1. INTRODUCTION

1.1. General mechanisms of eukaryotic gene transcription

Multicellular eukaryotes consist of hundreds of highly specialized cell types which differentiate following synthesis and accumulation of different sets of RNAs and proteins. Most of the abundant proteins are present at very similar levels in the various cell types, indicating that differential expression of only a small number of genes is responsible for the dramatic structural and functional differences seen.

Regulation of gene expression can be achieved at the level of RNA transcription, processing, transport, degradation as well as translation. However, in most cases it is the initial transcriptional step which is the major regulatory mechanism.

In eukaryotes, RNA transcription is performed by three different but structurally related DNA-dependent RNA polymerases (Pol). Pol I synthesizes the large ribosomal RNAs. Pol III makes the 5S ribosomal RNA and other small non-coding RNAs, most notably the transfer RNAs. Pol II synthesizes the messenger RNAs (mRNAs) which will eventually be translated into protein and most of the small nuclear RNAs which are involved in pre-mRNA splicing (Woychik and Hampsey, 2002; Cramer, 2002).

Unlike the bacterial enzyme, the eukaryotic Pols require additional proteins for gene transcription to take place, adding an additional level of complexity and regulation. The control elements of protein-coding genes typically lie in their upstream promoter region (Figure 1). Many of these promoters contain a TATA box normally located 25 bp upstream of the transcriptional start site. It is recognised by the TATA box-binding protein which is part of the large TFIID basal transcription complex. Following TFIID binding, other basal transcription factor complexes are sequentially added, ultimately leading to recruitment of Pol II. For initiation of transcription, Pol II gets phosphorylated and dissociates from the complex to start transcribing RNA. Another motif less frequently found is the CCAAT box located upstream of the TATA box. Some genes lack a proper TATA box and use less conserved elements like the initiator sequence to direct Pol II transcriptional activity (Nussinov, 1990).

The specific control of gene expression is achieved via regulatory DNA elements present in more distal enhancer regions. These motifs are recognized by transcription factors (TF) usually able to interact with the basal transcriptional complex to modulate

its activity. Beside the 5' upstream region, enhancers can also map in introns or in the 3' part of a gene.

TFs make direct contacts with the DNA, mainly in the major groove. Multiple protein-DNA contact sites exist ensuring strong and specific interactions. Structural studies have revealed the existence of only a few structural motifs in the DNA-binding regions of TFs. One of the most common ones is the helix-turn-helix motif found in many prokaryotic and eukaryotic DNA-binding proteins. It is present in the homeodomain (HD) of homeobox proteins which play a fundamental role in eukaryotic development. Another DNA-binding motif often found is the zinc finger. Two such zinc fingers are present in the members of the nuclear receptor family which includes the five steroid receptors. Finally, the leucine zipper and helix-loop-helix motifs are found in other TFs and mediate both DNA binding and protein dimerization (Latchman, 1998).

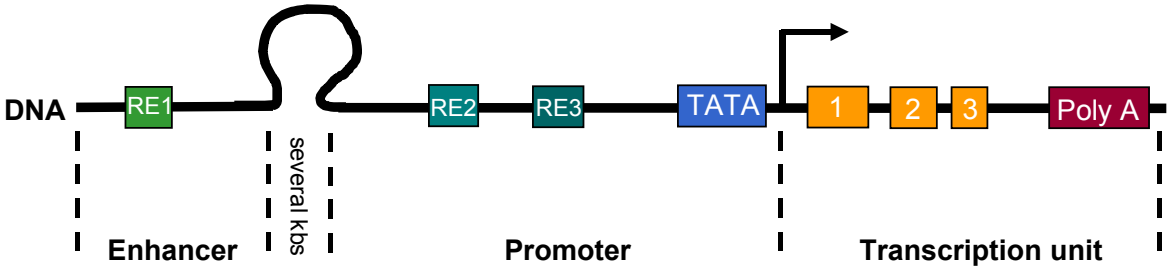

Figure 1: Schematic organization of a mammalian protein-encoding gene. *The transcription unit of the gene includes the transcriptional start site (arrow), three exons (1; 2; 3) and a poly-adenylation signal (Poly A). The regulatory portions of the gene are composed of the enhancer and the basal promoter which can be separated by several kbs. Both harbor DNA sequences that act as binding elements for transcription factors (RE1-3). The promoter normally contains a TATA box (TATA) to position Pol II for correct transcriptional initiation.*

1.2. General characteristics of the nuclear receptor family

Nuclear receptors are organized in a modular fashion and consist of four main domains. The N-terminal region contains one or several transactivation functions. The DNA-binding domain (DBD) is structured around two zinc fingers making contact with the DNA major groove. The hinge region is located C-terminal to the DBD and contains part of the nuclear localization signal (NLS). The C-terminal region of the receptor forms the ligand-binding domain (LBD) which is crucial for regulation of transcriptional activity.

Completion of the DNA sequence of the human and other genomes has revealed that the nuclear receptor family consists of 48 members in humans, 21 in *D. melanogaster* and more than 200 in *C. elegans* (Wilson and Moore, 2002). The human family can be subdivided into the high-affinity hormone binding receptors and the orphan receptors for which no or only low-affinity ligands are known (see Table 1; Wilson and Moore, 2002).

Trivial name	Official name	Natural ligand
ERα *	NR3A1	Estradiol
ERβ *	NR3A2	Estradiol
GR *	NR3C1	Cortisol
MR *	NR3C2	Aldosterone
PR *	NR3C3	Progesterone
AR *	NR3C4	Dihydrotestosterone
RARα	NR1B1	Trans-retinoic acid
RARβ	NR1B2	Trans-retinoic acid
RARγ	NR1B3	Trans-retinoic acid
TRα	NR1A1	Thyroxin
TRβ	NR1A2	Thyroxin
VDR	NR1I1	1,25-Dihydroxyvitamin D_3

Table 1: List of the principal human hormone nuclear receptors. *Twelve human hormone nuclear receptors are shown together with their natural ligand. The receptors marked with an asterisk belong to the steroid receptors, a subgroup of the nuclear hormone receptors.*

The steroid receptor family is an important and extensively studied subgroup of nuclear receptors. It is composed of the estrogen receptors α and β (ERα, ERβ), the progesterone receptor (PR), the androgen receptor (AR), the glucocorticoid receptor (GR) and the mineralocorticoid receptor (MR). All are ligand-dependent TFs with important roles in physiological processes as diverse as sexual development and reproduction, immunological response, mineral homeostasis and behavior (Lucas and Granner, 1992; Tsai and O'Malley, 1994).

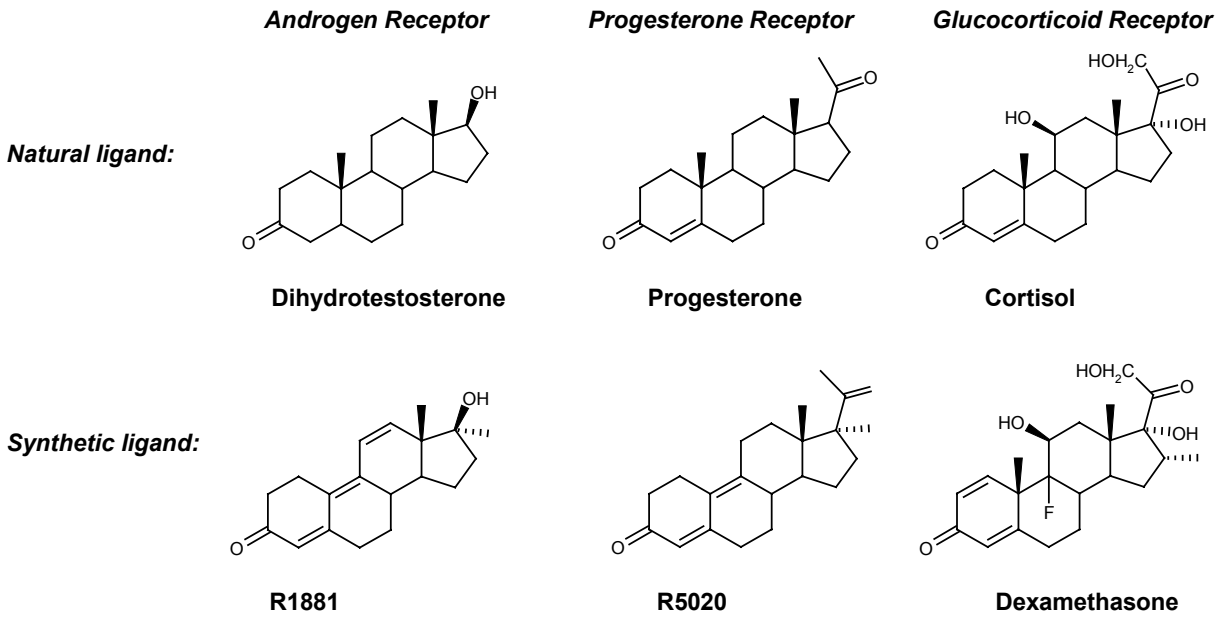

Figure 2: Ligands of AR, PR and GR. *The structures of the natural hormones and the synthetic ligands used in this study are shown.*

Since the cloning of the steroid receptors over 15 years ago, numerous studies have allowed to unravel the steps involved in receptor stimulation and activity. Briefly, the lipophilic ligand diffuses through the plasma membrane and binds to its specific receptor located in the cytoplasm complexed to chaperone proteins (Figure 3). This induces a conformational change leading to dissociation from the chaperones. The NLS thus gets exposed and translocation into the nucleus occurs. There the receptor binds to its cognate DNA response element as a homodimer. Cofactors are then recruited to form a multiprotein complex which affects chromatin structure and activates the preinitiation complex, thus leading to stimulation of gene transcription (Mangelsdorf et al., 1995; Latchman, 1998). For some steroid receptor target genes, repressive effects have been documented (McDonnell and Norris, 2002).

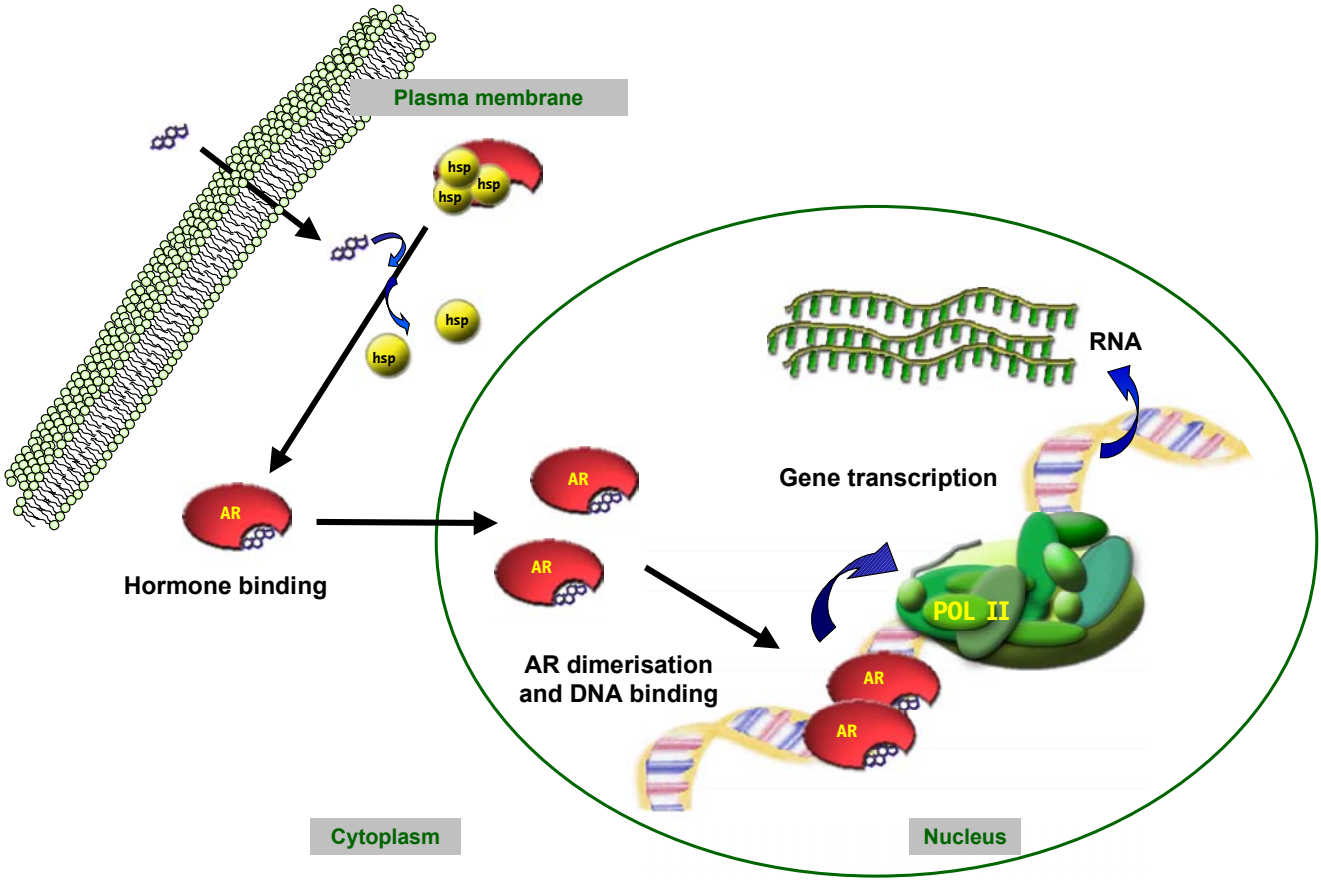

Figure 3: Overview of androgen action in mammalian cells. *Lipophilic androgen diffuses into the cell and binds to the androgen receptor (AR). Ligand binding induces conformational changes that lead to dissociation of the heat shock proteins (hsp). The receptor translocates into the nucleus where it dimerizes on the DNA response element, thereby stimulating gene transcription by Pol II.*

1.3. General characteristics of AR signaling

1.3.1. Structure of the AR

The *AR* gene is located on chromosome X and organized in eight exons (Figure 4). The first exon codes for the N-terminal domain (NTD) which contains the main transcriptional regulatory regions of the protein. Exons two and three code for the DBD, which comprises two zinc fingers. Exons four to eight code for the C-terminal LBD which contains the binding pocket for the two natural AR ligands, testosterone and dihydrotestosterone (Figure 2). The three-dimensional structure of the AR LBD reveals the presence of a fold resembling that of other steroid receptors (Matias et al., 2000). The LBD is structured in 12 helices which form the ligand-binding pocket. Helices 3, 5, and 11 interact directly with the hormone (Sack et al., 2001). For the ER it is known that in the agonist-bound state, helix 12 closes the pocket thereby

exposing new protein-binding sites which interact with cofactors (Shiau et al., 1998). A similar mechanism is anticipated for the AR. It is however thought that beside cofactor recruitment, the shift of helix 12 triggers an N-C-terminal interaction within the AR (He et al., 1999). Another difference to the other steroid receptors is the strong transactivation potential of the NTD of the AR (MacLean et al., 1997). Indeed the primary transactivation function (TAF) of the AR is located in the NTD while the LBD harbors a comparatively weak TAF (Simental et al., 1991).

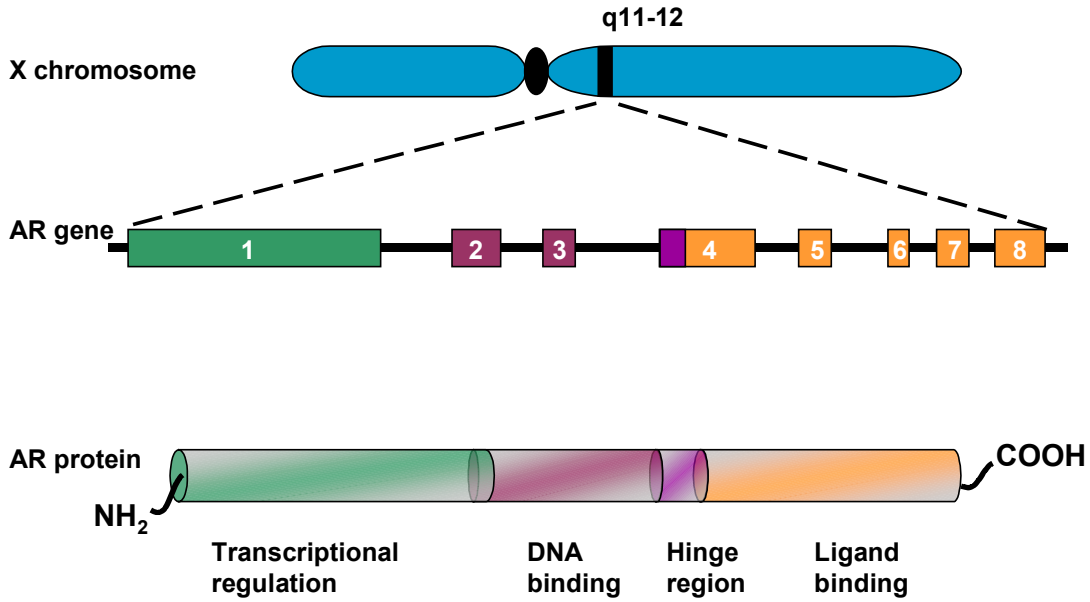

Figure 4: Human androgen receptor gene and protein. *The AR gene is located at q11-12 on the human chromosome X. It consists of 8 exons (1-8). The exons code for four distinct domains of the AR protein: The N-terminal domain (green) which harbors transactivation functions, the DBD (red) responsible for DNA binding, the hinge region (pink) which bears part of the NLS and the LBD (orange) responsible for ligand binding.*

The DBD is the best conserved region within the steroid receptor family. The DBD protein sequence is 79% identical between the AR and PR, and 76% between the AR and GR. From the crystal structure of the GR (Luisi et al., 1991) it appears that the more N-terminal zinc finger directly binds to the DNA response element. The second zinc finger is thought to be important for homodimerization by formation of salt bridges between two monomers and stabilization of DNA binding. It was originally supposed that the function of the DBD was solely to tether the AR to its DNA element, but this view has recently been challenged by the findings that the

DBD also plays a role in restricting self-synergy and in determining receptor selectivity (Liu et al., 1996; Schoenmakers et al., 1999).

1.3.2. Cofactors of the AR

The transcriptional activity of the AR depends on cofactors that regulate transcription by remodeling chromatin or by affecting the recruitment of the basal transcriptional machinery (Gelmann, 2002; Freedman, 1999; Haendler, 2002). These factors can either enhance or repress gene transcription.

The p160 family of cofactors (SRC-1, TIF2, SRC-3) plays an important role in mediating the activity of all steroid receptors. These factors have limited histone-acetylase activity and can recruit the histone acetylase CBP/p300, thus leading to chromatin remodeling into a transcriptionally active state (Leo and Chen, 2000). They interact with the LBD of most steroid receptors via their LXXLL motifs and additionally bind to the AR N-terminus (Bevan et al., 1999).

ARA70 and ARA55 are coactivators which display some degree of selectivity towards the AR (Yeh et al., 1999). ARA55 harbors an FXXLF motif probably implicated in interaction with the AR LBD. The molecular mechanism of ARA55 and ARA70 for stimulating AR function remains unclear.

A growing number of steroid receptor cofactors is implicated in ubiquitinylation and sumoylation (Conaway et al., 2002; Poukka et al., 1999). Ubiquitin and SUMO are peptide chains that can be covalently attached to many proteins. Beside their role in protein destruction, ubiquitin and SUMO are also implicated in regulation of RNA synthesis (Conaway et al., 2002; Courey, 2001). The SUMO-1 E2 ligase Ubc9 and members of the PIAS family of SUMO-1 E3 ligases interact with the AR DBD. AR is covalently modified by SUMO-1 in an androgen-enhanced fashion (Poukka et al., 2000). The acceptor sites reside in the N-terminus of AR and are identical with the transcriptional synergy control motifs that restrict the full transactivation potential of AR (Iniguez-Lluhi et Pearce, 2000).

Many additional cofactors regulating AR function have been identified by various methods including two–hybrid screens and coimmunoprecipitation. A comprehensive list can be found in the AR database in the web (www.mcgill.ca/androgendb).

1.3.3. DNA response elements

1.3.3.1. Consensus response elements

Two different consensus DNA response elements for steroid receptor binding have been identified, the estrogen response element (ERE) with the sequence 5´-GGTCAnnnTGACC-3´ for ERα/β and the steroid response element (SRE) with the sequence 5´-GGTACAnnnTGTTCT-3´ recognized by the other four steroid receptors: PR, AR, GR and MR (Lucas and Granner, 1992; Tsai and O'Malley, 1994; Freedman, 1992). The lack of selectivity of the SRE is explained by the high sequence conservation of the DBD of these four receptors, especially in the P- and D- boxes (Tsai and O'Malley, 1994; Freedman, 1992). Mutational analysis of steroid receptors identified amino acids of the first zinc finger α-helix that establish contact to specific nucleotides of the DNA element and other residues that prevent the receptor from binding to improper sites (Luisi et al., 1991). How the promiscuous SRE differentially transmits cues originating from steroid receptors with partially overlapping expression patterns but very different functions has so far not been elucidated.

1.3.3.2. Selectivity of androgen response

Specific steroid hormone action can be mediated through non-selective DNA elements in a variety of ways. The respective levels of receptor and hormone available in a given cell or tissue may play an important role (Strähle et al., 1998; Rundlett et al., 1995). Another mechanism is provided by differential chromatin remodeling as documented for the AR and GR, using the MMTV promoter (List et al., 1999). Several examples where the interplay of a steroid receptor with other transcription factors accounts for discriminating effects have furthermore been reported (Celis et al., 1993; Lu et al., 2000). The preferential interaction with cofactors represents another possibility but only a few receptor-specific ones have yet been identified (Yeh and Chang, 1996; Müller et al., 2000). Finally, cooperation among weak SREs and with auxiliary elements might lead to selective stimulation as shown for the sex-limited protein (Slp), the probasin (Pb), the prostate-specific antigen and the 20 kDa protein genes (Adler et al., 1993; Kasper et al., 1994; Cleutjens et al., 1996; Nelson et al., 1997; Ho et al., 1993). Studies on the Slp promoter furthermore substantiate that interactions between specific AR regions are

essential to enhance cooperativity at suboptimal DNA binding elements (Scheller et al., 1998).

Comprehensive *in vitro* selection procedures based on binding affinities have shown the optimal AR response element to be virtually identical to the SRE (Roche et al., 1992; Nelson et al., 1999). However highest DNA binding is not necessarily the primary property by which gene control is achieved *in vivo*. Indeed, suboptimal binding sites have been shown to allow specificity for a given steroid receptor by preventing recognition by other receptors (Verrijdt et al., 2002). A survey of natural functional androgen response elements (AREs) shows that they generally vary from the consensus SRE and display less binding affinity to the AR (Kokontis and Liao, 1992; Chang et al., 1995). Consequently, several such elements are frequently present in androgen-regulated genes and may be necessary for full-scale stimulation (Adler et al., 1993; Scarlett and Robins, 1995; Kasper et al., 1994; Cleutjens et al., 1996).

Very recently it has become apparent that unique variations within the DNA-binding sequence may have a dramatic impact on the recognition by a given nuclear receptor. Thus, response elements displaying androgen versus glucocorticoid selectivity have been identified in the promoter of the Pb, the secretory component (SC) and the Slp genes (Verrijdt et al., 2000; Claessens et al., 1996; Verrijdt et al., 1999). Unlike the SREs which form imperfect inverted repeats of the 5´-TGTTCT-3´ half-site, these selective androgen response elements have direct repeat features. The fact that these selective AREs do not exactly fit the consensus SRE, suggests that individual sequence variations may be instrumental in specificity of gene control.

1.3.4. AR mutations and diseases

Male sex differentiation is orchestrated by time-dependent androgen biosynthesis and action. Most critical for the differentiation of the male internal and external genitalia is the development of testis from the bipotential indifferent gonad (Hughes et al., 2001). The Wolffian duct derivatives (vas deferens, epididymis and seminal vesicle) as well as prostate growth depend on androgens. In fact the complete sexual development of males is coordinated by ligand-activated AR (Hughes et al., 2001).

The *AR* gene is located on chromosome X and therefore present only as a single copy gene in males. Mutations in this gene will thus cause a direct phenotypic manifestation not compensated by a wild-type codominant allele. Several diseases

have been correlated with mutations of the *AR*. Alterations that disturb AR ligand binding or AR transactivation function lead to partial or complete androgen insensitivity (Hughes et al., 2001). The phenotype of theses genetically male individuals spans from mild to complete infertility with otherwise normal male characteristics. In more severe cases, complete sex reversal with an external female phenotype has been reported (Hughes et al., 2001). Kennedy's disease (also called spinal and bulbar muscular atrophy) is caused by an extension of the CAG repeat present in the AR NTD (La Spada et al., 1991). This leads to the formation of protein aggregates in certain brain regions and consequently to brain damage. Finally many AR mutations have been reported in patients suffering from prostate carcinoma, especially at late stages of the disease and following anti-androgen treatment. Most of them are located in the LBD and in several instances they lead to alterations in the action of agonists and antagonists. For instance, the exchange of tyrosine to alanine at position 877 found in the LNCaP cell line makes the AR responsive to some anti-androgens (Sack et al., 2001). In other cases, stimulation of AR mutants by glucocorticoids has been documented (Zhao et al., 2000; Matias et al., 2002). Finally amplification of the whole *AR* gene has been reported in about 30% of late stages of prostate carcinoma (Visakorpi et al., 1995; Rini and Small, 2002).

1.4. The homeobox transcription factor family

Homeobox genes code for TFs that share a characteristic 60 amino-acid-long DNA-binding region, the HD. They act as master switches in developmental processes and are intimately associated with embryogenesis, as evidenced by extensive mutational analyses in the fly and mouse (Treisman et al., 1992; Gehring et al., 1994a). A critical role in development and functionality of reproductive organs has been reported for several homeobox genes. Examples include *Pax2* and *Lhx9* which are essential for gonad formation, and *Hoxa*-10 and *Hoxa*-11 which are necessary for testicular descent and implantation (Torres et al., 1995; Birk et al., 2000; Satokata et al., 1995; Gendron et al., 1997).

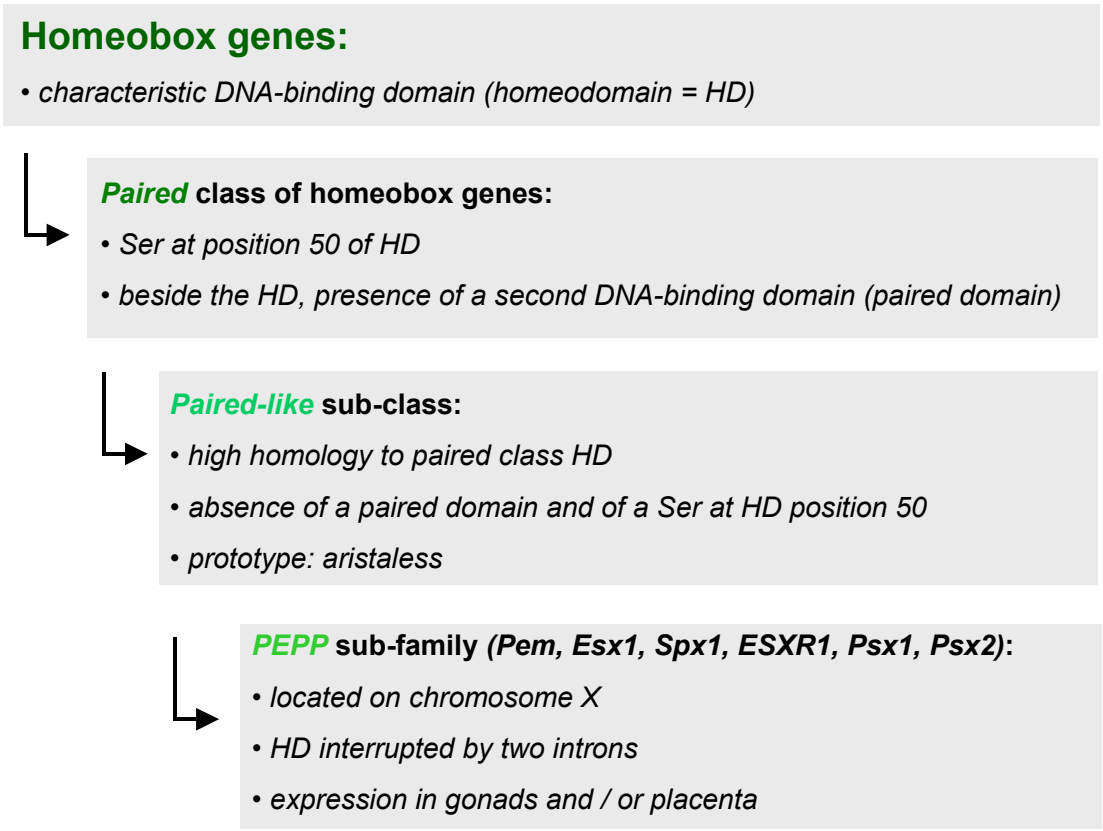

Figure 5: Classification of the PEPP homeobox genes. *Characteristic features of the different groups are given.*

Homeobox proteins are grouped in different classes (Figure 5), depending on their HD sequence and on the presence of additional motifs (Treisman et al., 1992; Gehring et al., 1994a). The most important ones are the HOX and the paired classes. The paired class is characterized by the presence of a serine at position 50 of the HD

and of an additional DNA-binding region, the paired domain (Figure 5). Absence of both these features is characteristic of the paired-like class for which the *Drosophila* aristaless protein is the prototype (Galliot et al., 1999).

1.4.1. The PEPP subfamily of the paired-like class of homeobox genes

The PEPP subfamily has recently emerged as a subset of the paired-like class of homeobox proteins (Maiti et al., 2001). It includes Pem, Esx1, Spx1 (a splice variant of Esx1), ESXR1, Psx1 and Psx2 (also called Gpbox) (Sasaki et al., 1991; Maiti et al., 1996a; Li et al., 1997; Branford et al., 1997; Fohn and Behringer, 2001; Han et al., 1998; Han et al., 2000; Takasaki et al., 2000). All the PEPP genes are located on chromosome X and share a similar structure, strongly suggestive of a common ancestor (Maiti et al., 2001; Takasaki et al., 2000).

Pem is the founding member of the PEPP homeobox subfamily. Its expression is controlled by two promoters which are independently regulated in a tissue-specific manner (Maiti et al., 1996a; Sutton et al., 1998). The proximal promoter is androgen-dependent and controls expression in the epididymis and testis. The distal promoter is androgen-independent and primarily expressed in the female reproductive tissues (placenta and ovary) as well as testis and muscle (Maiti et al., 1996b).

The function of the various PEPP proteins is largely unknown. Mice with a disrupted *Pem* gene have remarkably few phenotypic alterations when reared under normal laboratory conditions, suggesting redundant functions with other homeobox genes (Pitman et al., 1998). *ESXR1* is the only human *PEPP* gene identified and might represent the orthologue of murine *Esx1*. The existence of five murine *PEPP* genes suggests more human members with possibly important functions in human sexual development and reproduction to exist.

1.4.2. Homeobox response elements and specificity

Similarly to steroid receptors, the homeodomain proteins are very promiscous in their DNA-binding specificity. Most recognize the consensus DNA sequence ATTA (Gehring et al., 1994b), in line with the high conservation of their DBD. Given the essential but very distinct roles homeodomain proteins play during development, target gene specificity must be achieved *in vivo*. Possible mechanisms may be heterodimerization with other TFs, interactions with cofactors (activators or repressors) or posttranslational modifications (Pinsonneault et al., 1997; Ostendorff

et al., 2002; Biggin and McGinnis, 1997; Mann, 1995). In addition subtle differences in DNA-binding affinity and specificity may be instrumental in gene-specific action of homeobox proteins (Gehring et al., 1994b).

1.5. Aim of current work

AR function is essential for male development and fertility and is also implicated in several comparatively frequent pathologies. Most importantly, the AR plays a crucial role in prostate cancer, one of the most common cancer forms in men. In addition, germ-line AR mutations leading to partial or complete androgen insensitivity syndrome have been described.

The AR is a ligand-dependent TF which can be controlled by synthetic compounds and therefore represents a drug target for therapy. The mechanisms underlying gene regulation by the AR are only beginning to be unraveled as only few directly regulated genes are known and characterized. One aim of this work was therefore the identification of new androgen-regulated genes and the detailed analysis of their control elements.

The murine *Pem* gene was chosen as a model, due to the strong and tight regulation of its expression in typical androgen target organs. The identification of novel and selective response elements in the *Pem* promoter opened up the possibility that DNA elements were able to induce allosteric changes onto bound AR and thereby to influence cofactor recruitment as a way to achieve steroid receptor-specific gene control.

In this work, a detailed comparative analysis of the response elements found in the *Pem* promoter and in other androgen-regulated genes is presented to test this hypothesis (Figure 6).

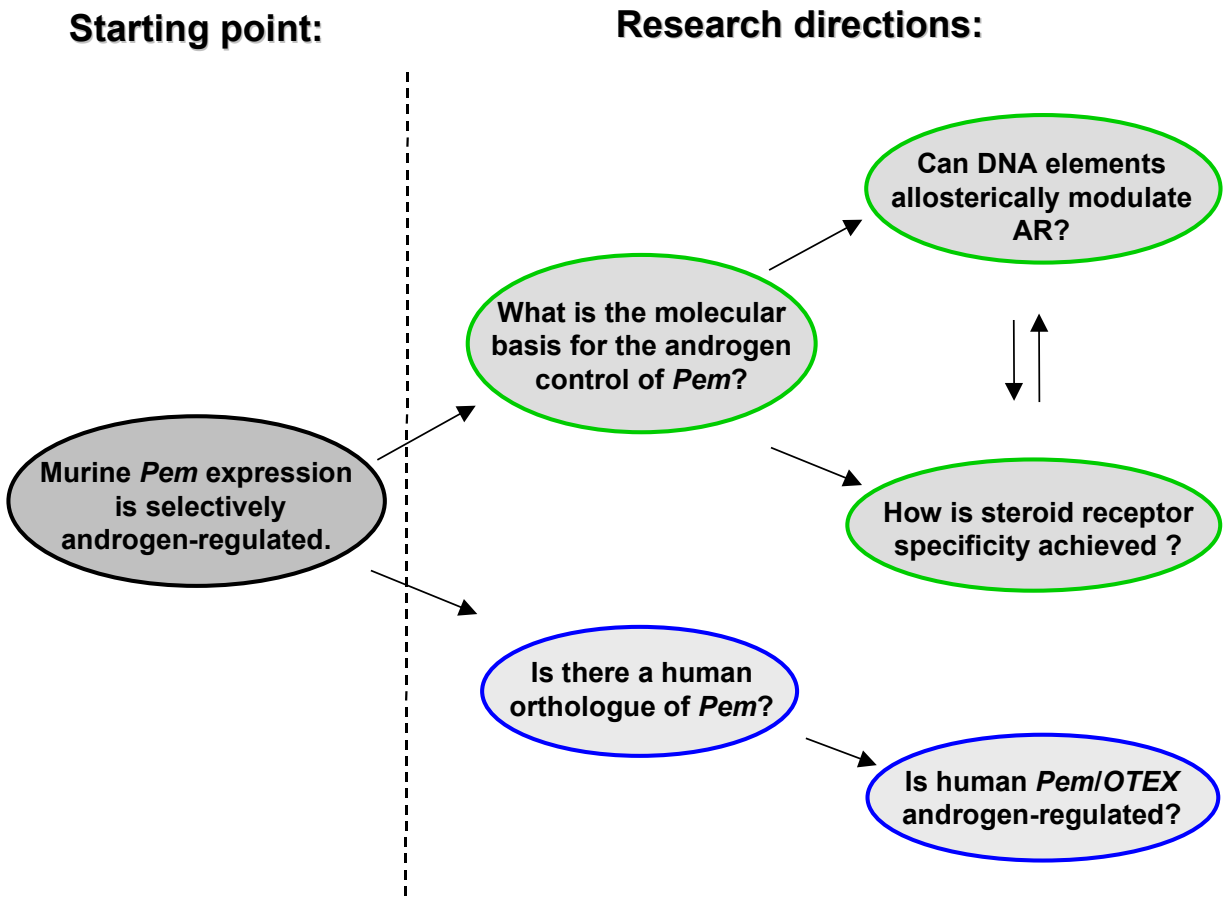

Figure 6: Schematic overview on the aim of this work.