Molecular determinants of mammalian sex

Milton H. Werner, Jeffrey R. Huth, Angela M. Gronenborn and G. Marius Clore

Mammalian male sex determination is controlled by a complex hierarchy of gene regulatory proteins and hormones, which promote male gonadal development and regression of the female primordia. At the core of this pathway lies the SRY protein, the master developmental switch for testicular differentiation and hence, the male sex. The three-dimensional structure of the SRY–DNA complex suggests a model of developmental gene regulation in which proteins that alter DNA structure and promote the assembly of higher-order nucleoprotein complexes play an essential role in the timing of cell specialization events.

IN MAMMALIAN SEXUAL differentiation, male and female developmental pathways are controlled by the presence or absence of hormones, which are produced by the male and female gonads in the developing fetus. Before hormonal secretion, XX and XY embryos develop two pairs of genital ducts associated with the undifferentiated gonads. The Müllerian ducts have the potential to differentiate into the fallopian tubes, uterus, cervix and upper vagina of the female reproductive tract. The Wolffian ducts become the vas deferens, epididymis and seminal vesicles of the male reproductive tract. Consequently, for normal male or female development to occur, one genital duct system must differentiate while the other must regress (Fig. 1).

Regression of the female pathway is achieved by the production of Müllerian inhibitory substance (MIS) (also known as the anti-Müllerian hormone) in the fetal testicular Sertoli cells. Subsequently, testosterone, produced by the Leydig cells of the testes, induces differentiation of the Wolffian ducts into male reproductive organs. In the absence of testes, and therefore in the absence of both MIS and testosterone, the Wolffian ducts passively regress, creating a permissive environment for the differentiation of the Müllerian ducts, and thereby, female reproductive organs.

Thus, MIS and testosterone impose a male pattern of development upon an inherently female program. As the male phenotype is controlled by testicular hormones, testicular differentiation is the first identifiable step in the male sex determination pathway. Central to the male pathway is the testis determining factor SRY, synthesis of which leads directly to the development of the testes, and thus, to male sex.

M. H. Werner, J. R. Huth, A. M. Gronenborn and **G. M. Clore** are at the Laboratory of Chemical Physics, Building 5, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892-0520, USA.

Regulated expression in the undifferentiated gonad

Three gene products appear to play regulatory roles in the development of the undifferentiated gonad, namely the Wilms' tumor suppressor protein WT1 (Ref. 1), the orphan nuclear receptor protein SF-1 (Ref. 2) and a high mobility group (HMG) family member, which is closely related to SRY and known as SOX9 (Ref. 3) (Fig. 1). Each of these proteins possesses a sequence-specific DNA-binding domain and a putative protein–protein interaction domain, attributes that suggest that these proteins regulate transcription. Expression patterns for the genes encoding these proteins indicate that their influence is exerted before the appearance of morphological differences between the sexes in the developing gonad. Mutations in WT1 and SOX9 give rise to disease phenotypes, which affect more than one organ system, indicative of an early role for these proteins in embryonic differentiation.

WT1 and transcriptional regulation. The DNA-binding domain of WT1 comprises four Cys_2His_2 (Krüppel-type) zinc fingers, which have significant homology to the EGR family of transcription factors. Chromosomal deletions first identified

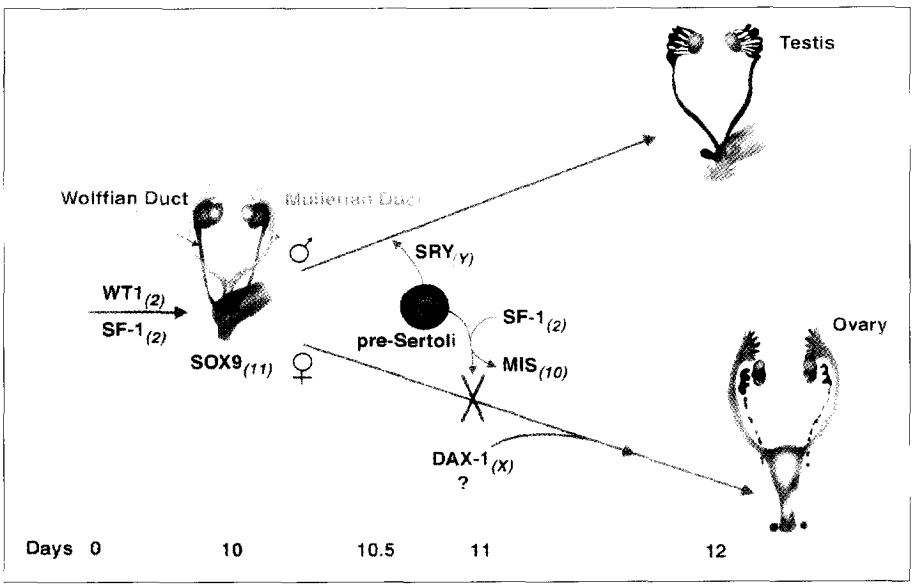

Figure 1

The proteins implicated in the mammalian sex determination pathway are illustrated for gonadal differentiation in the mouse. The approximate number of days *post coitum* at which transcripts for the indicated proteins are seen (bottom) is correlated with certain general events in differentiation. The mouse chromosome on which the corresponding gene has been mapped is indicated in parentheses. The anatomical drawings are for illustrative purposes only and do not represent the actual appearance of these structures in the developing embryo. The role of DAX-1 is considered speculative (see text) and is therefore indicated with a question mark. Abbreviation used: MIS, Müllerian inhibitory substance.

 0968–0004/96/$15.00 PII: S0968-0004(96)10032-3

in patients suffering from a collection of disorders known as WAGR syndrome (Wilms' tumors, Anridia, Genito-urinary abnormalities and mental Retardation) led to the mapping of the WT1 gene on human chromosome 11. Point mutations in WT1 often lead to the more severe Denys–Drash syndrome, characterized by childhood kidney cancers (Wilms' tumors), glomerular nephropathy and varying degrees of abnormal gonadal development[1]. Individuals heterozygous for certain point mutations display phenotypic features of XY sex-reversed females, in addition to other defects, reflecting the common embryological origin of the kidneys and gonads. Fingers 2–4 of WT1 share 60% sequence identity with the DNA-binding domain of Egr-1 (Ref. 4), whose structure provides a framework onto which WT1 mutations can be mapped (Fig. 2). These mutations occur either in residues forming part of the DNA-binding surface or in residues that coordinate zinc. Many of the mutations that result in Denys–Drash syndrome are expected to result in defective DNA-binding and, therefore, a loss of WT1 transcriptional regulation.

A minority of WT1 mutations affect the proportion of alternate splicing variants, which might provide insight into understanding the complex regulatory properties of the protein. Two different insertions naturally occur in WT1: the tripeptide Lys-Thr-Ser (KTS) can be inserted at position 407, between fingers 3 and 4 (Fig. 2); and an additional 17 amino acid insertion, with a potential phosphorylation site, can occur just amino-terminal of the residues that coordinate zinc. Thus, four alternate splicing variants, with either one of the insertions, both insertions or no insertions, are possible[5]. Alternatively spliced variants of WT1 have different promoter specificities and varying effects on levels of transcription[6]. Heterozygous mutations, which disrupt the activities of these isoforms, can have dominant negative effects by affecting the function of the wild-type allele, leading to nephropathy, intersex disorders and a predisposition to the development of Wilms' tumors.

The timing of expression of the gene encoding WT1 seems to be at the mesenchyme–epithelial transition in the developing embryo. An unusual proportion of Denys–Drash patients displaying abnormal genitalia (and other defects) appear to carry a single mutation in finger 3, which converts Arg394→Trp (Ref. 1) (Fig. 2). These individuals show large variability in the extent of the observed genital abnormalities, some with both Müllerian- and Wolffian-derived structures, others having neither. Murine homozygous nulls for *wt1* die at mid-gestation with complete failure in mesenchymal differentiation, which results in the absence of both kidneys and gonads. In addition, the histology seen in kidney stem cells that have lost WT1 function indicates that these cells continue to proliferate instead of going into terminal differentiation. These observations suggest that WT1 acts either to repress the expression of genes required to maintain cellular proliferation or to activate genes responsible for terminating differentiation.

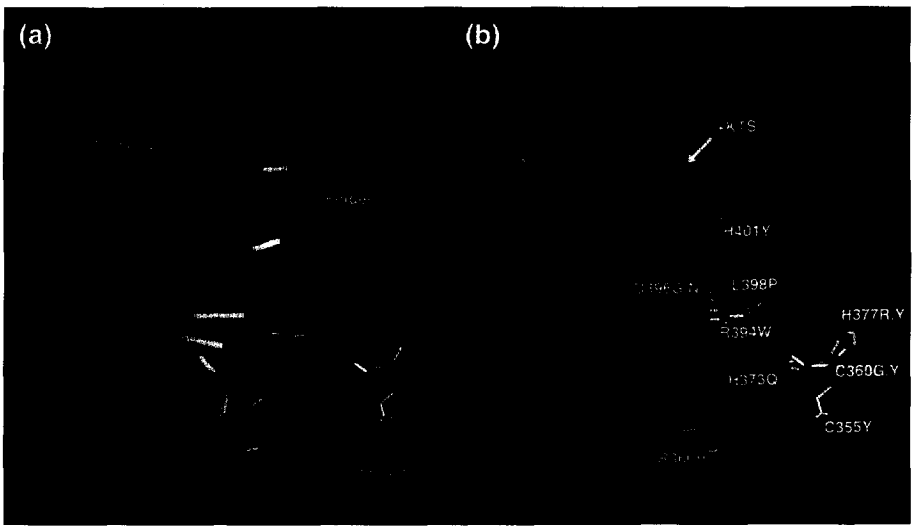

Figure 2

Mutations in WT1 associated with Denys–Drash syndrome affect the DNA-binding surface of the protein. Fingers 2–4 of WT1 are mapped on to the crystal structure of Egr-1 (Ref. 4). **(a)** Mutants that interact directly with DNA bases are indicated in yellow, those that coordinate zinc (blue balls) are shown in grey. **(b)** The point mutants identified in Denys–Drash syndrome are indicated. One additional point mutant is known in Finger 1 (not shown), namely C330Y and is involved in zinc coordination.

Co-transfection experiments, in which synthesis of WT1 was linked to a reporter construct in NIH-3T3 cells, indicated that a region of the protein outside of the DNA-binding domain is associated with strong transcriptional repression. Fusion of this domain, which contains a putative protein–protein interaction domain rich in proline and glutamine, to the DNA-binding domain of Egr-1 converted Egr-1 to a transcriptional repressor in the same assay system. Thus, WT1 appears to act as a transcriptional repressor[1]. As a consequence, WT1 is also a candidate tumor suppressor in the kidney (and perhaps gonad). To date, no candidate gene involved in gonadal differentiation has been identified whose expression is directly regulated by WT1.

SOX9. A second early gene product implicated in gonadal development is SOX9 (Ref. 7). Mutations in the *sox9* gene, located on human chromosome 17, have been linked to a severe dwarfism syndrome known as campomelic dysplasia (CD)[7]. Patients with CD display a number of congenital skeletal abnormalities and more than 30% are 46X,Y females with a gradation of genital defects. The phenotypes of these patients have been associated with mutations that lead to severely truncated proteins, or with chromosomal translocations that occur some distance upstream of the *sox9* gene, altering the level of gene expression[8]. As these mutations are present in only one allele of the gene, the skeletal and genital abnormalities seen in CD might be owing to haplo-insufficiency of the *sox9* gene product. The gene encodes a protein member of the HMG superfamily of DNA-binding proteins, which contains a DNA-binding domain that shares nearly 50% sequence identity to the DNA-binding domain of SRY. The similarity in the DNA-binding HMG domains of SOX9 and SRY suggests a common function for these gene products in transcriptional regulation. The three-dimensional structure of the SRY–DNA complex[9] therefore provides a framework from which to understand the biological function of SOX proteins in development.

Turning on testicular differentiation

The testis-determining factor SRY. During meiosis in males, aberrant crossing-over near the short arm telomeres of the X

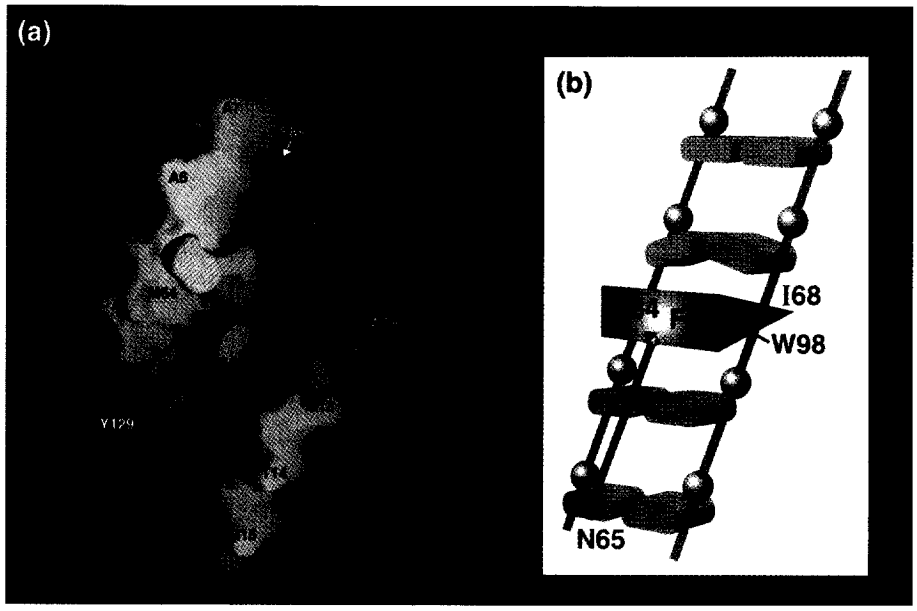

Figure 3

(a) Protein–DNA interactions in the SRY–HMG/DNA complex. Residues that directly interact with DNA bases are depicted as a yellow molecular surface attached to a green protein backbone. The intercalative wedge is formed by Met64, Asn65, Phe67, Ile68 and Trp98, and intercalates between A–T basepairs 5 and 6. **(b)** Schematic representation of the wedge indicating how it inserts into the DNA. The wedge is anchored by Asn65, which hydrogen-bonds to the C4–G13 basepair below the site of insertion.

and Y chromosomes can transfer Y-specific DNA into the genome of an otherwise XX individual. These XX 'sex-reversed' individuals, and their XY counterparts lacking a similar portion of the short-arm of the Y chromosome, led to the discovery of a factor responsible for testicular differentiation, the testis determining factor SRY (Refs 10, 11). SRY is a member of the HMG superfamily of DNA-binding proteins, a family characterized by an unusual L- or boomerang-shaped DNA-binding domain[9,12,13]. Sequence analyses of SRY proteins from a variety of mammals indicate that there is little sequence similarity, even among primate species, outside the DNA-binding HMG domain. Within the DNA-binding domain, however, sequence identity is greater than 60% (Refs 14–16). Mutations in SRY that lead to sex-reversal occur in the HMG box, with only one exception, strongly suggesting that the associated function of the SRY protein resides primarily in the DNA-binding domain itself. *In vitro* and *in vivo* studies of SRY have demonstrated that SRY binds specifically to the sequence AACAA(A/T)(G/C), bends DNA by more than 70° and is capable of transcriptional *trans*-activation[17,18].

The three-dimensional structure of the human SRY–HMG/DNA complex[9] confirmed earlier biochemical experiments suggesting that the protein binds exclusively in the minor groove[19]. Moreover, the structure revealed the presence of a 'mechanical wedge', formed by a tetrad of hydrophobic amino-acids that intercalates at an ApA basestep and unwinds the DNA helix[9] (Fig. 3). The structure of the SRY–HMG/ DNA complex provides an atomic view and consequently, affords an explanation for the molecular basis of 46X,Y sex reversal[9] (Fig. 4). Genetic defects in the SRY-HMG domain are expected to alter or disrupt the domain structure in three key regions: the hydrophobic triad formed by Val60, Tyr124 and Tyr127; the hydrophobic core of the domain; and the intercalative wedge comprising Met64, Asn65, Phe67, Ile68 and Trp98 (Fig. 4). The consequence of mutations in each of these three regions is a loss of DNA-binding, altered DNA-bending or both (Table I).

Targets for SRY activity. The ability of SRY to significantly alter DNA structure suggests that SRY has an architectural role in testicular differentiation by promoting the assembly of higher-order complexes of protein and DNA through DNA bending[20–22]. These complexes are likely to be constructed from one or more additional sequence-specific DNA-binding proteins that, together, form a functional regulatory unit. The promoter for the gene encoding the Müllerian inhibitory substance (MIS) is one potential target for SRY transcriptional regulation. The onset of MIS synthesis in mouse pre-Sertoli cells occurs early in gonadal differentiation, within 20 hours of the onset of SRY production[23].

Co-transfection assays with an SRY expression vector and a reporter construct containing the MIS promoter indicate that SRY is correlated with transcriptional activation at this promoter in a tissue culture model of the differentiating gonad[24]. In addition, SRY has a clear DNase I footprint covering 24 bp in the human MIS promoter, 6 bp of which (CACAAA) bear a strong resemblance to the consensus sequence derived from *in vitro* binding-site selection experiments, namely AACAA(A/T)[24]. Under some conditions, however, an oligonucleotide duplex comprising the DNase I footprint sequence fails to show an electrophoretic mobility shift[25]. This observation, coupled with the poor selection for C at position one in binding-site selection experiments, has led to the suggestion that SRY must bind weakly at the MIS promoter, and therefore could not have a direct role in transcriptional regulation of MIS. The three-dimensional structure of the complex of SRY with the DNA octamer derived from the MIS promoter[9], as well as additional biochemical data, suggest that SRY, or an SRY-like protein, can regulate MIS synthesis.

SRY clearly forms an extremely stable, specific complex (K_d in the nM range) with an octamer sequence contained within the DNase I footprint of SRY at the human MIS promoter (GCACAAAC)[9]. Complexes of the SRY- HMG box with this octamer duplex are readily observed in electrophoretic mobility shift experiments, and SRY protects basepairs 2–8 of this sequence from cleavage by hydroxyl radicals (M. H. Werner, A. M. Gronenborn, G. M. Clore and M. Bianchi, unpublished). The identical octamer sequence was used in the three-dimensional structure determination of the SRY–HMG/DNA complex, which was carried out at 37°C; approximately 13°C above the melting temperature of the free octamer duplex alone[9]. Moreover, the C–G basepair at position two, which is substituted for an A–T basepair in the consensus sequence derived from *in vitro* selection experiments, does not participate in any base specific interactions with SRY (Ref. 9). Thus, there is little question that SRY can form a stable, specific complex with a GCACAAAC duplex. In the context of the full MIS DNase I footprint, however, this sequence is degenerate (CACAAACAC), a fact that might complicate the

analysis of electrophoretic mobility experiments and could partially account for the failure to observe an SRY band-shift with an oligonucleotide duplex encompassing the DNase I footprint under certain circumstances[25]. Apparent DNA-binding affinity alone, therefore, might not adequately reveal the complexity of interaction between SRY and the MIS promoter.

Mutations within the DNase I footprint sequence in the MIS promoter, which reduced the apparent binding affinity of SRY to this sequence *in vitro*, did not suppress the transcriptional response of this mutant locus to SRY synthesis *in vivo*[24]. As the affinity of SRY for the mutant locus was significantly lowered, one interpretation of this observation is that the regulation of MIS synthesis *in vivo* is indirect, perhaps involving intervening factor(s) (SRYIFs) whose production is regulated by SRY (Ref. 24).

As already noted, it is inadequate to assess the activity of SRY in terms of apparent binding strength alone without considering the functional context of SRY action. An alternative interpretation is that SRY-induced DNA bending is not disrupted in the *in vivo* expression assay, even if the apparent *in vitro* affinity of SRY alone for a specific DNA-binding target is weakened. Hence, the key question is: if the affinity of SRY is reduced for a certain sequence, does this disrupt the ability of SRY to bend the DNA at that locus?

The ability of a weakly bound DNA-bending protein to support the assembly of a higher-order complex has recently been demonstrated for integration host factor (IHF), a protein in *Escherichia coli* that normally promotes the assembly of the phage λ intasome through sequence-specific bends[26]. In certain contexts, a non-sequence specific *E. coli* analog of IHF (known as the HU protein) localizes to the IHF site, where it can functionally replace IHF (Ref. 27), suggesting that the non-specifically bound protein can establish the necessary architecture at the bending locus during assembly of a functional complex. From this perspective, SRY might function at a mutant MIS promoter despite the apparent reduced affinity of SRY for the mutant DNA-binding site. High affinity DNA binding of SRY to MIS might, therefore, be important for efficient targeting of the DNA-bending protein to the optimal bending locus. However, even if the bending locus at the MIS promoter is comprised of a sub-optimal binding site, SRY might still have the ability to aid in the assembly of a functional complex; for example, the mutant Lys106→Ile binds 100-fold more weakly to an SRY-binding site, but induces a wild-type bend in the DNA *in vitro*[28]. This interpretation obviates the necessity of involving SRYIFs in transcription of the gene encoding MIS. Instead, it is possible that a DNA bend induced by SRY, or more generally, an SRY-like protein, is especially favorable for promoting the assembly of higher-order structures, even in a sub-optimal DNA-binding mode. Thus, it is plausible that either SRY or an SRY-related HMG-box protein (SOX protein) is an important component of MIS gene regulation.

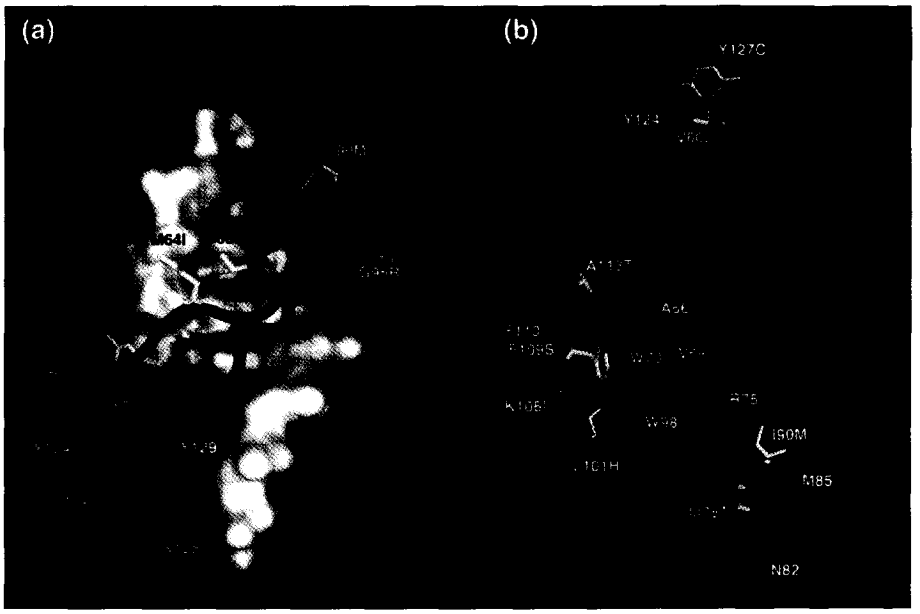

Figure 4

Mutations in SRY that result in 46X,Y sex reversal. **(a)** Mutations that impinge on the interaction of SRY with DNA are shown in yellow, non-mutated residues whose interaction is affected by other mutations are shown in magenta, and the hydrophobic triad stabilizing the amino and carboxyl terminii is shown in red. **(b)** Mutations that destabilize the high mobility group (HMG)-domain fold are shown in yellow, non-mutated residues that form the balance of the internal packing clusters in the domain are shown in red.

Suppression of the female phenotype

Androgen receptor. After development of the testes, commitment to the male sex involves two simultaneous pathways, one stimulatory and the other inhibitory. The stimulatory events require high levels of androgens and a functional androgen receptor (AR). The AR is a classical nuclear hormone receptor closely related to the estrogen receptor (ER), glucocorticoid receptor (GR) and mineralocorticoid receptor (MR) proteins. Mutations in the human gene encoding the AR, located on the X chromosome, lead to a gradation of genital defects owing to androgen insensitivity (androgen insensitivity syndrome, AIS)[29]. More than 150 mutations have been reported to date that primarily affect either the DNA-binding or the steroid-binding domains of the AR protein, of which those in the DNA-binding zinc-finger domain are expected to directly interfere with either DNA binding, protein dimerization or a combination of these[29,30].

Müllerian inhibitory substance. Regression of female reproductive structures is accomplished by MIS (Ref. 31). MIS is one of the first proteins produced by the developing testis before its morphological differentiation, and its synthesis is restricted to the Sertoli cells. In the adult, MIS synthesis is detectable in both the testicular Sertoli cells and at very low levels in the granulosa cells of the postnatal ovary. The observation of a sexually dimorphic pattern of synthesis linked MIS to the pathway for sex determination and, consequently, led to the identification of its gene on human chromosome 19.

MIS is a homodimeric glycoprotein of the transforming growth factor β (TGF-β) superfamily. Post-translational proteolysis of the 140 kDa dimer activates the hormone, but the two proteolytic fragments remain associated with one another, even though the carboxy-terminal 25 kDa dimer fragment is independently capable of regressing Müllerian ducts[32]. The extent of glycosylation approaches 18% by weight in the rat protein, but removal of carbohydrate by

Table I. The effect of point mutations in SRY that result in 46X,Y sex reversal[a]

Mutation	DNA binding	DNA bending	Structural defect	Ref(s)
De novo				
Arg62→Gly	NT	NT	Removes salt bridge to DNA backbone	45
Met64→Ile	<WT	<WT	Distorts intercalative hydrophobic wedge	28, 46
Ile68→Thr	«WT	NT	Distorts intercalative hydrophobic wedge	24
Met78→Thr	NT	NT	Destabilizes protein core	45
Gly95→Arg	«WT	NT	Distorts intercalative hydrophobic wedge and introduces steric clash with DNA backbone	28, 46
Leu101→His	NT	NT	Destabilizes protein core	47
Ala113→Thr	NT	NT	Destabilizes protein core	48
Tyr127→Cys	NT	NT	Destabilizes protein core	49
Lys133→Trp	NT	NT	Unknown[b]	45
Inherited				
Val60→Leu	«WT	NT	Destabilizes protein core	46
Ile90→Met	≈WT	NT	Destabilizes protein core	28, 46
Phe109→Ser	≈WT	NT	Destabilizes protein core	28
Unknown origin				
Lys106→Ile	«WT	≈WT	Destabilizes protein core	28, 46

[a]The apparent effects of point mutations on DNA binding and bending were analysed by electrophoretic mobility shift and circular permutation, respectively, and the references for these studies are indicated. *De novo* mutations are those not present in the paternal Y chromosome. Abbreviations used: NT, not tested; ≈WT, similar to wild type; <WT, complex visible by electrophoretic mobility shift, but with significantly reduced DNA-binding affinity; «WT, complex not visible by electrophoretic mobility shift. The Met64→Ile mutation reduces the bend angle by approximately 25%.
[b]Lys133 was outside the region of the hSRY–HMG domain whose structure was solved in the DNA complex[9].

endoglycosidase F does not alter its activity. Mutations in MIS have been found in some patients with persistent Müllerian duct syndrome (PMDS), a rare form of male pseudohermaphroditism, which is characterized by the presence of a uterus and Fallopian tubes in XY individuals that are overtly male in phenotype[33,34]. Other cases of PMDS, in which normal levels of serum MIS are detected, might be owing to mutations in the receptor for MIS (Ref. 35).

SF-1 is an orphan nuclear receptor protein, which has been implicated in the regulation of the steroid hydroxylases, and perhaps has a developmental role in the differentiation of the adrenals and gonads[2,25]. Although SF-1 transcripts appear as early as day nine in the mouse, quantitative analysis of the expression patterns of the genes encoding SF-1 and MIS in developing rat embryos demonstrate that expression levels are coincident during the interval of Müllerian duct regression. This evidence supports *in vitro* DNA-binding experiments that demonstrated SF-1 binding to sequences derived from the MIS promoter[25]. A luciferase reporter construct under the control of the MIS promoter in HeLa cells is unresponsive to SF-1, but is constitutively activated by an SF-1 deletion construct, indicating that a *trans*-regulatory domain exists in the deleted carboxy-terminal half of the protein[25]. This domain either binds an unidentified ligand specific to Sertoli cells or instead, might be a protein–protein interaction domain.

DAX-1. A binding site for SF-1 has recently been identified within the promoter of another gene implicated in the sex-determination pathway, *dax-1* (Ref. 36). Mutations in DAX-1 lead to adrenal hypoplasia congenita (AHC), an X-linked disorder, in which the adrenal glands fail to fully develop[37]. AHC is frequently associated with a failure of sexual maturation at puberty, a condition known as hypogonadotropic hypogonadism (HHG), connecting DAX-1 to the development of gonadal structures, although its influence appears to be indirect[37]. While SF-1 null mutants fail to develop adrenals or gonads, DAX-1 mutations lead only to arrested development, with the adrenals failing to differentiate beyond the fetal stage. Thus, the influence of DAX-1 must be downstream of SF-1 in the developmental hierarchy. The *dax1* gene maps to a region of the X chromosome (Xp21) that had previously been associated with dosage sensitive sex reversal (DSS), AHC and HHG, but does not comprise the entire Xp21 region. DSS is a consequence of a duplication of Xp21 and some have speculated that sex reversal in these cases results from a female-specific function, which is turned on in XY individuals owing to the extra dosage of the *dax-1* gene product[38,39]. In a normal XY male, the lower dosage of the *dax-1* gene product might be incapable of stimulating this response[40]. The *dax-1* gene encodes a protein believed to be a member of the nuclear hormone receptor superfamily, on the basis of a roughly 50% continuous similarity between its carboxy-terminal portion and the ligand-binding domain of the nuclear hormone receptors. The amino-terminal 258 amino acids comprise four incomplete repeats of a novel alanine/glycine-rich, 65–67-amino acid sequence motif[37]. Having only recently been cloned, no targets for DAX-1 have yet been identified.

A regulatory hypothesis for gonadal differentiation

The gene regulatory proteins in gonadal development identified to date cluster into two families, the (orphan) nuclear hormone receptors (WT1, SF-1 and DAX-1) and the HMG architectural proteins (SOX9 and SRY). The structure of the SRY–HMG/DNA complex suggests that there might be a more general role for SOX proteins as architectural elements in developmental gene expression. Approximately 20 SOX proteins have been identified to date, many of which are associated with regulating cell specialization events during development[41–43]. Although the organization of functional domains within SOX proteins varies, the SOX subfamily is characterized by a highly conserved DNA-binding HMG box. Taking into account the even higher degree of conservation for residues that contact the DNA, it is likely that the SOX subfamily will display nearly the same preference for the AACAA(A/T) sequence found for SRY. Thus, with respect to DNA binding and bending, SOX proteins could generally be expected to induce a similar kind of distortion at a specific DNA target as that observed in the SRY–HMG/DNA complex. What might distinguish one SOX protein from another is either the timing of its synthesis during stages of development, the role other domains within a SOX protein play in assembling a functional regulatory complex at a specific promoter, or both.

A significant component of the regulatory cascade in gonadal differentiation might, therefore, be the assembly of appropriately timed complexes consisting of a SOX protein and other promoter-specific DNA-binding proteins (Fig. 5). In these higher-order assemblies, additional sequence-specific DNA-binding proteins might or might not interact

with the SOX protein, but almost certainly interact with each other to form a functional complex. The amino acid sequences for both mouse and human SOX9 proteins contain conserved HMG DNA-binding domains and proline/glutamine-rich putative protein–protein interaction domains, suggesting that recruitment of regulatory proteins to the site of SOX-induced bending can involve direct contact with the architectural element. A serine-rich domain from SOX4 has been shown to *trans*-activate expression, confirming a role for protein–protein interactions in some SOX functional contexts[44]. It is anticipated, therefore, that many SOX gene products will contain conserved DNA-binding and protein-binding domains. As SOX proteins probably exhibit very similar DNA-sequence specificities, the fidelity of a SOX-induced nucleoprotein assembly might be defined by the protein–protein interaction surface(s) and the DNA sequence specificity of any additional proteins that act at the same control locus, rather than their own DNA specificity.

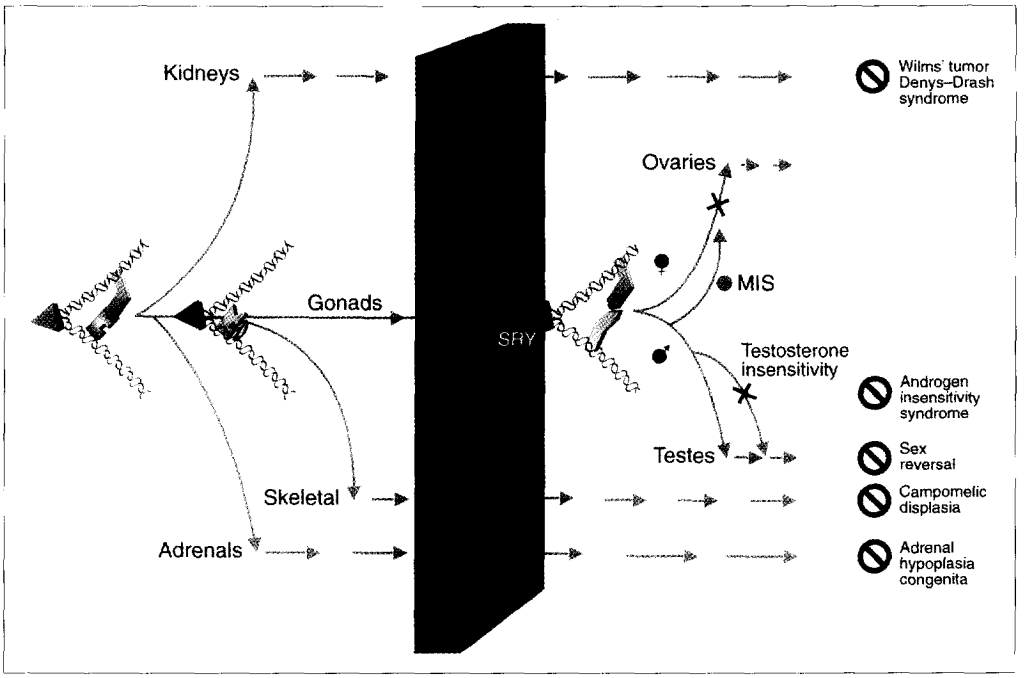

Figure 5

Putative model for an architectural regulatory hierarchy in development. The hierarchy of regulatory events in embryogenesis is thought to be coupled to the assembly of higher-order nucleoprotein complexes, whose assembly is promoted by SOX proteins (triangles). Early events, which might involve one or more (orphan) nuclear hormone receptor proteins, SOX proteins and other proteins separate kidney, gonad, adrenal and skeletal developmental pathways. Arrows represent different pathways, but do not imply exact timing or extent of development along that path. The monolith represents a point of commitment in embryonic development after which certain group(s) of cells are committed to specific organ system developmental pathways. The influence of SRY on testicular differentiation is expected to involve other regulatory proteins that might or might not interact with SRY itself. SRY stimulation of testicular differentiation shuts off the further development of Müllerian ducts as a consequence of Müllerian inhibitory substance (MIS) released from testicular Sertoli cells. Failures in the actions of several proteins discussed in the text lead to clinical disorders, which affect not only the indicated organ system, but also gonadal differentiation and, thereby, sex determination.

The delayed synthesis of MIS relative to the onset of SRY production allows for the possibility that an unidentified SOX protein, perhaps under the direct control of SRY, might actually be the architectural regulator at the MIS promoter, yielding the simplest explanation for SOX regulation of MIS. The hunt for SOX transcripts and/or proteins, together with their characterization, might therefore provide an avenue to identify many morecomponents of the sex determination pathway in mammals.

Acknowledgements

The authors wish to acknowledge H. Nash for suggesting an alternative interpretation of SRY activity at a mutant MIS promoter; M. Bianchi and K. Parker for stimulating discussions; M. Bianchi, R. Behringer, K. Williamson, N. Hastie, K. Parker and C. Quigley for providing additional information; and J. Aarons for artwork. This work was supported by the AIDS Targeted Antiviral Program of the Office of the Director of the National Institutes of Health (to G. M. C. and A. M. G.). All correspondence should be addressed to G. M. C. and A. M. G.

References

1 Hastie, N. D. (1994) *Annu. Rev. Genet.* 28, 523–558
2 Ikeda, Y., Shen, W-H., Ingraham, H. A. and Parker, K. L. (1994) *Mol. Endocrinol.* 8, 654–662
3 Foster, J. W. *et al.* (1994) *Nature* 372, 525–530
4 Pavletich, N. P. and Pabo, C. O. (1991) *Science* 252, 809–817
5 Drummond, I. A. *et al.* (1994) *Mol. Cell. Biol.* 14, 3800–3809
6 Reddy, J. C. *et al.* (1995) *J. Biol. Chem.* 270, 10878–10884
7 Wagner, T. *et al.* (1994) *Cell* 79, 1111–1120
8 Wright, E. *et al.* (1995) *Nat. Genet.* 9, 15–20
9 Werner, M. H., Huth, J..R., Gronenborn, A. M. and Clore, G. M. (1995) *Cell* 81, 705–714
10 Gubay, J. *et al.* (1990) *Nature* 345, 245–250
11 Sinclair, A. H. *et al.* (1990) *Nature* 346, 240–244
12 Weir, H. M. *et al.* (1993) *EMBO J.* 12, 1311–1319
13 Read, C. M. *et al.* (1995) *Nucleic Acids Mol. Biol.* 9, 222–250
14 Whitfield, L. S., Lovell-Badge, R. and Goodfellow, P. N. (1993) *Nature* 364, 713–715
15 Laudet, V., Stehelin, D. and Clevers, H. (1993) *Nucleic Acids Res.* 21, 2493–2501
16 Pontiggia, A. *et al.* (1995) *Gene* 154, 277–280
17 Ferrari, S. *et al.* (1992) *EMBO J.* 12, 4497–4506
18 van de Wetering, M. and Clevers, H. (1992) *EMBO J.* 11, 3039–3044
19 Giese, K., Cox, J. and Grosschedl, R. (1992) *Cell* 69, 185–195
20 Nash, H. A. (1990) *Trends Biochem. Sci.* 15, 222–227
21 Grosschedl, R., Giese, K. and Pagel, J. (1994) *Trends Genet.* 3, 94–100
22 Werner, M. H., Gronenborn, A. M. and Clore, G. M. (1996) *Science* 271, 778–784
23 Hacker, A., Capel, B., Goodfellow, P. N. and Lovell-Badge, R. (1995) *Development* 121, 1603–1614
24 Haqq, C. M. *et al.* (1994) *Science* 266, 1494–1500
25 Shen, W-H. *et al.* (1994) *Cell* 77, 651–661
26 Nash, H. A. in *Regulation of Gene Expression in Escherichia coli* (Lin, E. C. C. and Lynch, A. S., eds), R. G. Landes (in press)
27 Segall, A. M., Goodman, S. D. and Nash, H. A. (1994) *EMBO J.* 13, 4536–4548
28 Pontiggia, A. *et al.* (1994) *EMBO J.* 13, 6115–6124
29 Quigley, C. A. *et al.* (1995) *Endocr. Rev.* 16, 271–321
30 Zanaria, E. *et al.* (1994) *Nature* 373, 635–641
31 Behringer, R. R., Finegold, M. J. and Cate, R. L. (1994) *Cell* 79, 415–425
32 Gustafson, M. L. and Donahoe, P. K. (1994) *Annu. Rev. Med.* 45, 505–524
33 Guerrier, D. *et al.* (1989) *J. Clin. Endocrinol. Metab.* 68, 46–52
34 Imbeaud, S. *et al.* (1994) *Hum. Mol. Genet.* 3, 125–131
35 Behringer, R. R. (1995) *Curr. Top. Dev. Biol.* 29, 171–187
36 Burris, T. P., Guo, W., Le, T. and McCabe, E. R. B.

(1995) *Biochem. Biophys. Res. Commun.* 214, 576–581
37 Muscatelli, F. *et al.* (1994) *Nature* 372, 672–676
38 Bardoni, B. *et al.* (1994) *Nat. Genet.* 7, 497–501
39 King, V. *et al.* (1995) *Curr. Biol.* 5, 37–39
40 Ryner, L. C. and Swain, A. (1995) *Cell* 81, 483–493
41 Sockanathan, S., Cohen-Tannoudji, M., Colignon, J. and Lovell-Badge, R. (1993) *Genet. Res.* 61, 149
42 Kamachi, Y. *et al.* (1995) *EMBO J.* 14, 3510–3519
43 Denny, P. *et al.* (1992) *EMBO J.* 11, 3705–3712
44 van de Wetering, M., Oosterwegel, M., van Norren, K. and Clevers, H. (1993) *EMBO J.* 12, 3847–3854
45 Affara, N. A. *et al.* (1993) *Nucleic Acids Res.* 2, 785–789
46 Harley, V. R. *et al.* (1992) *Science* 255, 453–456
47 Braun, A. *et al.* (1993) *Am. J. Hum. Genet.* 52, 578–585
48 Zeng, Y. *et al.* (1993) *J. Med. Genet.* 30, 655–657
49 Poulat, F. *et al.* (1994) *Hum. Mutat.* 3, 200–204

Why mammalian cell surface proteins are glycoproteins

Carl G. Gahmberg and Martti Tolvanen

Most proteins presented at the external surface of mammalian cells contain carbohydrate. The reason for this is not fully understood, but recent work has shown that such carbohydrate has two major functions. Inside the cell, it helps proteins fold and assemble correctly in the endoplasmic reticulum, and it might also act as a signal for the correct migration of glycoproteins. Outside the cell, it provides specific recognition structures for interaction with a variety of external ligands.

EARLY WORK IN the 1960s on the morphology of mammalian cells showed that the external surface of the plasma membrane is rich in carbohydrate, whereas the inner side is devoid of conventional-type oligosaccharides[1]. Until recently, it was unclear whether the carbohydrate was confined to few or many different cell surface glycoconjugates[2], and the development of radioactive techniques in particular has allowed the carbohydrate portions of exposed cell surface glycoproteins and glycolipids to be labeled[3,4]. It was then possible to study the larger number of glycoconjugates specifically presented at the surface of various cells. Also, it became apparent that cell membranes contain a multitude of glycoproteins, many more than previously thought[5,6].

In 1976, after studying human erythrocytes and other cells, it was proposed that cell surface proteins are always glycoproteins[5]. A similar proposal was made independently by Bretscher and Raff[2]. As more and more mammalian cell membrane proteins and the genes encoding them have been characterized, cloned and sequenced, this proposal has turned out to be largely correct.

Most glycoproteins are *N*-glycosylated, i.e. they contain asparagine-linked oligosaccharides located at the peptide sequence(s) NxS/T (where x stands for any amino acid except for proline) at the external aspect of the membrane. Some membrane proteins are *O*-glycosylated, with the carbohydrate chains attached to serine or threonine residues, which are often clustered in distinct regions of the polypeptides. Most *O*-glycosylated proteins also contain one or more *N*-glycosidic oligosaccharides[7]. The importance and requirements for *O*-glycosylation have yet to be elucidated.

Exceptions to the rule

Early searches of the literature for unglycosylated surface proteins in mammalian cells met with little success. However, in 1982, the human red cell Rh(D) (Rhesus) protein, with an apparent molecular weight of 30–32 kDa, was identified[8,9]. Importantly, no evidence for the presence of carbohydrate was found[10], and subsequent cloning and sequencing of its cDNA and that of other polypeptides belonging to the Rh-blood group system, showed that they do indeed lack *N*-glycosylation sequences[11,12]. Thus, this protein seemed to be an exception to the glycosylation rule.

However, more recent work has shown that this is not the case. There is now evidence that the Rh-polypeptides form part of a large glycopolypeptide complex, including among others the Rh50 glycoproteins and the Landsteiner–Wiener (LW) blood group glycoprotein (intercellular adhesion molecule 4) (Refs 13, 14). This situation is similar to that of β2-microglobulin in class I transplantation antigens, where the unglycosylated protein associates with the heavy chains of the transplantation antigens. The importance of the association of the Rh-polypeptide with other glycosylated proteins is underscored by the fact that it has not yet been possible to express the Rh cDNA in any mammalian cell expression system.

In a recent survey of the SWISS-PROT database (release 33.0, April 1996) we found 1823 complete animal protein entries with reported extracellular features, of which 1671 (91.7%) were described as 'glycoproteins' in the keyword field; 1630 of these 1671 contained the *N*-glycosylation peptide sequence NxS/T. The remaining 8.3%, representing 152 potentially non-glycosylated plasma-membrane proteins, contained 116 proteins with multiple transmembrane regions, 15 proteins that are known to associate with glycosylated subunits in a complex such as CD3 chains and the Rhesus D-polypeptide, and seven that contained 5–38 potential *N*-glycosylation sites, i.e. polypeptides highly likely to be glycosylated, yet not marked as glycoproteins. This leaves only 14 sequences (0.7%) that are candidates for non-glycosylated, non-complexed plasma membrane proteins with a single transmembrane domain.

In another survey, we assessed whether this high representation of the *N*-glycosylation tripeptide sequence is more than should be found by chance alone*. To do this, we extracted all sequence features marked as extracellular domains from animal proteins in SWISS-PROT 33.0, which resulted in 4259 stretches of sequence from 1933

C. G. Gahmberg and **M. Tolvanen** are at the Department of Biosciences, Division of Biochemistry, P.O. Box 56, Viikinkaari 5, FIN-00014, University of Helsinki, Finland.

 0968–0004/96/$15.00 PII: S0968-0004(96)10034-7