J Mol Evol (1998) 47:334–342

JOURNAL OF MOLECULAR EVOLUTION

Evolution of the Primate Androgen Receptor: A Structural Basis for Disease

Catherine S. Choong,[1,*] Jon A. Kemppainen,[1] Elizabeth M. Wilson[1,2]

[1] Laboratories for Reproductive Biology and Department of Pediatrics, University of North Carolina, Chapel Hill, NC 27599, USA
[2] Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC 27599, USA

Received: 30 December 1997 / Accepted: 9 February 1998

Abstract. Androgen effects mediated by the androgen receptor (AR) are essential for male reproductive development and virilization. Comparison of AR DNA coding sequence from five primate species, *Homo sapiens* (human), *Pan troglodytes* (chimpanzee), *Papio hamadryas* (baboon), *Macaca fascicularis* (macaque), and *Eulemur fulvus collaris* (collared brown lemur), supports their phylogeny with complete conservation of the DNA and steroid binding domain protein sequence. A linear increase in trinucleotide repeat expansion of homologous CAG and GGC sequences occurs in the NH_2-terminal transcriptional activation region and is proportional to the time of species divergence. A serine phosphate/glutamine repeat interaction is observed where increasing CAG repeat length is associated with an increased rate of serine 94 phosphorylation. Disparity in the calculated and apparent molecular weight with CAG repeat expansion of an AR NH_2-terminal fragment suggests self-aggregation with increasing glutamine repeat length into the pathological range. These results suggest that a CAG/glutamine repeat expanded during divergence of the higher primate species, which may have a direct effect on AR structure and support a common pathway in CAG trigenic diseases in the pathophysiology of neurodegeneration observed in X-linked spinal bulbar and muscular atrophy.

Key words: Androgen receptor — Steroid receptor phosphorylation — Trinucleotide repeats — Glutamine repeats — Primate evolution — Neurodegeneration — Spinal and bulbar muscular atrophy — Transcriptional regulatory proteins

Introduction

The human androgen receptor (AR) is a transcriptional regulatory protein that is activated by androgen binding and is required for the development of male secondary sex characteristics. The AR is a member of a family of nuclear proteins of which the steroid receptors are a subgroup. In addition to its DNA and steroid binding domains, the AR has an NH_2-terminal region required for transcriptional activation (Simental et al. 1991) where the amino acid sequence diverges most markedly from that of other steroid receptors. Within the first exon of the AR gene that codes for the NH_2-terminal region are several trinucleotide repeat sequences of unknown function. A glutamine repeat beginning at codon 58 in human AR is polymorphic in the human population, with an average repeat length of 22 ± 2 residues (Edwards et al. 1992). Its expansion beyond 40 repeats is associated with spinal/bulbar muscular atrophy or Kennedy's disease (La Spada et al. 1991), while shorter CAG repeat lengths were found to correlate with increased risk of aggressive prostate cancer (Giovannucci et al. 1997; Hakimi et al. 1997; Irvine et al. 1995). We performed sequence analysis of the AR DNA in several primate species to determine critically conserved regions and to investigate the evolution and functional consequences of polymorphic variation within the AR. We examined the effect of varying glutamine repeat lengths on AR mobility in sodium

* *Present address:* Department of Paediatric Endocrinology, Mater Children's Hospital, Annerley Road, South Brisbane, 4101 Australia-
Correspondence to: E.M. Wilson; *e-mail:* emw@med.unc.edu

dodecyl sulfate (SDS) gels and receptor phosphorylation to investigate possible consequences of CAG expansion on AR protein structure and function.

Materials and Methods

DNA Sequence Analysis. DNA was extracted (John et al. 1991) from peripheral white blood cells from three female chimpanzees (animal 407, 155, and 116), *Pan troglodytes;* six wild-caught macaques, *Macaca fascicularis;* six unrelated male baboons, *Papio hamadryas;* and five male lemurs of the following subspecies—*Lemur catta* (6268; ring-tailed lemur), *Eulemur fulvus collaris* (6225; collared brown lemur), *Eulemur fulvus sanfordi* (6391; Sanford's brown lemur), *Eulemur macaco macaco* (6274; black lemur), and *Varecia variegata rubra* (6310; red ruffed lemur). Blood collections were performed using procedures approved by the participating institutions. Exon 1 was amplified using internal exon primers (Lubahn et al. 1989) and subcloned in pGEM-T (Promega) (Choong et al. 1996b). Flanking intron primers were used to amplify exons B–H for direct sequencing of PCR products. Analysis of lemur DNA required minor modification. Exons D and H were analyzed as described above using intronic human primers (Lubahn et al. 1989). Exons B, C, and E–G could not be amplified by this method, most likely because of the differing flanking intron sequence in lemurs compared with other primates. Total RNA was reverse transcribed using the AMV Reverse Transcription System (Promega) according to the manufacturer's instructions and oligo-(dT) primers. cDNA was amplified in a standard PCR reaction (Choong et al. 1996b) with the following specific primers: AR-LEM-1790-25-S, GTAACCCCCTATGGTTACTCTCGGC; and AR H2-29-AS, GGAACATGTTCATGACAGACTGTACATCA. Lemur PCR product was sequenced with AR-LEM-1790-25-S and AR-6225LEM-2559-23-S: CAGATGGCAGTCATTCAATATTC. The full AR DNA sequence was determined from two chimps (407 and 155) and one baboon (Bert Papham); macaque (4086), and lemur (6225) at the University of North Carolina at Chapel Hill Automated DNA Sequencing Facility on a Model 373A DNA Sequencer (Applied Biosystems) using the Taq DyeDeoxy Terminator Cycle Sequencing Kit (Applied Biosystems).

Data Deposition. Primate AR sequences were submitted to Genbank under the following accession numbers: *Pan troglodytes,* U94177; *Papio hamadryas,* U94176; *Macaca fascicularis,* U94179; and *Eulemur fulvus collaris,* U94178.

Cell Culture and Labeling Studies. [^{35}S]Methionine/cysteine labeling of transiently expressed AR was performed (Kemppainen et al. 1992) by incubating monkey kidney COS-1 cells (1.2×10^6/10-cm dish) with 100 μCi of [^{35}S]methionine/cysteine Trans label for 30 min at 37°C. Cells were washed in phosphate-buffered saline and incubated in complete medium for 1.7 h at 37°C for isotope dilution. The cells were also incubated with 150 μCi of [^{32}P]orthophosphate for 2 h at 37°C and harvested in RIPA buffer (Kemppainen et al. 1992), and the AR was immunoprecipitated using antipeptide antibody AR52 and analyzed on 9 or 12% acrylamide gels containing SDS.

Results

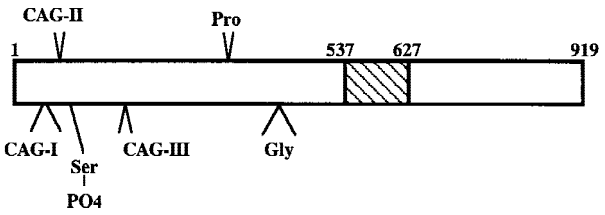
AR Sequence Comparison. Within the AR NH_2-terminal region are polymorphic triplet repeats of unknown function (Fig. 1A). These include CAG repeats coding for glutamine and a GGC repeat coding for glycine. Of these, CAG-I is highly polymorphic in the human population (Edwards et al. 1992), and its expansion beyond 40 repeats is associated with X-linked spinal/bulbar muscular atrophy, a neurodegenerative disease in adult males (La Spada et al. 1991). We found that CAG-I (human population range, 11–31; mean, 22 ± 2) (Edwards et al. 1992) decreases exponentially ($r = 0.947$) with evolutionary distance from human (Figs. 1B, 2B, and 3A). The less polymorphic GGC repeat, preceded by $(GGT)_3GGG(GGT)_2$, which together code for 24 consecutive glycine residues in human AR, also shortens inversely and exponentially ($r = 0.974$) with evolutionary distance from human, undergoing concurrent GGC and GGG transition to GGT as the GGC repeat expands (Figs. 1B, 2B and C, and 3B and C). The shorter GGC repeats in lower primates do not alter AR mRNA expression (results not shown) as we observed previously for the CAG-I repeat, where increasing CAG repeat length decreased AR mRNA expression (Choong et al. 1996a).

Other trinucleotide repeat sequences in the AR coding region include an eight-residue proline repeat that is conserved among primates despite nucleotide changes in the CCG/CCC/CCT iteration (Figs. 1B and 2B and C). In the rat the proline repeat is interrupted by histidine and flanked by threonine (Figs. 1B and 2C). A single leucine codon preceding CAG-I in rat, lemur, macaque, and baboon is expanded to three leucine residues in chimpanzee and four leucine residues in human (Figs. 1B and 2B). CAG-II and CAG-III are more stable through species divergence, although in rat, CAG-II is replaced by a five-residue arginine (CGG) repeat and CAG-III is expanded to 22 repeats (Figs. 1B and 2B). At a similar sequence in human AR (residues 112–118), where the rat glucocorticoid receptor coding sequence has 21 CAG repeats (Gearing et al., 1993), rat and lemur AR CAG codons exceed by one the four closely spaced glutamine residues in the higher primates.

An independently corroborated phylogeny with known genes from six taxa shows three internal branches supported by a high level of confidence indicated by 100% bootstrap replication frequencies (Fig. 4). The pattern of evolution for the major repeats mapped onto the noncontroversial phylogeny shows an increased repeat number along the branches leading to the primates, especially the hominids. CAG-I and GGC repeats expand predominantly after branching of Old World monkeys, whereas CTG repeat expansion was more recent.

The NH_2-terminal 1–53 and 360–429 amino acid residues in the human AR are completely conserved among the primates studied. As these regions have not been found critical for AR transactivation (Zhou et al. 1993), their high conservation may reflect the complete sequence conservation of the steroid binding domain (Fig. 1B), as we demonstrated previously that these NH_2-terminal regions interact with the androgen-bound AR steroid binding domain during AR dimerization (Langley et al. 1995). The greatest divergence of amino acid sequence occurs in exons A and D, whereas exons G and H

A



B

```
                a                                                                                                I                  II
human     1     MEVQLGLGRVYPRPPSKTYRGAFQNLFQSVREVIQNPGPRHPEAASAAPPGASLLLLQQQQQQQQQQQQQQQQQQQQQQETSPRQQQQQQ-GED
chimp                                                                                                  --QQQQQQQQQQQQQQQQQQQQQQ     QQQQQ--
baboon                                                                                                 ----------------QQQQQQQQQQQ     QQQQQQ-
macaque                                                                                                ----------------------QQQQQQQQ     QQQQQQQ
lemur                                                                                            R     ----------------------QQQQ     PQQQQQQQ
rat       1                                        A              I     C     ----------------------QQRQ     RRRRQQHP

human     93    GSPQAHRRGPTGYLVLDEEQQPSQPQSALECHPERGCVPEPGAAVAASKGLPQQLPAPPDEDDSAAPSTLSLLGPTFPGLSSCSADLKDILSEAS
chimp                                      P
baboon                                     P                    G
macaque                                    P                    G                                                T
lemur               QS       A        Q         S         A      Q  P   S      V                                    G
rat       74         I  T     A E      Q   S G   S  L      T PG      P     Q                                  I       G
                              III
human     188   TMQLL----------------QQQQQEAVSEGSSSGRAREASGAPTSSKDNYLGGTSTISDNAKELCKAVSVSMGLGVEALEHLSPGEQLRGDC
chimp               ----------------QQQQQ                                        S
baboon              ----------------QQQQQ                                        S
macaque             ----------------QQQQQ                                        S
lemur               ----------------QQQQQ                  A        S          S                  T
rat       169       QQQQQQQQQQQQQQQQQQQQQ VI      V     T   S    S    N     S

human     266   MYAPLLGVPPAVRPTPCAPLAECKGSLLDDSAGKSTEDTAEYSPFKGGYTKGLEGESLGCSGSAAAGSSGTLELPSTLSLYKSGALDEAAAYQSR
chimp
baboon              V
macaque             V
lemur                  G                           D G  EP   T    S  Q            SE                        E   S
rat       264       S   G                  LS  EGP  G  E     S     A             SE        I  S        V       N

human     361   DYYNFPLALAGPPPPPPPPHPHARIKLENPLDYGSAWAAAAAQCRYGDLASLHGAGAAGPGSGSPSAAASSSWHTLFTAEEGQLYGPCGGGGGGG
chimp                                                                                                                -----GG
baboon                                                                                                               -------
macaque                                                                                                              -------
lemur                        L                    S         F        G  T           A                              -------
rat       359          S   H    T           S                        GSV   ST   P T                                 --------
                                                                                                                    ab
human     456   GGGGGGGGGGGGGGGGG-----EAGAVAPYGYTRPPQGLAGQESDFTAPDVWYPGGMVSRVPYPSPTCVKSEMGPWMDSYSGPYGDMRLETARDH
chimp           GGGGGGGGGGGGGGGGG-----                    G
baboon          ----GGGGGGGGGGGG---AG                     G
macaque         ----GGGGGGGGGGGG---AG                     G
lemur           ----------GGGGGG---TS     T    S         G  P       S V        S         E       V
rat       446   ------------GGGGGSSSPSD  P             S  G  S SE      V N       S         EN          DST
                          *                              bc                                  *  cd
hcbml     546   VLPIDYYFPPQKTCLICGDEASGCHYGALTCGSCKVFFKRAAEGKQKYLCASRNDCTIDKFRRKNCPSCRLRKCYEAGMTLGARKLKKLGNLKLQ
rat       529
                                                    →                                                          de
human     641   EEGEASSTTSPTEETTQKLTVSHIEGYECQPIFLNVLEAIEPGVVCAGHDNNQPDSFAALLSSLNELGERQLVHVVKWAKALPGFRNLHVDDQMA
chimp
baboon                       A
macaque                      A
lemur               A        SS
rat       624       N  AG     DPS  M
                                                ef                                                    fg
hcbml     736   VIQYSWMGLMVFAMGWRSFTNVNSRMLYFAPDLVFNEYRMHKSRMYSQCVRMRHLSQEFGWLQITPQEFLCMKALLLFSIIPVDGLKNQKFFDEL
rat       719
                                                      gh                                                   ←h
hcbml     831   RMNYIKELDRIIACKRKNPTSCSRRFYQLTKLLDSVQPIARELHQFTFDLLIKSHMVSVDFPEMMAEIISVQVPKILSGKVKPIYFHTQ   919
rat       814                                                                                                902
```

Fig. 1. AR amino acid sequence comparison among primates and rat. **A** Schematic of the AR coding sequence with trinucleotide repeat regions and DNA binding domain (*hatched area*) indicated in the 919-amino acid residue human AR. **B** AR amino acid sequence from human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), baboon (*Papio hamadryas*), long-tail or crab-eating macaque (cynomolgus monkey, *Macaca fascicularis*), collared brown lemur (*Eulemur fulvus collaris*), and Sprague–Dawley white rat. Absence of the single-amino acid code indicates sequence identity with the human AR except for glutamine (Q) and glycine (G) repeats, where the same residues are indicated and *dashes* signify amino acid omission. Amino acid numbering is indicated for the human (Lubahn et al. 1988) and rat (Tan et al. 1988) AR; CAG repeats I, II, and III, *above the sequence;* DNA binding domain boundaries, by *asterisks;* ligand binding domain, by *arrows;* and exon boundaries, by *lowercase letters.*

A. Nucleotide differences from human AR by exon

Exon	A*	B	C	D	E	F	G	H	nt total		aa total	
human AR aa	1- 537	538- 588	589- 627	628- 723	724- 771	772- 815	816- 868	869- 919	2757	%	919	%
human	0	0	0	0	0	0	0	0	0	--	0	--
chimp	7	0	0	0	0	0	0		7	99.8	3	99.7
baboon	28	0	2	5	1	1	2	0	39	98.6	8	99.1
macaque	27	2	2	5	1	1	2	1	41	98.5	9	99.0
lemur	138	5	7	17	13	4	3	4	191	93.0	49	94.7
rat	208	13	8	39	15	10	5	6	304	89.0	86	90.6

human AR aa	NH_2-terminal domain	DNA binding domain	hinge region	steroid binding domain
	1-558	559-624	625-676	677-919

B. Major trinucleotide repeats

	CTG(TTG)	CAG-I	CAG-II	CAG-III	CCG/CCC/CCT	GGT/GGC	n
	Leu	Gln	Gln	Gln	Pro	Gly/GGC	
human AR	54-57	58-78	84-89	193-197	372-379	449-472/455-472	
human	4	22	6	5	8	24/18	1
chimp	3	18 n=1 20 n=1 21 n=2 23 n=1	5	5	8	17/11 n=1 19/13 n=1	2
baboon	1	8 n=4 10 n=2	6	5	8	13/6	6
macaque	1	7 n=5 9 n=1	6	5	8	13/6	6
lemur	1	4	7 n=3 9 n=1	5	7	6/1 n=4 7/2 n=1	5
rat	1	2	2	22	6	6/2	1

C. Proline and glycine repeats

human	372	Pro	Pro	Pro	Pro	Pro	Pro	Pro	Pro	449	Gly	Gly	Gly	Gly	Gly	Gly	Gly	Gly
human		CCG	CCG	CCC	CCT	CCG	CCG	CCT	CCC		GGT	GGT	GGT	GGG	GGT	GGT	GGC	$(GGC)_{17}$
chimp													G					$(GGC)_{10,12}$
baboon						A							G	C	C	C	T	$(GGC)_6$
macaque						A							G	C	C	C	T	$(GGC)_6$
lemur		A	A				TT Leu		T		---		G	$T_{n=3}$ $C_{n=2}$	C	$C_{n=3}$ $T_{n=2}$		---
rat				A	C		C		A		G	CCA	A	C	G	C	C	---
rat	370	Pro	Pro	His	Pro	Pro	Pro	Pro	Thr	444	Gly	Pro	Gly	Gly	Gly	Gly	Gly	---

Fig. 2. Summary of AR nucleotide and amino acid sequence divergence. **A** AR sequence divergence is indicated by the number of nucleotide changes by exon, with amino acid boundaries indicated *above* and domain boundaries *below*. Lemur is *Eulemur fulvus collaris.* *Sequence changes in exon A exclude trinucleotide repeat expansions or contractions at CAG-I,-II, and -III and GGC. **B** AR trinucleotide repeat lengths of CAG-I, CAG-II, and GGC and amino acid repeat lengths of CAG-III, leucine, proline, and glycine. The glutamine repeat length is CAG_{n+1} at CAG-I for human, chimp, baboon, and macaque and at CAG-II for macaque because CAA coding for glutamine follows CAG-I and precedes CAG-II in these species. The number (n) of unrelated animals assayed for repeat length is indicated at the *right* or elsewhere. **C** AR NH_2-terminal proline (Pro) and glycine (Gly) repeat nucleotide and amino acid sequence in primates and rat. At the *left* are the amino acid residue numbers for the human (Lubahn et al. 1988) and rat (Tan et al. 1988) AR that begin the repeats. All trinucleotides code for Pro, except CAC for His and ACC for Thr in the rat AR (*left panel*); all sequences code for Gly, except CCA for Pro in the rat AR (*right panel*). Sequences are identical within species unless otherwise indicated.

in the steroid binding domain are the most highly conserved. The only two amino acids unique to the human AR, asparagine 233 and serine 494, replace serine and glycine, respectively, in lower primates and rat (Fig. 1B). Synonymous nucleotide changes in the DNA and steroid binding domains in the macaque and baboon AR are conserved in exons C–G (Fig. 2A), whereas the lemur sequence in this region is intermediate between that of rat and that of higher primates, supporting its phylogeny (Fig. 4). Threonine 656 in the human AR hinge region (exon D) is alanine in baboon and macaque and serine in lemur and rat (Fig. 1B). These transitions have no measurable effect on AR transcriptional activity (Zhou ZX, Kemppainen JA, and Wilson EM, unpublished studies).

Structural Effects of CAG-I Repeat Expansion. To determine the effects of the expanding CAG-I expansion on AR structure, expression vectors coding for the transcriptionally active NH_2-terminal and DNA binding domain fragment AR1-660 (Zhou et al. 1994) were constructed to contain 0 to 64 CAG-I repeat lengths. The 64 CAG-I repeat derives from a patient with spinal/bulbar muscular

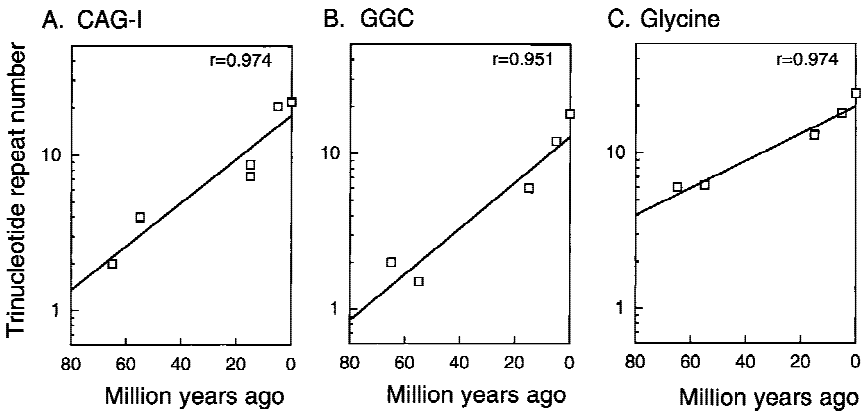


Fig. 3. Graphical representation of AR major repeat expansions during primate divergence. Shown on a semilog scale is the expansion of **(A)** CAG-I coding for glutamine, **(B)** GGC coding for Gly, and **(C)** the full Gly repeat, relative to time of divergence of the major primate species and rat (Tan et al. 1988).

atrophy (La Spada et al. 1992) and shorter repeats mimic those of the nonhuman primates.

Labeling for 0.5 h with [^{35}S]methionine/cysteine results in double-bands for all CAG-I repeat lengths in AR1-660 (Fig. 5, lanes 1–10), which is characteristic of the full-length human AR (Zhou et al. 1995). Spacing between the double-bands is proportional to CAG-I repeat length. Shorter CAG repeats in AR1-660 migrate as closely spaced bands (Fig. 5A, lanes 1 and 3), whereas expanding repeats are more divergent (Fig. 5A, lanes 5 and 7). The largest difference in band migration occurs for CAG 23 (Fig. 5A, lane 7), a predominant AR allele in the human population, and decreases as the CAG-I repeat expands to the pathological range associated with spinal/bulbar muscular atrophy (CAG 64; Fig. 5A, lane 9).

As shown in Fig. 5A (lane 11) and as reported previously (Zhou et al. 1995), the upper band results from phosphorylation of serine 94 and is eliminated by nucleotide conversion coding for alanine. The shift to the slower-migrating, phosphorylated form is also lost in Δ14–83 AR1-660, which deletes the CAG-I repeat region (Fig. 5A, lane 13). The results indicate that AR migration on SDS–acrylamide gels is influenced by CAG repeat length and serine 94 phosphorylation and raises the possibility of a structural interaction that persists in the presence of SDS.

With subsequent 1.7-h isotope dilution at 37°C after the 0.5-h labeling period with [^{35}S]methionine/cysteine, the AR shifts to the slower-migrating, phosphorylated form (Fig. 5A). From the extent of band shift, phosphorylation occurs more rapidly as the CAG-I repeat expands. After 1.7 h, approximately 90% of the lower band shifts to the upper phosphorylated form for CAG 64, whereas CAG 4 from lemur remains predominantly in the lower unphosphorylated form, suggesting that the CAG repeat length alters the rate of serine 94 phosphorylation. This conclusion is supported by the lack of change in migration during isotope dilution of the serine-to-alanine mutation and by deletion of the CAG repeat in Δ14–83.

To establish whether the difference in band migration results from the extent of phosphorylation or depends on glutamine repeat length and serine 94 phosphorylation,

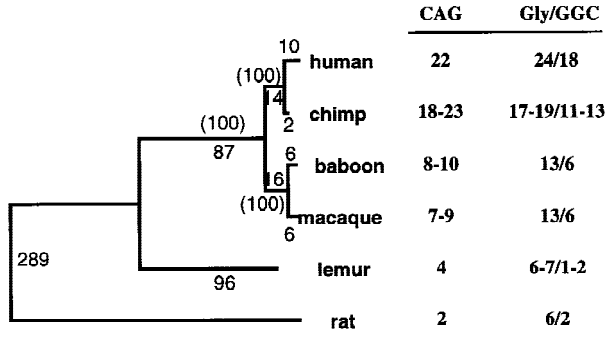


Fig. 4. Phylogeny of the AR DNA sequence. Full coding sequences excluding the trinucleotide repeat regions were compared with maximum parsimony by phylogenetic analysis using parsimony (PAUP) (Swofford, 1993). Shown are CAG-I and Gly/GGC repeat lengths relative to species.

the effect of CAG repeat length was determined on steady-state levels of AR phosphorylation in an NH_2-terminal 171 residue AR fragment. With increasing CAG repeat lengths, steady-state phosphorylation measured by the incorporation of [^{32}P]-orthophosphate is essentially equivalent for CAG 0, 4, 14, and 23 (Fig. 5B, lanes 2, 4, 6, and 8; lower panel). Reduced phosphorylation is noted, however for CAG 64 (Fig. 5B, lane 10). This reduction is not accounted for by reduced expression of CAG 64 (Fig. 5B, upper panel, lane 10), as there was nearly equivalent expression by [^{35}S]methionine/cysteine incorporation (Fig. 5B, upper panel). The DEAE-dextran transfection method used in these studies results in high AR expression levels so differences in AR mRNA and protein expression were not evident with expanding CAG repeat as reported previously (Choong et al. 1996a). In each case phosphorylation is eliminated by the serine 94-to-alanine mutation (Fig. 5B, lower panel, lanes 3, 5, 7, 9, and 11), indicating that phosphorylation occurs at serine 94. Thus although the rate of serine 94 phosphorylation is increased with increasing CAG-I repeat length, steady-state phosphorylation levels decrease as CAG repeat lengths expand to the pathological range associated with spinal bulbar muscular atrophy.

An increasing disparity between calculated and apparent molecular weight with CAG-I repeat expansion oc-

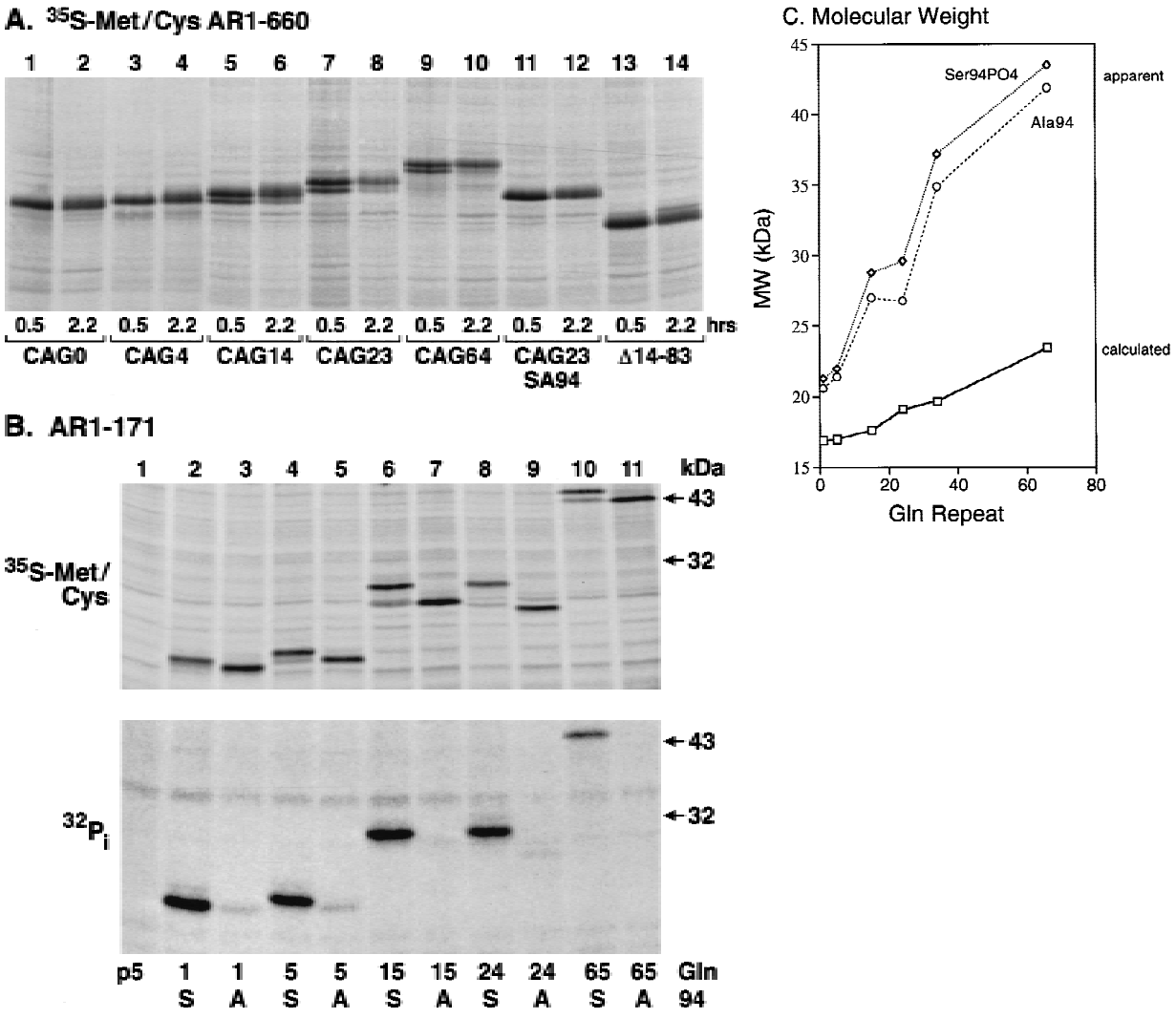


Fig. 5. Structural change in the AR NH_2-terminal domain with CAG repeat expansion and AR phosphorylation. **A** AR1-660 expression vector DNA (10 μg) coding for AR NH_2-terminal and DNA binding domain residues 1–660 and increasing CAG-I repeat lengths from 0 to 64 (lanes 1–10), Ser 94-to-Ala mutation in CAG 23 (lanes 11 and 12), and deletion of residues 14–83 (deletes the CAG-I repeat; lanes 13 and 14) were expressed in monkey kidney COS-1 cells as described. **B** AR1-171 expression vector DNA coding for NH_2-terminal residues 1–171 with increasing CAG-I repeats constructed by truncating pCMVhAR with *AflII* were expressed in COS-1 cells. Transfected cells were incubated with 200 μCi [^{35}S]methionine/cysteine Trans-label (*top*) or 150 μCi [^{32}P]orthophosphate (*bottom*) for 2 h at 37°C as described (Kemppainen et al. 1992) and harvested. Samples were immune precipitated using AR52 antibody and analyzed on 12% acrylamide gels containing SDS. **C** Apparent molecular weight (kDa) of AR NH_2-terminal fragment 1–171 with increasing CAG-I repeat lengths in the Ser 94 phosphorylated (◇) and Ala 94 (○) forms measured in 12% acrylamide gels containing SDS versus the calculated MW (□).

curs with a 171-residue AR NH_2-terminal fragment. The calculated molecular weight of 17 kDa for CAG 0 and 4 in AR1-171 is similar to its migration in SDS–acrylamide gels at 22 kDa (Fig. 5C). The disparity in calculated and apparent molecular weight increases dramatically, however, for CAG 14, 23, and 64 in AR1-171, where the apparent molecular weight for CAG 64 is twice the calculated size (23 versus 43 kDa) (Fig. 5B). This increase in apparent molecular weight of the AR NH_2-terminal fragment is not evident in the 660-amino acid residue NH_2-terminal and DNA binding domain fragment coded by AR1-660, even though the shift of the phosphorylated and nonphosphorylated forms in the 171-residue fragment (Figs. 5B–D) recapitulates that observed for AR1-660 (Fig. 5A). The results suggest that self-aggregation of the 1–171 NH_2-terminal fragment occurs with increasing glutamine repeat length into the pathological range.

Discussion

Several trinucleotide repeat expansions occur in the AR during primate evolution. These include an increase from 4 CAG-I repeats coding for glutamine in lemur to 23 in human and a similar increase from 1 to 18 GGC coding

for glycine in these species. These and other natural mutations are congruent with the known time of diversification among the earliest-evolved primates, the prosimian lemurs 55 million years ago, the branching of Old World monkeys 37 million years ago, and the evolving macaque and baboon 15 million years ago and chimpanzee 5 million years ago (Collura and Stewart 1995). Prosimian lemurs arose during primate evolution from lower mammals about 64 million years ago (Mittermeier et al. 1995) and have been isolated and protected on the Island of Madagascar. The comparison of the lemur AR DNA sequence with those of the rat and the higher primates is in keeping with their transitional status among evolving primates. Expansion of the CAG-I and GGC trinucleotide repeats during primate evolution indicates that strict conservation of these coding regions is not essential for AR function.

Previous studies investigating the effect of CAG repeat length on AR transactivation are controversial. Some observations suggest that increasing CAG repeat length correlates with reduced AR function (Chamberlain et al. 1994; Mhatre et al. 1993), however, we (Choong et al. 1996a) and others (Neuschmid-Kaspar et al. 1996) observed no direct effect of CAG expansion on AR transactivation compared with wild-type AR. Jenster et al. (1994) noted that the increased functional activity of an AR mutant containing a deletion of CAG-I compared to wild-type AR was reporter specific. Our previous studies demonstrating increased AR mRNA and protein expression with decreasing CAG repeat length (Choong et al. 1996a) raised the possibility that higher AR expression in nonhuman primates with shorter CAG-I repeats may confer these species with a reproductive or survival advantage. Correlation between CAG repeat length and increased spacing between the doublet bands has been reported previously by others (Jenster et al. 1994). Results from the present report support these findings and indicate that the double-band migration of the full-length AR in SDS gels depends on both serine 94 phosphorylation and the expanded glutamine repeat.

In some studies of human prostate cancer, shorter CAG-I repeats correlate with more aggressive disease (Giovannucci et al. 1997; Irvine et al. 1995; Hakimi et al. 1997). Association of the AR with other transcription factors such as the protooncogene c-*Jun* (Bubulya et al. 1996), members of the Ets-related family (Bubulya et al. 1996; Schneikert et al. 1996; Zhang et al. 1997), and the NF-K B proteins (Palvimo et al. 1996) suggests a complex mechanism for AR activation of specific genes. Notably the interaction between the AR and the latter two proteins involves the NH_2-terminal region of the AR and results in repression of gene expression. The functional effect of the serine phosphate–glutamine repeat interaction remains to be elucidated. It is possible that altered AR conformation and phosphorylation influence the expression of these and other protooncogenes and contribute to carcinogenesis.

Expansion of CAG-I within the AR gene is thought to be the causative mutation for X-linked spinal and bulbar muscular atrophy. Our results regarding the altered migration of AR NH_2-terminal fragments with varying glutamine repeat lengths are the first to suggest abnormal aggregation of AR fragments containing expanded CAG repeats. Recent reports investigating the pathological mechanism for neurodegeneration in other trigenic diseases associated with CAG expansion support our results and indicate that SDS fails to disrupt the protein conformation engendered by the CAG repeat. Ikeda et al. (1996), using a truncated 123-amino acid Machado–Joseph disease protein containing 79 glutamines, noted higher protein migration than the calculated molecular weight, which persisted following boiling in the presence of SDS. Scherzinger et al. (1997) reported that purified glutathione-*S*-transferase–huntingtin fusion proteins with increasing polyglutamine lengths form high molecular weight insoluble protein aggregates in SDS when the protein contains a polyglutamine expansion in the pathological range between 35 and 40 repeats. Electron micrographs revealed fibrillar structures reminiscent of β-amyloid fibrils in Alzheimer's disease which resemble intranuclear inclusion bodies detected in symptomatic mice transgenic for the Huntington's disease mutation (Davies et al. 1997). Recombinant huntingtin fusion proteins migrated at a size corresponding to the calculated molecular weight in SDS–acrylamide gels, but removal of the glutathione-*S*-transferase tag resulted in two protein products, at 37 and 60 kDa. These findings resemble the results of our experiments using AR1-660 and AR1-171, where the latter fragment migrated with a higher apparent molecular weight as the glutamine repeat expanded, whereas the same phenomenon was not observed with the larger AR fragment. We and others have observed that the SDS gel mass of the full-length AR containing expanded CAG-I repeats is close to the calculated mass (Choong et al. 1996a; Jenster et al. 1994). It is conceivable that protein aggregation occurs after AR proteolysis to smaller fragments encompassing an expanded CAG-I repeat and that this aggregation contributes to loss of neuronal cell function. A pathophysiological mechanism such as this is consistent with the observation that the full-length AR with a CAG-I repeat sequence expanded into the pathological range has functional activity similar to that of the wild-type AR and supports normal male sexual development and function until adulthood.

A recent report by Ordway et al. (1997) of late-onset neurological disease associated with neuronal intranuclear inclusions in transgenic mice expressing a mutant hypoxanthine phosphoribosyltransferase protein containing an expanded CAG repeat of 146 residues en-

coded with exon 3 suggests that the neurotoxicity of the polyglutamine repeat may relate to its presence in smaller or truncated proteins. Together, these results support a common pathogenic mechanism for neurodegeneration due to CAG triplet repeat diseases.

Perutz et al. (1994) suggested that glutamine repeats function as polar zippers that join DNA-bound transcription factors and that disease results from increased nonspecific interactions or from gradual precipitation of affected proteins in neurons. Enhanced protein aggregation with an expanded polyglutamine tract due to transglutaminase cross-linking was also proposed as a mechanism for cellular toxicity (Green, 1993). Accumulation over time of toxic products or sequestration of critical proteins from sites of normal function may account for the relationship between glutamine repeat length and age at onset of the neurodegenerative phenotype.

Short reiterated microsatellite sequences of one to five nucleotides tend to lengthen in humans relative to other primates, suggesting that microsatellite evolution is directional (Rubinsztein et al. 1995a) and may result from a mutational bias toward longer repeats. In nonhuman primates CAG iteration is shorter in the AR and other genes associated with human neurodegenerative diseases (Djian et al. 1996; Rubinsztein et al. 1995b). Although no universal model accounts for expanding trinucleotide repeats during mammalian evolution, the older age at reproduction in human males compared to nonhuman primates or the larger pool of heterozygote alleles associated with a diverse population could contribute to the higher mutation rates (Rubinsztein et al. 1995a). Natural selection tends to eliminate harmful mutations and maintain advantageous ones, yet silent mutations with no adaptive advantage can become fixed in a genome and reflect the time of divergence of a species (Ayala 1995). In light of our results, we propose that evolutionary expansion of the AR CAG-I repeat sequence could become detrimental if expansion continues into the range associated with androgen insensitivity and neurodegeneration. The deleterious effects on reduced AR expression and gain-of-function defects involving peptide aggregation would be expected to impair human fertility and thus should arrest the expansion through natural selection.

Acknowledgments. We are grateful to the following individuals for providing primate blood samples: Michael Adams, Debbie Golden, and Claude L. Hughes, Department of Comparative Medicine, Bowman Gray School of Medicine, Wake Forest University, for macaque; Lorraine Meller, North Carolina Zoological Park, for chimpanzee; Kenneth Glander and Cathy Williams, Duke University Primate Center, for lemur; and David J. Handelsman, Royal Prince Alfred Hospital, University of Sydney, for baboon from the self-sustaining National Health and Medical Research Council National Baboon breeding colony, supported by the Australian National Health and Medical Research Council, Bullens Primate Research Foundation, and Medical Foundation of the University of Sydney. We thank Rytas Vilgalys at Duke University for preparing the phylogenetic tree, K. Michelle Cobb for technical assistance, and John R. Dyer and Frank S. French for reading the manuscript. This work was supported by Grants HD16910 and P30 HD18968 from the National Institute of Child Health and Human Development Center for Population Research, Research Award 95–193 from the Genentech Foundation for Growth and Development, and the Andrew W. Mellon Foundation.

References

Ayala FJ (1995) The myth of Eve: molecular biology and human origins. Science 270:1930–1936

Bubulya A, Wise SC, Shen XQ, Burmeister LA, Shemshedini L (1996) c-Jun can mediate androgen receptor-induced transactivation. J Biol Chem 271:24583–24589

Chamberlain NL, Driver ED, Miesfeld RL (1994) The length and location of CAG trinucleotide repeats in the androgen receptor N-terminal domain affect transactivation function. Nucleic Acids Res 22:3181–3186

Choong CS, Kemppainen JA, Zhou ZX, Wilson EM (1996a) Reduced androgen receptor gene expression with first exon CAG repeat expansion. Mol Endocrinol 10:1527–1535

Choong CS, Sturm MJ, Strophair JA, McCulloch RK, Tilley WD, Leedman PJ, Hurley DM (1996b) Partial androgen insensitivity caused by an androgen receptor mutation at amino acid 907 (Gly → Arg) that results in decreased ligand binding affinity and reduced androgen receptor messenger ribonucleic acid levels. J Clin Endocrinol Metab 81:236–243

Collura RV, Stewart CB (1995) Insertions and duplication of mtDNA in the nuclear genomes of Old World monkeys and hominoids. Nature 378:485–492

Davies SW, Turmaine M, Cozens BA, Difiglia M, Sharp AH, Ross CA, Scherzinger E, Wanker EE, Mangiarini L, Bates GP (1997) Formation of neuronal intranuclear inclusions underlies the neurological dysfunction in mice transgenic for the Hd mutation. Cell 90: 537–548

Djian P, Hancock JM, Chana HS (1996) Codon repeats in genes associated with human diseases: fewer repeats in the genes of nonhuman primates and nucleotide substitutions concentrated at the sites of reiteration. Proc Natl Acad Sci USA 93:417–421

Edwards A, Hammond HA, Jin L, Caskey CT, Chakraborty R (1992) Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics 12:241–253

Gearing KL, Gustafsson JA, Okret S (1993) Heterogeneity in the polyglutamine tract of the glucocorticoid receptor from different rat strains. Nucleic Acids Res 21:2014

Giovannucci E, Stampfer M, Krithivas K, Brown M, Brufsky A, Talcott J, Hennekens C, Kantoff P (1997) The CAG repeat within the androgen receptor and its relationship to prostate cancer. Proc Natl Acad Sci USA 94:3320–3323

Green H (1993) Human genetic diseases due to reiteration: relationship to an evolutionary mechanism. Cell 74:955–956

Hakimi JM, Schoenberg MP, Rondinelli RH, Piantadosi S, Barrack ER (1997) Androgen receptor variants with short glutamine or glycine repeats may identify unique subpopulations of men with prostate cancer. Clin Cancer Res 3:1599–1608

Ikeda H, Yamaguchi M, Sugai S, Aze Y, Narumiya S, Kakizuka A (1996) Expanded polyglutamine in the Machado-Joseph disease protein induces cell death in vitro and in vivo. Nature Genet 13: 196–202

Irvine RA, Yu MC, Ross RK, Coetzee GA (1995) The CAG and GGC microsatellites of the androgen receptor gene are in linkage disequilibrium in men with prostate cancer. Cancer Res 55:1937–1940

Jenster G, de Ruiter PE, van der Korput HA, Kuiper GG, Trapman J, Brinkmann AO (1994) Changes in the abundance of androgen receptor isotypes: effects of ligand treatment, glutamine-stretch variation, and mutation of putative phosphorylation sites. Biochemistry 33:14064–14072

John SWM, Weitzner G, Rozen R (1991) A rapid procedure for extracting genomic DNA from leukocytes. Nucleic Acids Res 19:48

Kemppainen JA, Lane MV, Sar M, Wilson EM (1992) Androgen receptor phosphorylation, turnover, nuclear transport, and transcriptional activation. Specificity for steroids and antihormones. J Biol Chem 267:968–974

Langley E, Zhou ZX, Wilson EM (1995) Evidence for an anti-parallel orientation of the ligand-activated human androgen receptor dimer. J Biol Chem 270:29983–29990

La Spada AR, Wilson EM, Lubahn DB, Harding AE, Fischbeck KH (1991) Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature 352:77–79

La Spada AR, Roling DB, Harding AE, Warner CL, Spiegel R, Hausmanowa-Petrusewicz I, Yee WC, Fischbeck KH (1992) Meiotic stability and genotype-phenotype correlation of the trinucleotide repeat in X-linked spinal and bulbar muscular atrophy. Nature Genet 2:301–304

Lubahn DB, Joseph DR, Sar M, Tan J, Higgs HN, Larson RE, French FS, Wilson EM (1988) The human androgen receptor: complementary deoxyribonucleic acid cloning, sequence analysis and gene expression in prostate. Mol Endocrinol 2:1265–1275

Lubahn DB, Brown TR, Simental JA, Higgs HN, Migeon CJ, Wilson EM, French FS (1989) Sequence of the intron/exon junctions of the coding region of the human androgen receptor gene and identification of a point mutation in a family with complete androgen insensitivity. Proc Natl Acad Sci USA 86:9534–9538

Mhatre AN, Trifiro MA, Kaufman M, Kazemi-Esfarjani P, Figlewicz D, Rouleau G, Pinsky L (1993) Reduced transcriptional regulatory competence of the androgen receptor in X-linked spinal and bulbar muscular atrophy. Nature Genet 5:184–188

Mittermeier RA, Tattersall I, Konstant WR, Meyers DM, Mast RB (1995) Lemurs of Madagascar. Tropical Field Guide Series, Conservation International, Washington, DC

Neuschmid-Kaspar F, Gast A, Peterziel H, Schneikert J, Muigg A, Ransmayr G, Klocker H, Bartsch G, Cato AC (1996) CAG-repeat expansion in androgen receptor in Kennedy's disease is not a loss of function mutation. Mol Cell Endocrinol 117:149–156

Ordway JM, Tallaksen-Greene S, Gutekunst C-A, Bernstein EM, Cearley JA, Weiner HW, Dure LS, Lindsey R, Hersch SM, Jope RS, Albin RL, Detloff PJ (1997) Ectopically expressed CAG repeats cause intranuclear inclusions and a progressive late onset neurological phenotype in the mouse. Cell 91:753–763

Palvimo JJ, Reinikainen P, Ikonen T, Kallio PJ, Moilanen A, Janne OA (1996) Mutual transcriptional interference between RelA and androgen receptor. J Biol Chem 271:24151–24156

Perutz MF, Johnson T, Suzuki M, Finch JT (1994) Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. Proc Natl Acad Sci USA 91:5355–5358

Rubinsztein DC, Amos WJL, Goodburn S, Jain S, Li SH, Margolis RL, Ross CA (1995a) Microsatellite evolution evidence for directionality and variation in rate between species. Nature Genet 10:337–343

Rubinsztein DC, Leggo J, Coetzee GA, Irvine RA, Buckley M, Ferguson-Smith MA (1995b) Sequence variation and size ranges of CAG repeats in the Machado-Joseph disease, spinocerebellar ataxia type 1 and androgen receptor genes. Hum Mol Genet 4:1585–1590

Scherzinger E, Lurz R, Turmaine M, Mangiarini L, Hollenbach B, Hasenbank R, Bates GP, Davies SW, Lehrach H, Wanker EE (1997) Huntingtin-encoded polyglutamine expansions form amyloid-like protein aggregates in vitro and in vivo. Cell 90:549–558

Schneikert J, Peterziel H, Defossez PA, Klocker H, Launoit Y, Cato AC (1996) Androgen receptor-Ets protein interaction is a novel mechanism for steroid hormone-mediated down-modulation of matrix metalloproteinase expression. J Biol Chem 271:23907–23913

Simental JA, Sar M, Lane MV, French FS, Wilson EM (1991) Transcriptional activation and nuclear targeting signals of the human androgen receptor. J Biol Chem 266:510–518

Swofford DL (1993) PAUP: phylogenetic analysis using parsimony. Smithsonian Institute Press, Washington, DC

Tan JA, Joseph DR, Quarmby VE, Lubahn DB, Sar M, French FS, Wilson EM (1988) The rat androgen receptor: primary structure, autoregulation of its messenger ribonucleic acid, and immunocytochemical localization of the receptor protein. Mol Endocrinol 2: 1276–1285

Zhang M, Magit D, Sager R (1997) Expression of maspin in prostate cells is regulated by a positive Ets element and a negative hormonal responsive element site recognized by androgen receptor. Proc Natl Acad Sci USA 94:5673–5678

Zhou ZX, Sar M, French FS, Wilson EM (1993) Molecular biological aspects of the human androgen receptor relating to disease. In: Steroid hormone receptors. Vk Moudgil (ed) Birkhauser, Boston, pp 407–426

Zhou ZX, Sar M, Simental JA, Lane MV, Wilson EM (1994) A ligand-dependent bipartite nuclear targeting signal in the human androgen receptor: requirement for the DNA-binding domain and modulation by NH_2-terminal and carboxyl-terminal sequences. J Biol Chem 269:13115–13123

Zhou ZX, Kemppainen JA, Wilson EM (1995) Identification of three proline-directed phosphorylation sites in the human androgen receptor. Mol Endocrinol 9:605–615