Evolution of the androgen receptor: structure–function implications

Joseph W. Thornton and Darcy B. Kelley*

Summary

To shed light on the nature and evolution of structure–function relations in the androgen receptor (AR), we have undertaken a comparative analysis of all available AR and other steroid receptor sequences. We have identified a group of amino acids that ''diagnose'' the clade of androgen receptors—residues that are strictly conserved among the ARs but not shared with other receptors. We hypothesize that these amino acids, clustered in a few regions of the protein, confer upon the androgen receptor its unique functions, including recognition of specific DNA response elements and affinity for androgens, rather than other steroid hormones. The four domains of the AR display markedly different rates of evolutionary divergence; conserved portions of the sequence, including small stable stretches within otherwise divergent regions, may be essential to receptor function. Current data from experimental, crystallographic, and clinical studies support these hypotheses, which can now be further tested in the laboratory. *BioEssays* **20**:860–869, 1998. © 1998 John Wiley & Sons, Inc.

Introduction: the evolution of structure and function

Nucleotide and amino acid sequence data record the process by which structure–function relations have originated, evolved, and been maintained through millions of years of mutation and selection. They provide a convenient and powerful source of hypotheses about the relation between a protein's structure and its function. Evolution can be viewed as a vast genetic experiment; although uncontrolled, its large-scale trials have generated a prodigious amount of data, which can now be scanned for revealing patterns. In this essay, we show how a comparative evolutionary analysis of protein sequences can uncover traces of the evolutionary process and provide a basis for the formulation of structure–function hypotheses to guide experimental work.

Department of Biological Sciences and Center for Environmental Research and Conservation, Columbia University, New York, New York.
Funding agency: National Institutes of Health; Grant number: NS19949;
Funding agency: National Science Foundation. Grant number: 9870055
*Correspondence to: Darcy B. Kelley, Department of Biological Sciences, 911 Fairchild Center, Mail Code 2432, Columbia University, New York, NY 10027; E-mail: dbk3@columbia.edu

We focus on the androgen receptor (AR) protein, which mediates the response of individual cells to circulating androgenic hormones, including testosterone and dihydrotestosterone (DHT). These hormones direct male-specific aspects of development, physiology, reproduction, and behavior and are associated with a number of important human diseases and conditions. The AR[1,2] is a ligand-activated intracellular transcriptional regulator that belongs to the nuclear receptor superfamily,[3,4] a large group of related proteins believed to have evolved by a process of duplication and divergence from a common ancestral gene.[5,6] Like all nuclear receptors, the AR has a modular structure, consisting of an N-terminal domain (NTD) involved in transcriptional activation, a zinc-finger DNA-binding domain (DBD) that binds to specific genomic response elements of target genes, a flexible hinge region, and a largely helical ligand-binding domain (LBD), which is also involved in receptor dimerization and transcriptional regulation. (Fig. 1). When the receptor binds its ligand, the protein is thought to undergo a conformational change that facilitates the formation of AR homodimers; this complex can then bind to specific palindromic DNA response elements and interact with the basal transcription machinery

BioEssays 20:860–869, © 1998 John Wiley & Sons, Inc.

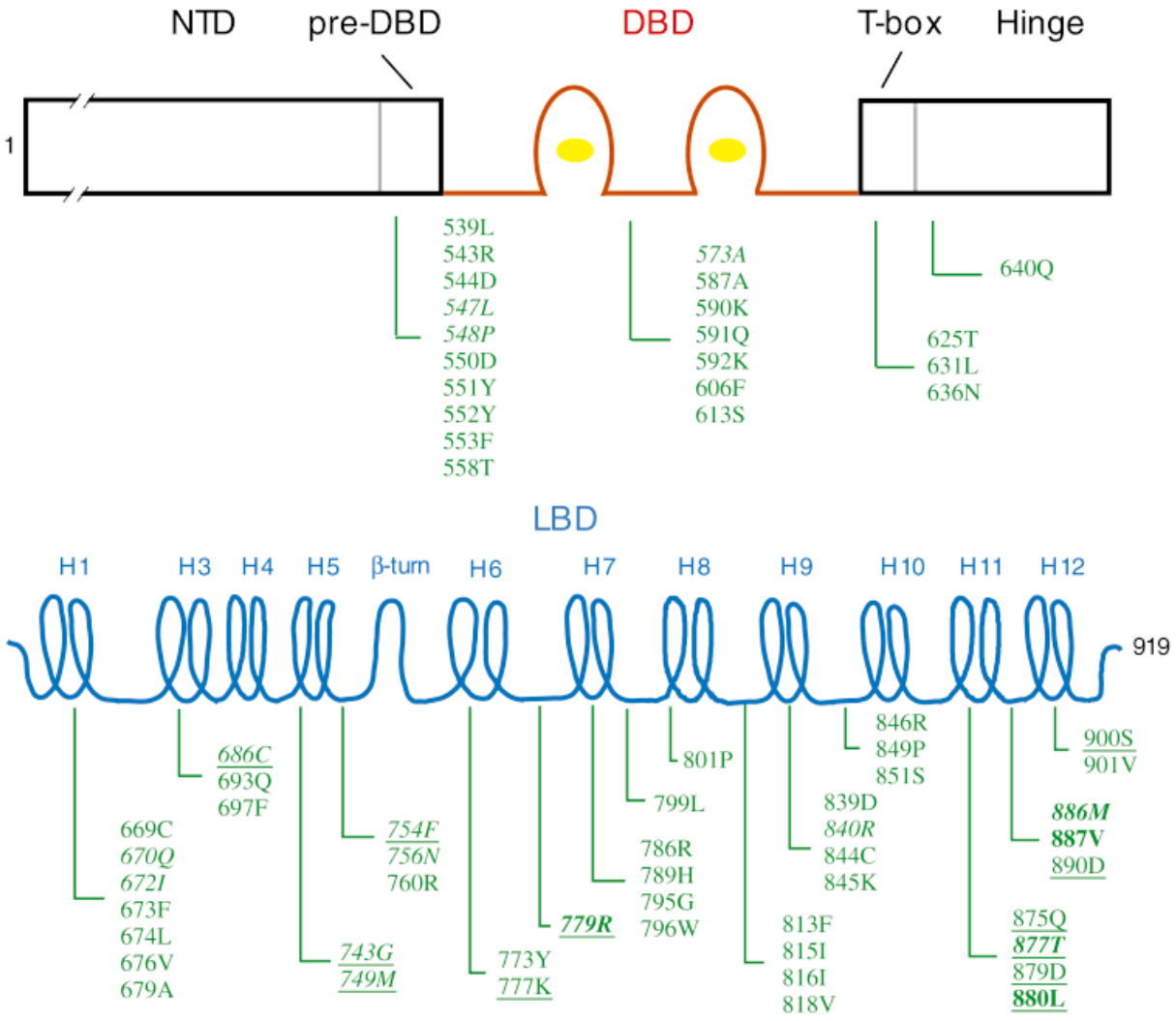

Figure 1. The androgen receptor protein. Amino acids that unambiguously diagnose the clade of androgen receptors are shown in *green*, numbered by position in the human AR. Diagnostic amino acids were identified using MacClade, version 3.01, software and documentation.[42] Only residues that are also conserved within the clade of andrdogen receptors are shown. We hypothesize that these residues are involved in the evolution of AR-specific aspects of receptor function, including ligand-binding and response-element recognition. *Italics*, diagnostic residues that are also mutated in androgen-insensitivity syndrome or prostate cancer.[10,11] Underlined residues are at or no farther than 2 amino acids away from positions predicted to be in contact with ligand, based on crystallographic[13] analysis of other nuclear receptors. **Bold** residues are predicted to be in contact with ligand, based on genetic[25–27] or biochemical[28,29,43] studies of other steroid receptors. The N-terminal domain (NTD), shown in *black*, is involved in transcriptional activation. The zinc fingers of the DNA-binding domain, shown in *red*, mediate binding to specific palindromic DNA response elements and play a role in homodimerization. The hinge domain includes sequences for dimerization and nuclear transport, as well as a short region, called the T-box, that participates in DNA-binding in other nuclear receptors.[19] The C-terminal ligand-binding domain (LBD), shown separately in *blue* with its predicted helical structure,[13] mediates binding of testosterone and dihydrotestosterone and is involved in dimerization and transcriptional regulation.

and numerous cofactors to activate transcription of target genes.

The receptors for androgens, estrogens (ER), progestins (PR), glucocorticoids (GR), and mineralocorticoids (MR), are closely related evolutionarily, and the sequences of their DBDs and LBDs are highly conserved.[3] Despite this sequence similarity, the steroid receptors have taken on considerable functional specificity. For instance, the mammalian AR has a high affinity for testosterone and DHT, but not for the structurally similar steroids that activate other receptors.[1]

Further, the androgen receptor specifically activates androgen-responsive genes, despite binding to a DNA response element also bound by several other steroid receptors.[7] Tissue-specific receptor expression has been proposed as one mechanism for preventing the improper activation of gene targets by other steroid receptors,[7] but the androgen receptor mRNA is expressed in virtually all tissues, including many in which PR, GR, MR, or ER are also expressed.[8,9]

What aspects of the AR sequence determine the specific functions of the androgen receptor, and how did they evolve? Insight into the mechanics of AR function has come from several sources. First, reverse genetic, biochemical, and crystallographic studies of AR and other steroid receptors provide experimental evidence that particular regions of the protein's primary structure are involved in distinct aspects of its function. Second, some naturally occurring mutations in the human AR result in partial or complete insensitivity to androgen (androgen insensitivity syndrome or AIS)[10,11]; individuals with complete AIS develop as phenotypic females despite being genotypically male. Other mutations are associated with reproductive tract cancers, particularly prostatic cancer.[10,11] The set of natural AR mutations with profound phenotypic effects provides further insight into structure–function relations. These data do not indicate, however, what amino acid changes during the course of evolution made possible the emergence of the AR from an ancestral steroid receptor, nor do we understand at the molecular level the selection pressures that have affected the AR in the hundreds of millions of years since its origin. For biomedical purposes, the reductionist goal of understanding which amino acids determine the unique functions of the AR remains distant.

Evolution of the steroid receptors

A comparative analysis of AR sequences from a number of species should illuminate the domain- and site-specific variability and conservation of the androgen receptor sequence and provide a basis for functional and evolutionary inference. Further comparison to other steroid receptors should shed light on aspects of the AR sequence that contribute to functional specificity. Complete AR sequences, however, have been available for only three species (mouse, rat, and human), representing a single vertebrate class. The resulting phylogenetic narrowness—two rodents and a primate, separated from their common ancestors by only an estimated 20 and 80 million years, respectively—has precluded such an analysis; it has not been possible to distinguish conservation of sequence due to selective constraints from conservation due to inadequate time for divergence. Fragmentary sequences are available from a bird, a squamate reptile, and several mammals, but they do not include important domains of the protein (e.g., the NTD and LBD). To remedy this problem, Flavio Kamenetz, Diana Catz, and Leslie Fischer from this laboratory have cloned and sequenced a complete androgen receptor cDNA from the African clawed frog, *Xenopus laevis* (Genbank accession number U67129). Anurans diverged from the lineage that led to reptiles, birds, and mammals in the early Carboniferous, some 350 million years ago[12]; we have thus expanded the temporal range of complete androgen receptor sequences more than fourfold.

The first step in comparative analysis is multiple sequences alignment—the insertion of gaps in one or more sequences to propose homology statements among the amino acids at each position in the various sequences. We aligned the inferred peptide sequences of all available androgen receptors with those of other steroid receptors from rodent, human, and frog.[1] The resulting alignment is virtually identical to the recently published ''canonical'' alignment of nuclear receptor LBDs based on the crystal structure of the human retinoic acid receptor RAR-γ.[13][2]

[1]All available AR amino acid sequences were obtained from the Genbank nucleotide or protein databases using the Entrez browser. Other steroid receptor sequences from *Xenopus*, mouse, human, and rat were obtained in the same way. Abbreviations and accessions for androgen receptor sequences used in this analysis: androgen receptors from frog (*Xenopus laevis*, xenAR, U67129), mouse (*Mus musculus*, musAR, 109558), rat (*Rattus norvegicus,* rat AR, 292896), human (*Homo sapiens*, humAR, 105325), rabbit (*Oryctolagus cuniculus*, rabbitAR, 577829), cow (*Bos taurus*, cowAR, Z75313, Z75314, Z75315), canary (*Serinus canaria*, canaryAR, 414734), and whiptail lizard (*Cnemidophous uniparens*, lizardAR, 1195596). The cowAR sequence, which was available only as nucleotide sequence with some intronic DNA, was edited and translated to yield an inferred partial amino acid sequence. Other steroid sequences used were musGR (90514), humGR (72116), xenGR (2144900), humMR (88157), ratMR (111971), xenMR (994838), humPR (130894), musPR (130895), ratPR (2119671), musER (625327), humER (72114), xenER (625330), humER-β (1518263), musER-β (1912468), humERR1 (36609), musERR1 (1916861), humERR2 (119561), musERR2 (1703648), humSF-1 (2119673), musSF-1 (1805353).

[2]The complete alignment, annotated with available structure–function data, is available on-line at http:www.columbia.edu/cu/biology/faculty/kelley.html. The sequences of steroid and related receptors, with the highly divergent NTD removed (but including 20 residues N-terminal to the first cysteine of the DNA-binding domain) were aligned using clustalX software (46) and guided by a phylogenetic tree of the established relations among taxa[45] and the results of a parsimony-based phylogenetic analysis of nuclear receptor sequences.[15] Minimal manual adjustment of the LBD of the ER group was necessary to match the canonical nuclear receptor alignment proposed based on crystallographic data.[13] The alignment of the the AR NTDs, which are too divergent to be aligned with

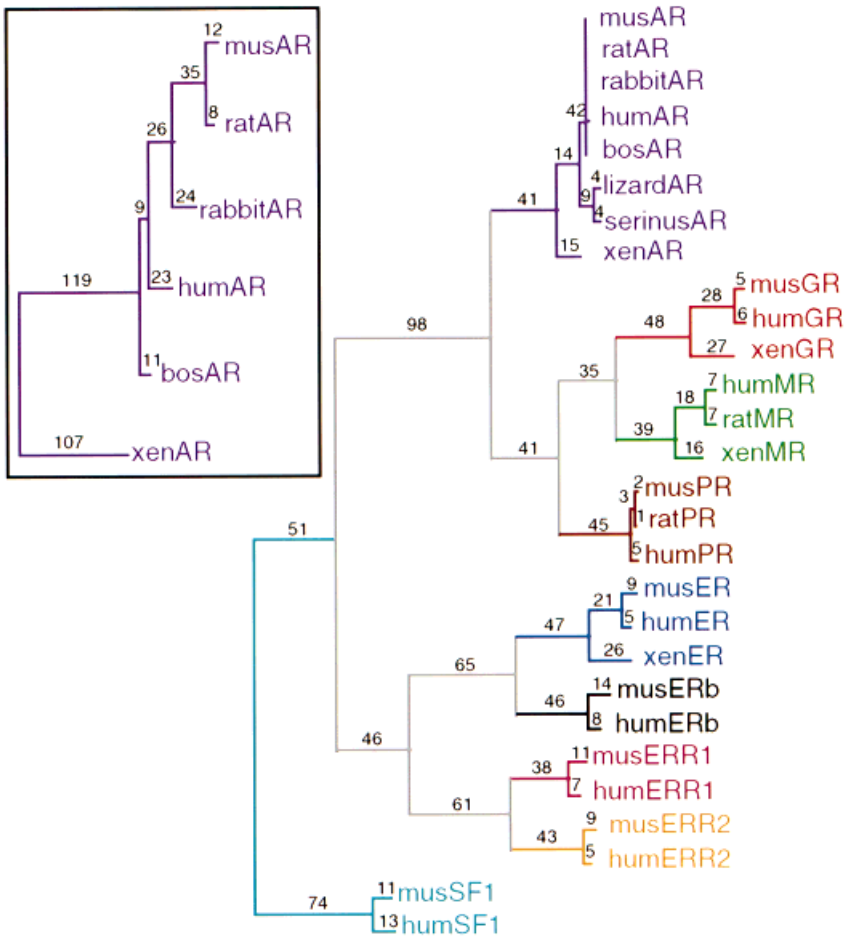

Figure 2. Steroid receptor phylogeny. Aligned sequences (see text discussion under Evolution of the Steroid Receptors) of steroid and related receptors, excluding the NTD, were used as a matrix for phylogenetic analysis based on parsimony, using the branch and bound search option of PAUP software.[44] The analysis yielded three equally most parsimonious trees, which varied only in the resolution of relations within the clade of androgen receptors; these were resolved according to the established taxonomic phylogeny of species from which these sequences were obtained. The tree was rooted on SF-1, as suggested by phylogenetic analyses of a larger set of nuclear receptors.[15] Branch lengths are proportional to the number of amino acid changes in the ligand binding domain occurring on each branch, indicated by numerical branch labels. **Inset:** phylogram of androgen receptor NTDs. Branch lengths are proportional to the number of amino acid changes in the N-terminal domain occurring on each branch. Only sequences for which complete NTD sequences were available are shown.

Comparative analysis is meaningful only within the context of a phylogeny.[14] If the various steroid receptors have evolved, as appears to be the case, by a process of duplication and divergence, the methods used to reconstruct phylogeny among organismal taxa can also be used to infer the relations among paralogous sequences (those resulting from gene duplication) in a gene family. Using the aligned sequences as a data matrix, we sought the most parsimonious phylogenetic reconstruction of relations among the steroid receptors. The tree was rooted on the sequence for the related nuclear receptor steroidogenic factor-1 (SF-1), as suggested by phylogenetic analyses of a larger set of nuclear receptors[15] and by the fact that SF-1, which is common to both protostomes and deuterostomes, is clearly a more ancient protein than any of the steroid receptors, which appear to be restricted to the vertebrate lineage.[16] The resulting phylogeny (Fig. 2) suggests that the estrogen and estrogen-related receptors diverged first from the other steroid receptors, followed, at some later time, by the androgen receptor. The progesterone receptor split off next from the lineage leading to the corticosteroid receptors GR and MR. The mineralocorticoid receptor, which—like its ligand aldosterone but unlike the other steroid receptors and their ligands[16,17]—is not known to be present in fishes or more basal vertebrate lineages, appears to be the most recently evolved of this group of proteins.

the NTD of other receptors, was prepared separately by the same method.

Amino acids that diagnose the androgen receptor

By making only the Darwinian assumption that shared character states (in this case, shared amino acids at homologous positions) are due to descent from a common ancestor, the parsimony-based approach allows the amino acid sequence at any ancestral node in the phylogeny to be hypothetically reconstructed. It is thus possible to analyze in detail the sequence of changes that have taken place during the evolution of a gene family. We have used this approach and

our phylogeny of the steroid receptors (Fig. 2) to identify amino acids that diagnose the clade of androgen receptors—those residues that have changed on the branch of the phylogeny leading from the other steroid receptors to the group of androgen receptors. In order to focus on amino acids that are truly characteristic of the AR, we have ignored residues that are not strictly conserved in all androgen receptors or that contain the same amino acid in other steroid receptors. We include, however, amino acids that the AR sequences share only with SF-1—by unique retention of, or reversal to, the ancestral amino acid—because these also distinguish the AR from all other steroid receptors.

The resulting group of 65 diagnostic amino acids are not randomly distributed but are clustered in specific regions of the AR sequence (Fig. 1). In particular, there are numerous diagnostic residues in the pre-DBD (the 20 residues immediately N-terminal to the DBD), the zinc-finger region itself, the T-box (the 12 residues immediately C-terminal to the DBD), and in portions of the LBD—specifically, in those parts of the sequence corresponding to helices 1, 7, 9, and 11, and loops L1–3, L8–9, L9–10, and L11–12 in the ''canonical'' alignment of nuclear receptor LBDs.[13]

Predicting structure/function relations: DNA binding and recognition

We hypothesize that these diagnostic amino acids determine the unique functions that differentiate the androgen receptor from other steroid receptors, including response-element recognition, dimerization behavior, and ligand binding. How, for example, does each steroid receptor activate a unique set of target genes, since four of the five proteins (AR, PR, GR, and MR) bind to the same 15-nucleotide palindromic response element? The accepted picture of receptor function ascribes DNA recognition to the ''P-box,'' a six-amino acid portion of the DBD known to mediate recognition of response elements.[3] The P-box sequence, however, is identical in all known sequences of the AR, PR, GR, and MR, so this region cannot explain the functional specificity of each receptor.

A clue comes from recent research on the glucocorticoid receptor, which makes clear the importance of ''context effects,'' by which sequences up and downstream from the palindrome are involved in determining the affinity of the receptor for canonical response elements in different genomic contexts.[18] The parts of the steroid receptor involved in these interactions with DNA have not been identified, however. We hypothesize that the large number of diagnostic amino acids clustered in the androgen receptor's pre-DBD and the T-box—11 and 3 residues, respectively—are involved in determining the AR's ability to recognize and bind to unique response elements upstream from androgen-induced genes which, though they share the canonical palindromic repeat with REs for other steroid receptors, may have AR-specific upstream or downstream ''contexts.''

This hypothesis can be evaluated with data from human mutations in the AR and biochemical and crystallographic studies of other nuclear receptor family members. Both the pre-DBD and the T-box are adjacent to the zinc-finger region, which makes direct contact with DNA in all receptors thus far examined. More specifically, the T-box of other nuclear receptors forms a helix that is involved in DNA binding[19–21]; residues of the estrogen receptor including, and N-terminal to, the T-box are required for maximum stability on estrogen-response elements.[22] Less information is available on the function of the pre-DBD, but crystallography shows that at least one residue in this region of human thyroid receptor-β makes indirect contact with DNA.[21] In the orphan receptor ROR-α, the DNA-binding specificity of alternatively spliced isoforms maps to a large region N-terminal to the zinc-finger region.[23] At least two of the residues in the pre-DBD—L547 and P548—must contribute to AR function, because their mutation results in partial or mild androgen insensitivity syndrome.[10,11]

Diagnostic amino acids in the zinc-finger region itself may also contribute to the unique response-element affinity—and possibly the dimerization behavior—of the androgen receptor. Of the seven diagnostic amino acids in the AR DBD, only one is in a position that serves a known function in any nuclear receptor. This residue—A573 in the human AR (throughout this paper, all amino acid numbering refers to position in the 919-amino acid human AR)—lies in the first zinc-finger, which is involved in the formation of the dimer interface in the vitamin D and glucocorticoid receptors.[24] In all other steroid receptors, this position is occupied by a hydrophobic valine residue; the presence of an alanine may alter this dimer interface in a way that facilitates AR homodimerization and prevents significant heterodimerization of AR with PR, GR, MR, or ER. The functional role of the other diagnostic amino acids in the AR DBD remains to be tested in the laboratory.

Affinity and specificity for androgens

In the ligand-binding domain, we hypothesize that amino acids uniquely diagnostic of the androgen receptor are involved in determining AR's higher affinity for testosterone and DHT than for other steroids; they may also be involved in determining the protein's unique dimerization behavior. The existing empirical data are largely consistent with this hypothesis. Of 43 AR-diagnostic residues in the LBD, 12 are at, or no more than two amino acids away from, positions known to be in close contact with ligand in the RAR holoreceptor crystal structure.[13] These residues cluster in predicted helices 5, 11, 12, and the loop between helices 6 and 7 (underlined in Fig. 1). Helix 1 (H1), also rich in diagnostic amino acids, is not predicted to contact the ligand; it may contribute to the overall fold of the domain in a way that facilitates unique binding to

androgens, or the prediction based on homology to RAR-γ may be incorrect.

Of particular interest is the region 866–889, spanning H11 and the loop between H11 and H12, which harbors six diagnostic residues. This region of the AR corresponds to amino acids in the ER shown by site-directed mutagenesis to be involved in determining the affinity of the receptor for its ligand and, in particular, for determining the differential affinity of the protein for various agonists and antagonists.[25,26] Two of the AR-diagnostic residues in this region (T877 and L880) align directly with amino acids in the ER that determine that receptor's affinity for estradiol, while a third (V887) aligns with an amino acid that determines ER's affinity for trans-hydroxytamoxifen and hexestrol.

T877 may be a particularly important amino acid. In addition to its role in ER ligand-binding, mutation of this residue to alanine in the AR of LNCaP prostate cancer cells reduces binding specificity for androgens and causes the receptor to become transcriptionally active upon binding estrogens, progestins, and anti-androgens.[27] Affinity labeling studies suggest that the homologous cysteine residue of the glucocorticoid receptor makes contact with ligand,[28] while crystallography shows that the homologous position in RAR-γ does so, as well.[13] Evolution of a unique threonine at this position in the ancestral AR sequence may have been a key event in the emergence of a receptor that uniquely binds testosterone and its metabolites, but not other steroids.

R779, in the loop between predicted helices 6 and 7, may also play a key role in the specificity of the androgen receptor. This residue occurs at a position predicted to make contact with ligand, based on homology to RAR-γ. Further, affinity labeling shows that the corresponding cysteine in the rat GR contacts synthetic glucocorticoids,[28] and mutation of this residue to glycine or serine produces a "super-GR" that has increased affinity for glucocorticoids but reduced affinity for other steroids.[29] Finally, mutation of this amino acid in the human AR is associated with complete androgen insensitivity syndrome.[10,11] During the emergence of the ancestral androgen receptor, replacement of a hydrophobic residue with a basic arginine at this position may have been essential to the evolution of a ligand-binding domain with high specificity for testosterone and DHT, but not other steroids.

Human patients with androgen insensitivity syndrome provide additional data to test our hypothesis. Indeed, eight of the amino acids diagnostic of the androgen receptor LBD are associated with AIS in human patients.[10,11] Several such residues (italicized in Fig. 1) are of particular interest. Mutation G743V is associated with complete AIS and a significant reduction in ligand-binding affinity; the glycine here, adjacent to a position predicted to be in contact with ligand,[13] may thus be particularly important for androgen recognition. Mutation F754V is associated with complete androgen insensitivity and zero affinity for ligand, while R840C/H is associated with partial AIS and reduced ligand affinity. Other diagnostic amino acids that produce human AIS when mutated, but for which no effect on ligand binding has been established, include C686, M749, and N756.

Divergence and conservation in the N-terminal domain

The NTD of the androgen receptor is so divergent that it could not be aligned with other steroid receptors, so it was not possible to specify diagnostic amino acids in this domain. NTDs from androgen receptors of various species can be aligned to each other, however, and similarly "orthologous" alignments can be prepared for each of the other steroid receptors, as well. These alignments allowed us to compare the rate of sequence divergence in the AR vis-à-vis other steroid receptors and identify conserved regions, which are presumably of functional importance, within this generally divergent domain. To evaluate the rate at which receptor sequences have evolved—and in turn, to gain insight into the strength of selection that has constrained sequence change at a fine scale—we have evaluated the degree of divergence among the receptors using a random model of molecular evolution,[30,31] which allows a statistical approach to sequence analysis. For each pair of receptor sequences, we first evaluate the simple phenetic distance, measured as the proportion of pairwise amino acid differences (D); for instance, two peptides 10 amino acids long, identical at eight sites, would give $D = 0.20$. Because multiple hits at the same sites can never be observed as more than a single difference, however, observed differences systematically underestimate the actual rate at which amino acids have been replaced. Thus, we have calculated the corrected proportion of amino acid replacements (K), using a simple probabilistic model.[30] This approach also allows us to calculate the variance of K, which increases with shorter sequences or higher degrees of divergence; we can thus statistically evaluate hypotheses that regions of a protein—or the same protein in different lineages—may have diverged at different rates. For instance, it is clear that the degree of divergence (and, by implication, the strength of selection that causes sequence conservation) of the androgen receptor varies wildly among the domains, ranging from almost complete conservation in the DBD to nearly total divergence in the NTD; these differences are far greater than expected by chance alone (Fig. 3).

The NTD of the *Xenopus* AR is highly divergent from the other AR sequences (corrected pairwise divergence between musAR and xenAR, $K = 0.88 \pm 0.060$, $D = 0.58$). Comparison of the rate of divergence of the NTD to other pairs of mouse–*Xenopus* steroid receptors reveals that the NTD of the AR is significantly more variable than the same region of other steroid receptors—more than twice as diverged, in some cases (Fig. 3). The NTD of the human AR is polymorphic, with polyglutamine and polyproline tracts of variable

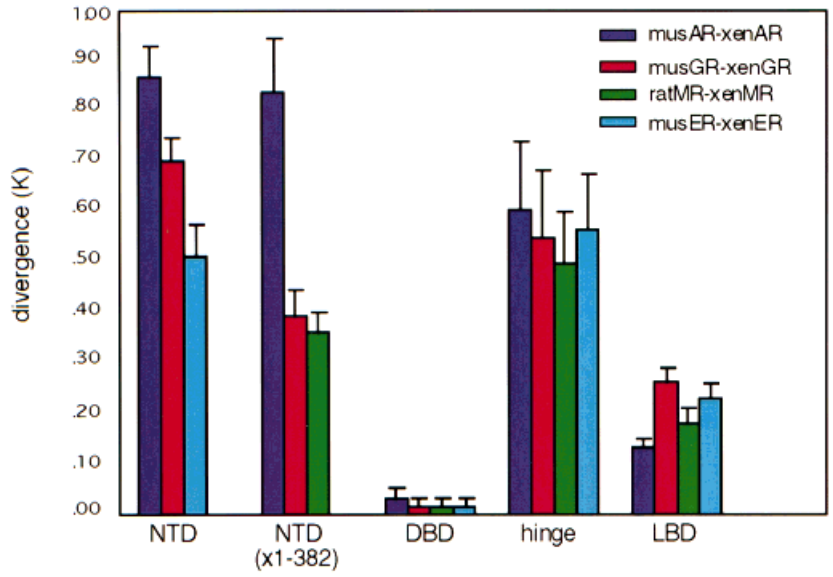

Figure 3. Evolutionary divergence of steroid receptor domains. The proportion of corrected pairwise amino acid differences (K) between the *Xenopus laevis* and rodent androgen receptors were calculated for each domain of the protein, and for the glucocorticoid, mineralocorticoid, and estrogen receptors. Only a C-terminal portion of the NTD of the xenMR is available, but a motif present in this fragment is conserved among AR, MR, and GR, allowing K to be calculated for the portion of the NTD C-terminal to this motif (NTDx1–382). We calculated K from the proportion of observed amino acid differences (D) between each pair of aligned sequences (gaps excluded), based on the equation $K = -\ln(1-D)$.[30] Error bars show the standard deviation of $K = [D/[(1-D)L]^{1/2}$, where L is the length of the pair of sequences.[31] Comparisons are based on separate alignments for each type of receptor, prepared using ClustalX software.

length; these repeats are absent in the *X. laevis* AR, suggesting that these tracts were not present in the ancestral androgen receptor but have evolved since the lineage leading to mammals split from the ancestral anurans.

The great divergence of the AR NTD—particularly when compared to the extraordinary conservation of the rest of the protein—might suggest that the hypervariability of this region was generated not by sequence divergence but by exon shuffling or some other form of sequence transfer, such that novel sequences would have been grafted onto the N-terminus of an otherwise conserved receptor cassette. Our analysis, however, reveals five small conserved regions throughout the NTD, as detailed below. This pattern clearly indicates that the entire *Xenopus* and mammalian ARs evolved from a single full-length ancestral sequence, N-terminal domain included. The hypervariability of the NTD is not due to exon shuffling but must be the result of a significantly faster rate of sequence divergence than has occurred in the other domains of the AR.

The great divergence of the AR NTD is particularly remarkable, considering that this region of the protein is essential to full transactivation potential.[32] Proper receptor function—particularly the transactivation activity, dubbed AF-1, which has been localized to this domain[33]—may thus require only several conserved regions within the NTD, with the rest of the domain necessary for proper folding but with much less stringent sequence requirements. Local regions within the domain that are quite conserved between *Xenopus* and human suggest strong selective constraints only in these stretches, which we hypothesize are essential to the function of the NTD. Amino acids corresponding to positions 1–35 of the human AR are far less variant ($K=0.377\pm0.114$) than the rest of the NTD, as are amino acids 230–268 ($K=0.134\pm0.060$). The latter, serine-rich sequence appears similar to a region of the ER that contains a number of serine residues that serve as phosphorylation sites necessary for AF-1 activity and ligand-independent activation of transcription.[34] Like ER, AR is a phosphoprotein that can be activated by the mitogen-induced kinase cascade, becoming competent to activate transcription even in the absence of ligand.[35] This short region may include phosphorylation sites for these kinases. Several shorter motifs appear conserved between the frog and mammalian ARs, but the statistical significance of the difference in divergence compared to the rest of the N-terminal domain is less clear, precisely because the sequences are short. A total of 7 of the 11 residues in the motif 382–393 and 8 of 12 amino acids 433–446 are conserved between the anuran and mammalian ARs. It is notable that a mutation of P390 in the former motif is associated with complete androgen insensitivity syndrome,[11] supporting the hypothesis that this region is essential to proper AF-1 function or regulation.

Finally, the pre-DBD is highly conserved among the androgen receptors, with only one of 16 sites variant in this region ($K=0.065\pm0.065$). Though less strictly invariant, an even larger region comprising residues 510–558 is also conserved ($K=0.336\pm0.090$). The extremely high degree of conservation of this region supports the hypothesis that it contributes to an essential and unique function in the AR, such as recognition of androgen response elements.

The DNA-binding domain and hinge

Like the DBD of all steroid receptors, the zinc-finger region of the *Xenopus* AR is remarkably conserved relative to its mouse orthologue ($K=0.03\pm0.02$, $D=0.03$). There are only two amino acid substitutions in the DBD of the *Xenopus* androgen receptor, both threonine-for-serine replacements.

The N-terminal portion of the hinge is highly conserved in the AR. Of the 18 amino acids immediately C-terminal to the DBD, only one is different between *Xenopus* and mouse. The first 12 amino acids—positions homologous to the T-box—are invariant among all androgen receptors, with the exception of a single substitution in the cow AR. The high degree of conservation C-terminal to the T-box as traditionally defined, however, suggests that an even larger region of the hinge is essential to AR's unique functions, possibly related to DNA-binding, as discussed above, or to the unique nuclear localization behavior of the AR,[1] a function in which the hinge is known to participate.[36]

The rest of the hinge region of the androgen receptor is extremely variable. The proportion of amino acid replacements between the *Xenopus* and mouse AR hinge ($K=0.60\pm0.147$, $D=0.45$) is somewhat higher than for the hinge of other steroid receptors, but the difference is not statistically significant, due to the relatively short length of the hinge and the high degree of divergence. Although portions of the hinge have been found to be involved in transcriptional repression by TR-β[37] and in the overall orientation of the DNA-bound receptor in ROR-α,[38] the specific function of the hinge remains undetermined. Our analysis does not suggest any function strictly dependent on primary structure.

Unusual conservation of the ligand-binding domain

The ligand-binding domain of the *Xenopus* androgen receptor is highly conserved in relation to other androgen receptors (musAR-xenAR, $K=0.13\pm0.023$, $D=0.12$). The LBD of the androgen receptor has diverged significantly less during evolution than the same region of other steroid receptors; the divergence between the androgen receptor LBDs of frog and mouse is less than one-half that between the estrogen receptor LBDs from the same two species (Fig. 3). The phylogenetic reconstruction based on parsimony (Fig. 2) supports this inference, since the number of amino acid changes on branches leading to the various androgen receptors is much lower than on the corresponding branches leading to other steroid receptors in the same species. That the AR LBD has evolved so much more slowly than other steroid receptors suggests that the amino acid requirements for binding androgens are extraordinarily strict, more so than required for other receptors to bind their ligands. The great number of natural and synthetic chemicals that are estrogen receptor agonists[39]—in contrast to the much smaller number of compounds that interact with the androgen receptor—is consistent with this view.

We have suggested that amino acids conserved among androgen receptor orthologues separated by hundreds of million years have been strongly constrained by selection—an argument that implies that the vast majority of the residues in the AR LBD are functionally important. To refute this hypothesis, we sought sites that appear to be functionally important because they are mutated in individuals with AIS or prostate cancer, [10,11] but that are not conserved between the mammalian and anuran AR proteins. Of 81 AIS sites in the LBD, 74 are conserved between the human and *Xenopus* receptors. Of the seven positions that are not identical, four are conservative hydrophobic substitutions (V746I, M787L, I841V, and L881F). Of the three nonconservative substitutions, two (G683V and A748P) involve replacements in the *Xenopus* lineage with residues different from those known to produce deficient phenotypes. At the remaining position, A870, a glycine is found in the AR of the frog and of a patient with complete AIS. Why the same mutation does not result in androgen insensitivity in the wild-type *Xenopus laevis* is unknown; androgen target tissues in *X. laevis* bind DHT and R1881 with high affinity.[40,41] (The 23 AIS sites not in the LBD are also conserved in the frog receptor, with the exception of two conservative substitutions—one in the DBD (S597T) and one in the hinge (I664L)—and a nonconservative substitution in the hinge (A645G), in which the amino acid in the *Xenopus* receptor differs from the aspartate that produces partial AIS). Of 23 sites mutated in the androgen receptor of prostate cancer cells,[10] only one is variable in the *Xenopus* receptor, a substitution of serine for glutamine at the C-terminus of the protein (Q919); prostate cancer, in contrast, is associated with replacement by arginine.

With the exception of a single amino acid at position 870, then, empirical data are consistent with the hypothesis that conserved amino acids play important roles in the function of the protein. It appears that the selective constraints suggested by mutant phenotypes in contemporary humans have operated over hundreds of millions of years of evolution to conserve the AR sequence at these sites. The lack of strict conservation at a few AIS or prostate cancer positions in the *Xenopus* receptor may have several explanations. In some cases, the amino acid in the frog AR is different from that which produces the AIS phenotype. In others, amino acid replacements may produce functional changes at the molecular level, but these may have been successfully integrated into the physiology of the organism. For instance, the affinity of laryngeal tissue from male *X. laevis* for testosterone is much lower than that of androgen-responsive mammalian tissues, but DHT, which the larynx binds with very high affinity, appears to be the physiologically important hormone in the frog[40,41]; amino acid substitutions that reduce affinity for

testosterone may thus interfere with male sexual differentiation in humans but not in *Xenopus*. Finally, a lack of sequence conservation may be due to different contexts in which natural selection operates in the two species. Risk of prostate cancer, for instance, may not be an important selective pressure in wild *Xenopus laevis*.

Rate of AR amino acid evolution

Assuming that the mammalian–amphibian divergence took place approximately 350 million years ago, we can use the sequence divergence (K) between the *Xenopus* and mouse sequences to estimate the average (but not necessarily constant) rate of AR evolution and compare it to that of other proteins. The entire AR has evolved at an average rate of 1.22 substitutions per site per billion years, somewhat greater than that of many other highly conserved proteins and peptide hormones (Table 1). Most of this difference, however, is due to a very high rate of divergence in the NTD. When considered separately, the DBD of the AR protein is extraordinarily conserved, with an evolutionary rate estimated at 0.04 substitutions per site per billion years, a figure that places the DBD among the most slowly evolving proteins known, such as the ribosomal proteins and myosin-β. The LBD also appears to evolve at the unusually slow rate of 0.19 substitutions per site per billion years, as expected of a receptor that serves functions essential to reproduction and binds a ligand strictly conserved over hundreds of millions of years of evolution.

TABLE 1. Rate of Evolution of the Androgen Receptor*

Protein	Replacements per amino acid position per billion years: Rate	SD
AR (entire)	0.69	0.04
AR NTD	1.25	0.08
AR DBD	0.04	0.03
AR hinge	0.86	0.20
AR LBD	0.19	0.04
Histone 4	0.00	0.00
Somatostatin 28	0.00	0.00
Actin-α	0.01	0.01
Ribosomal S14	0.02	0.02
Ribosomal S17	0.06	0.04
Aldolase A	0.09	0.03
Myosin-β heavy chain	0.10	0.01
Creatine kinase M	0.15	0.03
Lactate dehydrogenase A	0.19	0.04
Insulin	0.20	0.10
Thymidine kinase	0.43	0.08
α-Globin	0.56	0.11
ILGF-2	0.57	0.11
Amylase	0.63	0.06
Erythropoietin	0.77	0.12
Parathyroid hormone	1.00	0.20
Luteinizing hormone	1.05	0.17
Growth hormone	1.34	0.17
Interferon-α1	1.47	0.19
Relaxin	2.59	0.51

*The average rate of amino acid sequence divergence for each domain of the androgen receptor was estimated by dividing the proportion of corrected pairwise amino acid differences (K) between the *Xenopus laevis* and mouse androgen receptors by the estimated evolutionary time since divergence, assuming that the anuran and rodent lineages split 350 million years ago. Estimated evolutionary rates for other selected proteins[30] are shown for comparison.

Hypothesis or hypotheses?

The identification of individual amino acids and conserved regions of the AR can serve as a focal point for future investigations into the relation of androgen receptor structure and function, which is of both fundamental and biomedical importance. We have hypothesized that many of these diagnostic amino acids confer upon the AR the functional specificity that differentiates it from other steroid receptors. It is likely that some of these amino acids will turn out to have been conserved not by natural selection but simply by chance; however, we expect that the number of diagnostic amino acids that are uninvolved in determining AR-specific function should be small, given the great age of the steroid receptors.

We should be precise, then, that we are not formulating a single hypothesis that all diagnostic amino acids and conserved regions of the protein serve differentiating functions in the androgen receptor. This hypothesis would be falsified by a single diagnostic amino acid that was apparently neutral in function. Rather, our analysis proposes a large number of specific hypotheses, each concerning the role of individual residues (or stretches of multiple residues), which can be individually tested in the laboratory. The specificity and refutability of the hypotheses generated by our analysis is, indeed, one of the great strengths of this character-based approach.

A rigorous comparative approach to sequence data can thus make an important contribution to the hypothesis-testing framework for experimental molecular biology. On the basis of our analysis, we have also ventured hypotheses about the process and dynamics of evolution, such as the rate of sequence divergence and its causes, both proximate (molecular mechanisms) and ultimate (the strength and nature of selection). These insights are less amenable to experimental testing because of evolution's historical and contingent nature. Eager as we are to contribute to biology's Popperian approach to the elucidation of molecular mechanisms, we do not discount the value of less testable hypotheses about

evolutionary history. The comparative approach can thus yield strong hypotheses about the way proteins work and contribute to our still murky understanding of how, over hundreds of millions of years of evolution, they learned to do so.

Acknowledgments

We thank Rob DeSalle, Flavio Kamenetz, Larry Chasin, Ben Evans, Robert Pollack, Diane Robins, and Vincent Laudet for helpful comments on the manuscript. J.T. is supported by an NSF Graduate Research Fellowship.

References

1 Zhou Z-X, Wong C-I, Sar M, Wilson EM (1994) The androgen receptor: An overview. *Recent Prog Horm Res* 49:249–274.
2 Keller ET, Ershler WB, Chang C (1996) The androgen receptor: A mediator of diverse responses. *Front Biosci* 1:59–71.
3 Gronemeyer H, Laudet V (1995) Transcription factors 3: Nuclear receptors. *Prot Profiles* 2:1173–1308.
4 Manglesdorf DJ, Thummel CS, Beato M, Herrlich P, Schutz,G, Umesono K, Blumberg B, Kastner P, Mark M, Chambon P, Evans RM (1995) The nuclear receptor superfamily: The second decade. *Cell* 83:835–839.
5 Amero SA, Kretzinger RH, Moncrief ND, Yamamoto KR, Pearson WR (1992) The origin of nuclear receptor proteins: A single precursor distinct from other transcription factors. *Mol Endocrinol* 6:3–7.
6 Laudet V, Hanni C, Coll J, Catzeflis F, Stehelin D (1992) Evolution of the nuclear receptor gene superfamily. *EMBO J* 11:1003–1013.
7 Gronemeyer H (1992) Control of transcription activation by steroid hormone receptors. *FASEB J* 6:2524–2529.
8 Takeda H, Chodak G, Mutchnik S, Nakamoto T, Chang C (1990) Immunohistochemical localization of androgen receptors with mono and polyclonal antibodies to androgen receptors. *J Endocrinol* 126:17–25.
9 Young WJ, Roecker EB, Weindruch R, Chang C (1991) Quantitation of androgen receptor mRNA by competitive reverse transcription polymerase chain reaction. *Endocr J* 2:321–329.
10 Quigley CA, De Bellis A, Marschke KB, el-Awady MK, Wilson EM, French FS (1995) Androgen receptor defects: Historical, clinical, and molecular perspectives. *Endocr Rev* 16:271–321.
11 Gottlieb B, Trifiro M, Lumbroso R, Pinsky L (1997) The androgen receptor gene mutations database. *Nucleic Acids Res* 25:158–162 (Database: www.mcgill.ca/androgendb).
12 Carroll RL (1988) *Vertebrate Paleontology and Evolution.* New York: Freeman.
13 Wurtz J-M, Bourguet W, Renaud J-P, Vivat V, Chambon P, Moras D, Gronemeyer H (1996) A canonical structure for the ligand-binding domain of nuclear receptors. *Nature Struct Biol* 4:87–94.
14 Harvey PH, Pagel MD (1991) *The Comparative Method in Evolutionary Biology*. Oxford: Oxford University Press.
15 Thornton J, DeSalle R (1997) Molecular systematics of the nuclear receptor superfamily. In *Structure and Function of the Nuclear Receptors*. Proceedings of the EMBO Workshop, Erice, Italy, p 71 (abst).
16 Escriva H, Safi R, Hanni C, Langlois MC, Saumitou-Laprade P, Stehelin D, Capron A, Pierce R, Laudet V (1997) Ligand binding was acquired during evolution of nuclear receptors. *Proc Natl Acad Sci USA* 94:6803–6808.
17 Norris DO (1980) *Vertebrate Endocrinology*. New York: Lea & Febiger.
18 LaBaer J, Yamamoto KR (1994) Analysis of the DNA-binding affinity, sequence specificity and context dependence of the glucocorticoid receptor zinc finger region. *J Mol Biol* 239:668–668.
19 Wilson TE, Pauls RE, Padgett KA, Milbrandt J (1992) Participation of non-zinc finger residues in DNA binding by two nuclear orphan receptors. *Science* 256:107–110.
20 Lee MS, Kliewer SA, Provencal J, Wright PE, Evans RM (1993) Structure of the Retinoid X Receptor alpha DNA binding domain: A helix required for homodimeric DNA binding. *Science* 260:1117–1121.
21 Rastinejad F, Perlmann T, Evans RM, Sigler PB (1995) Structural determinants of nuclear receptor assembly on DNA direct repeats. *Nature* 375:203–211.
22 Mader S, Chambon P, White JH (1993) Defining a minimal estrogen receptor DNA binding domain. *Nucleic Acids Res* 21:1125–1132.
23 Giguere V, Tini M, Flock G, Ong E, Evans RM, Otulakowski G (1994) Isoform-specific amino-terminal domains dictate DNA-binding properties of RORα, a novel family of orphan hormone nuclear receptors. *Genes Dev* 8:538–553.
24 Towers TL, Luisis BF, Asianov A, Freedman LP (1993) DNA target selectivity by the vitamin D3 receptor: Mechanism of dimer binding to an asymmetric repeat element. *Proc Natl Acad Sci USA* 90:6310–6314.
25 Ekenat K, Weis KE, Katzenellenbogen JA, Katzenellenbogen BS (1996) Identification of amino acids in the hormone binding domain of the human estrogen receptor important in estrogen binding. *J Biol Chem* 271: 20053–20059.
26 Ekenat K, Weis KE, Katzenellenbogen JA, Katzenellenbogen BS (1997) Different residues of the human estrogen receptor are involved in the recognition of structurally diverse estrogens and antiestrogens. *J Biol Chem* 272:5069–5075.
27 Veldscholte J, Ris-Stalpers C, Kuiper GGJM, Jenster G, Barrevoets C, Claassen E, van Rooij HCJ, Trapman J, Brinkmann AO, Mulder E (1990) A mutation in the ligand binding domain of the androgen receptor of human LNCaP cells affects steroid binding characteristics and response to anti-androgens. *Biochem Biophys Res Commun* 173:534–540.
28 Carlstedt-Duke J, Stromstedt P-E, Persson B, Cederlund E, Gustaffson JA, Jornvall HJ (1988) Identification of hormone-interacting amino acid residues within the steroid binding domain of the glucocorticoid and other steroid hormone receptors. *J Biol Chem* 263:6842–6848.
29 Chakraborti PK, Garabedian MJ, Yamamoto KR, Simons SS (1991) Creation of super glucocorticoid receptors by point mutations in the steroid binding domain. *J Biol Chem* 256:22075–22078.
30 Li W-H (1997) *Molecular Evolution.* Sunderland, MA: Sinauer.
31 Hartl D (1991) *Primer of Population Genetics.* Sunderland, MA: Sinauer.
32 Doesburg P, Kuil C, Berrevoets CA, Steketee K, Faber PW, Mulder E, Brinkmann AO, Trapman J (1997) Functional in vivo interaction between the amino-terminal transactivation domain and the ligand binding domain of the androgen receptor. *Biochemistry* 36:1052–1064.
33 Chamberlain NL, Whitacre DC, Miesfeld RL (1996) Delineation of two distinct type 1 activation functions in the andrdogen receptor amino-terminal domain. *J Biol Chem* 271:26772–26778.
34 Ali S, Metzger D, Bornert J-M, Chambon P (1993) Modulation of transcriptional activation by ligand-dependent phosphorylation of the human oestrogen receptor A/B region. *EMBO J* 12:1153–1160.
35 Reinikainen P, Palvimo JJ, Janne OA (1996) Effects of mitogens on androgen receptor-mediated transactivation. *Endocrinology* 137:4351–4357.
36 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.
37 Nawaz Z, Tsai M-J, O'Malley BW (1995) Specific mutations in the ligand binding domain selectively abolish the silencing functions of human thyroid hormone receptor beta. *Proc Natl Acad Sci USA* 92:11691–11695.
38 McBroom LD, Flock G, Giguere V (1995) The nonconserved hinge region and distinct amino-terminal domains of the RORα orphan nuclear receptor isoforms are required for proper DNA bending and RORa–DNA interactions. *Mol Cell Biol* 15:796–808.
39 Katzenellenbogen JA (1995) The structural pervasiveness of estrogen activity. *Environ Health Perspect* 103:99–101.
40 Segil N, Silverman L, Kelley DB (1987) Androgen-binding levels in a sexually dimorphic muscle of *Xenopus laevis*. *Gen Comp Endocrinol* 66:95–101.
41 Kelley DB, Sassoon D, Segil N, Scudder M (1989) Development and hormone regulation of androgen receptor levels in the sexually dimorphic larynx of *Xenopus laevis*. *Dev Biol* 131:111–118.
42 Maddison WP, Maddison DR (1992) MacClade, version 3.01, software and documentation. Sunderland, MA: Sinauer.
43 Stromstedt P-E, Berkenstam A, Jornvall H, Gustafsson J-A, Carlstedt-Duke J (1990) Radiosequence analysis of the human progestin receptor charged with [3H]promegestone. *J Biol Chem* 265:12973–12977.
44 Swofford D (1993) *Phylogenetic Analysis Using Parsimony (PAUP)*. Version 3.1.1, software and documentation. Sunderland, MA: Sinauer.
45 Novacek MJ (1992) Mammalian phylogeny: Shaking the tree. *Nature* 356:121–125.
46 Thompson JD, Higgins DG, Gibson TJ (1997) Clust al X 1-51 software for power PC. Heidelberg, European Molecular Biology Organization.