-chips.com)

a) mRNAs from many different sources are pooled and reverse transcribed into cDNA. The cDNAs are sequenced and those of interest are spotted onto the array. b) To subsequently measure the difference in the transcriptome between cell types A and B mRNAs are extracted seperately from either sample, converted into cDNA and labelled with red and green dye respectively. c) The cDNAs are pooled and poured onto the microarray where they bind to their corresponding spots. If a given cDNA species is more abundant in sample A or sample the spot will emit primarily green or red light respectively. Spots that correspond to cDNAs that are equally frequent in either sample will emit yellow light (the mixture of red and green).

mRNA is between 2 and 3 kilobases long. Second, reverse transcriptase often produces truncated copies of the original mRNA (Das et al., 2001, Bashiardes et al., 2001). To overcome these problems and to assess the transcriptome of a cell in a high throughput fashion large numbers of cDNA clones are randomly picking and single shot sequenced. The obtained 200 to 500 bp long and error prone sequence fragments are referred to as expressed sequence tags or ESTs (Boguski et al., 1995). cDNAs that have been completely sequenced in dedicated experiments are referred to as full length cDNAs. In contrast to microarrays, ESTs have to be newly made and assessed for each cell type under investigation making the technology more costly and laborious.

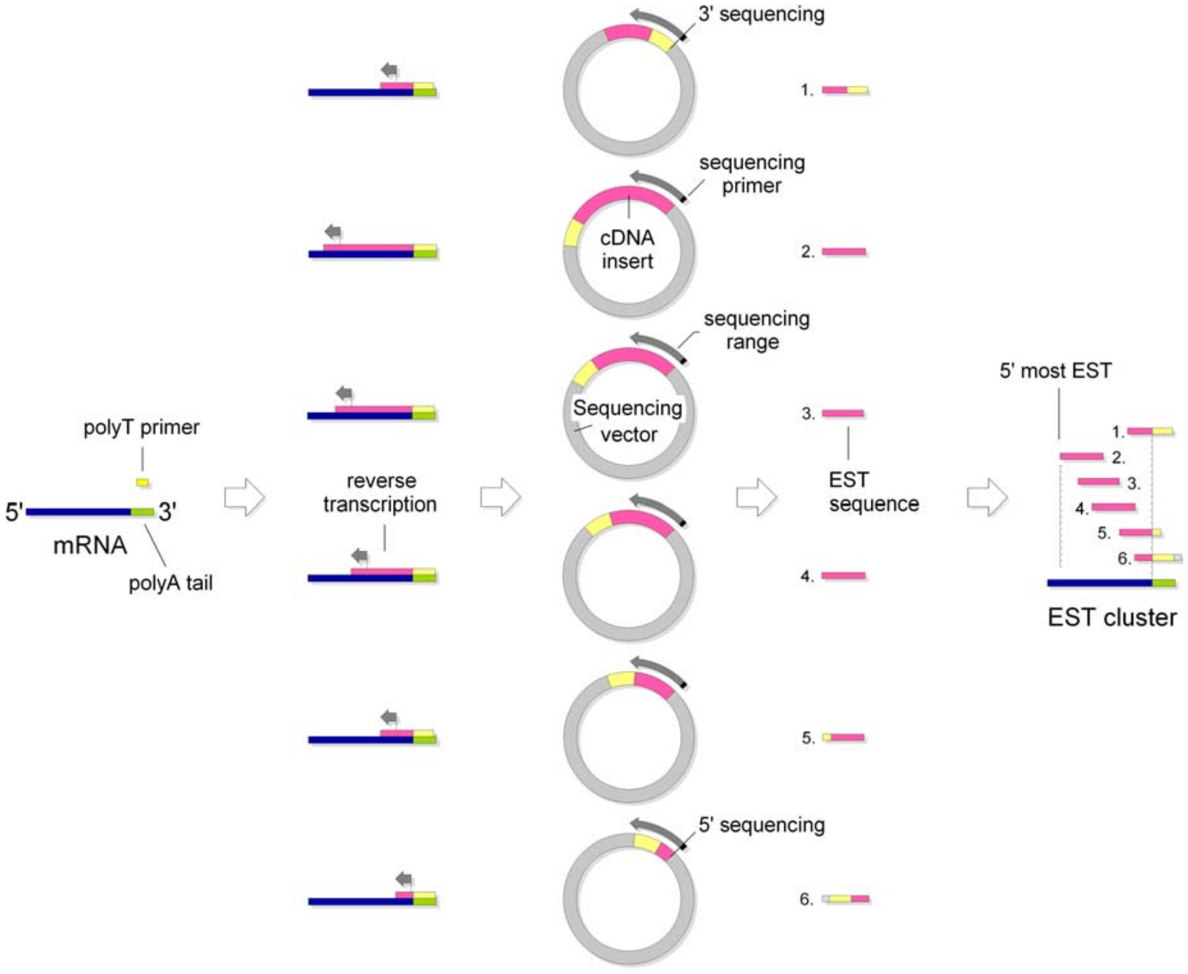
Clustering ESTs and inferring tissue specificity of a gene

The production of ESTs corresponds to sampling from the population of all possible mRNA subsequences. In general, the more ESTs are found for a given gene the higher will be its expression level and thus the contribution of its mRNA to the transcriptome of the cell. Before one can make statistical tests on the significance of the number of ESTs found for a gene it is necessary to accurately assign the ESTs to their genes (Haas et al., 2002). This task is not trivial even if the genome of an organism is known due to the presence of a large number of possible splice variants for many genes. However, if enough ESTs have been obtained so that their sequences overlap they can be clustered together to form a contig ideally spanning an entire mRNA (Figure 2.10). Although clustering is simplified if a fully assembeled genome is available mapping EST singletons to a particular gene oftentimes remains a difficult task as a given EST might span an exon junction and thus cannot be aligned with the genome without the introduction of alignment-gaps. Once the ESTs have been assigned to the respective genes the expression level of a given gene can be estimated based on the number of ESTs assigned to it in respect to all ESTs obtained from the cell (Haas et al., 2000). Aside from making it possible to estimate the expression level of genes whose sequence is not known *a priori* this procedure also leads to the identification of the splice variants present in a cell. With the advent of high throughput pyrosequencing techniques, which lower the cost for obtaining large number of sequences, EST based methods will become ever more important for analysing the transcriptome of cells.

The GeneNest database

Over the last decade millions of ESTs have been produced from a large number of vertebrate tissues (Haas et al., 2002). By combining the sequences obtained from all these experiments a more and more reliable picture of the transcriptome present in the different tissues has been established. One EST repository that hosts information for vertebrate tissue is the GeneNest database located at http://genenest.molgen.mpg.de (Haas et al., 2002). In

Figure 2.10 – From mRNAs to EST clusters



Purified mRNAs are first transcribed into cDNA using a poly T primer as initiation points for reverse transcriptase. Due to incomplete reverse transcription the produced cDNAs will terminate at different points along their encoding mRNAs. The produced double stranded cDNA is cloned into a sequencing vector in either 3' or 5' direction. In the former case obtained sequences will start with the sequence corresponding to the polyA tail of the mRNA, in the latter case the start of sequences will correspond to the middle of the mRNA. If the obtained ESTs overlap they can be clustered into a contig. Alternatively, due to the availability of the full genomic sequence ESTs can be directly mapped to the genome.

GeneNest ESTs have been taken from the UniGene database and subsequently assembled into contigs reflecting parts of putative transcripts of the respective gene. For each EST GeneNest stores the tissue from which it was derived (colour coded squares on the left of Figure 2.11). A sample screenshot from the GeneNest webpage is shown in Figure 2.11 for the EST cluster representing Otx2 a known eye and pineal gland specific transcription factor.

All shown ESTs (indicated in blue and cyan) for this gene belong to the same contig, that is, each EST overlaps with at least one other EST from the cluster. In addition, full length cDNA sequences obtained in dedicated experiments are shown as green lines on top. We will return to the GeneNest database later on, for the analysis of tissue specific gene regulation.

2.5 Measuring transcription factor binding

2.5.1 Small scale experiments

Historically, binding sites of TFs have been identified in experiments dedicated to verifying the binding of TF to specific promoters. Most candidate promoters have been identified either via mutation studies or based on sequence conservation between orthologous target genes of the transcription factor in different species. To verify the binding of the TF to the candidate promoter various techniques are being employed.

Promoter bashing

In promoter bashing experiments the promoter of a gene is cloned into a vector and its ability to drive the expression of a reporter gene is investigated. The expression of the reporter gene given the full promoter is taken as base line. Then, subsequences of the promoter are deleted from the vector and the expression of the reporter gene is again measured. Subsections causing a strong change in promoter activity when deleted are candidate regions for TF binding (Engstrom et al., 2004).

Electrophoresis mobility shift assay (EMSA)

To further narrow down the binding region of the TF, pieces of the promoter are excised and run on an agarose gel with or without the presence of the transcription factor. If the factor binds to the DNA piece the DNA-protein complex migrates slower through the gel than the DNA piece alone. In addition, an antibody against the TF can be added thereby further increasing the molecular weight of the complex and further slowing the migration speed. The corresponding supershift of the DNA band on the gel is a clear indication that the TF of interest binds to the sequence under investigation (Garner et a. 1981, Fried et al., 1981).

DNase footprinting assay

For footprinting experiments purified TFs are mixed with DNA pieces and are allowed to bind. Subsequently DNAse is added which cleaves DNA at random positions. Only the region protected by the bound TF will be cleaved less often. After the partial digest the DNA fragments are separated on a gel whereby an area on the gel which is void of DNA bands indicates the region protected by the bound TF (Galas et al., 1978).

2.5.2 Large scale approaches

Recent advances in microarray and sequencing technology have made it possible to analyse the binding pattern of a TF all across the genome *in vivo* or *in vitro*. The following methodologies are currently applied:

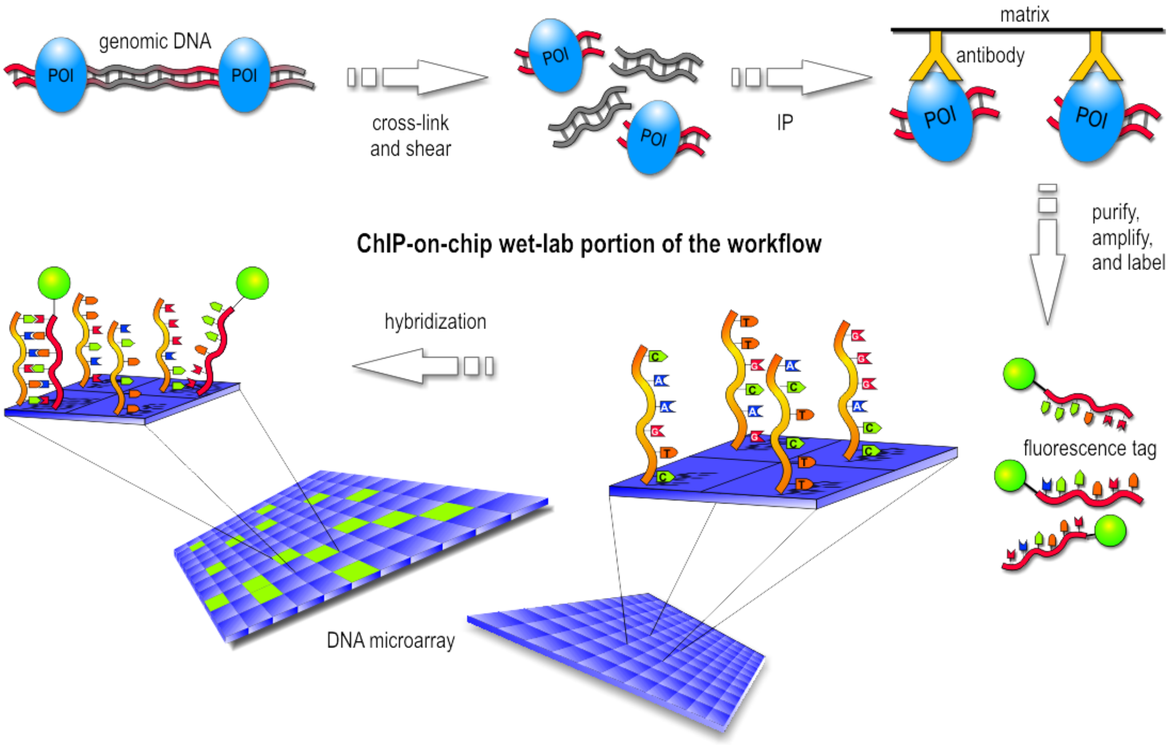
SELEX

In a SELEX (Systematic Evolution of Ligands by Exponential Enrichment, Tuerk et al., 1990, Ellington et al., 1990) experiment first purified TF molecules of given kind are immobilized on a matrix. In a second step, short random DNA oligos are synthesized and poured over the TFs. Oligos that resemble the binding site of the TF will be preferentially retained by the immobilized TF molecules while unbound oligos are removed by washing steps. The retained oligos are eluted, amplified and again poured over the TFs. The procedure is repeated several times each time enriching the oligos with those who tightly bind to the TFs. The oligos left at the end are sequenced. Traditionally, at least 6 rounds of purification are performed in order to limit the number of oligos that need to be sequenced. This in turn has the disadvantage that one obtains only the most strongly binding oligos. Therefore, many sites with intermediate affinity might not be recovered (Stoltenburg et al., 2007). Since high throughput sequencing techniques have become available this problem can be alleviated by sequencing all bound oligos already at earlier purification steps thus identifying also weaker bound oligos. A remaining problem is however that the binding behaviour of the immobilized TF *in vivo* does not always reflect the sequence preference *in vitro* for a number of reasons such as posttranslational modifications of the TF or the required presence of co-factors.

Protein binding arrays

For this technology purified transcription factor molecules are labelled with a fluorescent dye and poured onto a microarray (Mukherjee et al., 2004). The array can thereby contain for instance all known promoter regions of a given species. Importantly, the more labelled factors binds to a given promoter sequence the more light is emitted from the corresponding spot on the microarray. This readily allows identifying possible target genes for the TF. Alternatively, it is possible to spot all possible *k*-mers of a given length onto the array. In this case it is possible to determine precisely the affinity of the factor to all possible sequences of given length. Currently the binding preference for hundreds of TF is being investigated by the lab of Martha Bulyk (Mukherjee et al., 2004). While both approaches are well suited for

Figure 2.12 – Workflow of a ChIP-chip experiment



DNA is cross linked to bound TFs and sheared. Using a specific anti body the TFs and the DNA fragments bound to them are precipitated. The stronger a binding site for the TF the more of the corresponding fragments will be occupied across the used cells and thus the higher will be their fraction among all the pulled down pieces. The precipitated DNA is then amplified, labelled with a fluorescent dye and hybridized to a chip. (Image obtained from http://en.wikipedia.org/wiki/ChIP-on-chip).

detecting all potential binding sites of a TF they suffer from two shortcomings. For one, as in SELEX experiments, TFs are artificially expressed, that is, they might lack important post translational modifications present *in vivo*. Secondly, the potential binding pattern identified in this approach might diverge considerably from what is found in a given cell type due to the absence of other DNA binding proteins.

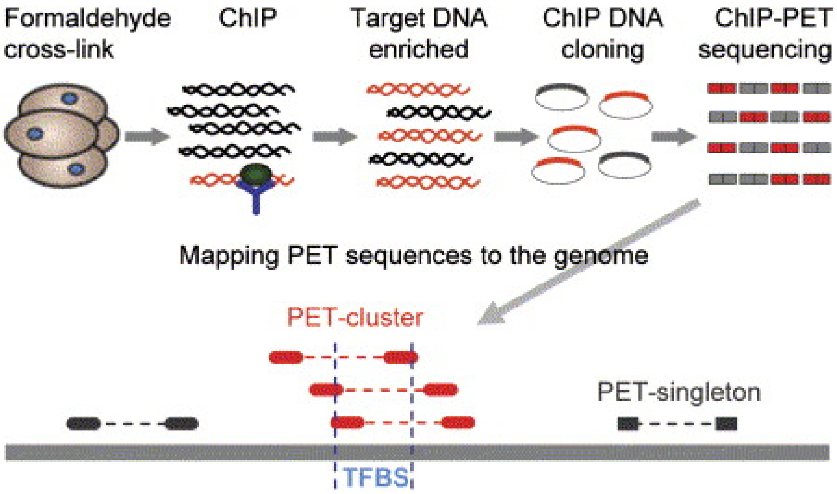
ChIP-chip

Chromatin immunoprecipitation followed by a microarray analysis (ChIP on chip) can be used to assess the binding behaviour of a TF across the genome *in vivo* (e.g. Harbison et al., 2004). To this end TF are cross-linked with the DNA by adding Formaldehyde to the cell culture. Subsequently, the genomic DNA from all cells in the sample is sheared and the TF of interest is immunoprecipitated using an immobilized antibody (see Figure 2.12). DNA pieces linked to the TF will thereby be coprecipitated with the factor while unbound DNA fragments will be washed away. Importantly, the more often a given genomic region is bound by a given TF *in vivo* the more of the corresponding DNA fragments will be retained. After the cross linking has been removed the purified DNA fragments are amplified and labelled with a red fluorescent dye. To minimize the effect of amplification bias the fragments are mixed with green labelled control fragments obtained from an aliquot of the shared DNA that was not subjected to the immunoprecipitation step. The fragment mixture is then hybridized to a microarray. Spots with sequences complementary to retained fragments will emit a higher portion of red fluorescent light compared to spots that do not correspond to sequences bound by the TF (Harbison et al., 2004). Chip-chip suffers from the same problems as normal microarrays. Due to the relatively high level of noise introduced by the amplification step and the hybridization a considerable number of false positive binding sites are usually predicted while true binding sites are often missed.

ChIP-PET

Similar to measuring the expression levels of genes the noise introduced by the microarray hybridisation step in ChIP-chip experiments can be avoided by employing high throughput sequencing techniques to directly obtain the sequences of the DNA fragments retained in the immunoprecipitation step. This procedure is referred to as chromatin immunoprecipitation followed by paired end tag sequencing (ChIP-PET, Wei et al., 2006). Again sequencing individual fragments corresponds to sampling from the population of all retained DNA fragments. Therefore, the more DNA fragments are sequenced the better will be the estimate on how frequently a given fragment is bound by the TF in the cells. While the procedure cannot determine the absolute fraction of bound fragments of a given type it can very accurately determine the relative binding fraction between different fragment species. The sequencing and mapping procedure for the retained DNA fragments is explained in detail in Figure 2.13 (Wei et al., 2006). Overlapping sequences mapped to the genome are referred to as PET clusters. The size of a cluster is thereby indicative of the strongly a given genomic site is bound by the TF. The binding site of the TF is thereby assumed to reside in the intersection of all sequences assigned to the PET cluster. Given much lower noise levels and the fast development of cheap high throughput sequencing techniques it appears likely that this technique will soon replace most microarray based approaches. However, at this moment there exist only a small number of data sets available from ChIP-PET or similar sequencing based procedures such as such as ChIP-seq.

Figure 2.13 – Experimental procedure of ChIP-PET



DNA and bound transcription factors are cross-linked in vivo and the DNA is extracted and sheared. Antibodies raised against the TF of interest are used to pull down the TF molecules and with them all bound DNA pieces. The stronger a binding site for the TF the more of these sites will be occupied across the used cells and thus the higher will be the fraction of this piece among all the pulled down fragments. Precipitated pieces are subsequently cloned into a plasmid. A restriction enzyme is used to cut out most of the inserted DNA leaving only about 20 base pairs of the insert on both the 5’ and 3’ end of the insert. The two short end tags are then ligated back together and sequenced. As is indicated on the bottom the fact that the end tags are paired - that is, on the genomic sequence they cannot lie further away from each other than a given maximal distance (depending on the intensity of the DNA shearing) – facilitates the mapping to the genome. Strong binding site for the TF are indicated by large clusters of PETs where the true binding site lies somewhere in the intersection of the mapped PETs. Statistical analysis is required to determine how big a PET cluster must be in order to infer a true binding site in its centre. (Image obtained from Wei et al., 2006).

The verified binding sequence of a TF obtained in any of the above experiments can be used to make computer based predictions of where in the genome a given TF might bind. To this end the sequences are aligned and a TF motif is generated. Later on, this motif is used to scan the genome and to identify potential new binding sites of the TF. Motif derivation and computer based binding site discovery will be the focus of the next chapter.