0888-8809/02/$15.00/0
Printed in U.S.A.

Molecular Endocrinology 16(6):1269–1279

NUBIScan, an *in Silico* Approach for Prediction of Nuclear Receptor Response Elements

MICHAEL PODVINEC, MICHEL R. KAUFMANN, CHRISTOPH HANDSCHIN, AND URS A. MEYER

Division of Pharmacology/Neurobiology, Biozentrum of the University of Basel, CH-4056 Basel, Switzerland

Nuclear receptors (NRs) are transcription factors activated by a multitude of hormones, other endogenous substances, and exogenous molecules. These proteins modulate the regulation of target genes by contacting their promoter or enhancer sequences at specific recognition sites. The identification of these response elements is the first step toward detailed insight into the regulatory mechanisms affecting a gene. We have developed NUBIScan, a computer algorithm to predict DNA recognition sites for NRs in the regulatory regions of genes. The algorithm is based on weighted nucleotide distribution matrices and combines scores from both half-sites necessary for NR dimer binding. It provides more specific identification of functional sites than previous *in silico* approaches, as evidenced by scanning published regulatory regions of drug-inducible genes and comparing the obtained predictions with experimental results. In prospective analyses, NUBIScan consistently identified new functional NR binding sites in sets of large sequences, which had eluded previous analyses. This is exemplified by the detailed functional analysis of the flanking region of two genes. This approach therefore facilitates the selection of likely sites of gene regulation for subsequent experimental analysis. (*Molecular Endocrinology* 16: 1269–1279, 2002)

THE LARGE FAMILY of intracellular receptors and transcription factors collectively called nuclear receptors (NRs) has raised considerable research interest in past years (1, 2). It comprises more than 300 members, which mediate transcriptional regulation of target genes by binding to DNA response elements (REs) in promoter or enhancer regions of the regulated gene. Ligand binding of the receptor triggers the recruitment of coactivators and dissociation of corepressors and leads to changes in chromatin structure. This allows increased binding of the transcription initiation complex to the promoter and increases the amount of mRNA transcribed from the gene (3, 4).

The ligands for NRs are extremely diverse, including steroid hormones, vitamin D_3, thyroid hormones, fatty acids, eicosanoids, RA, bile acids, sterols, and numerous drugs and other xenochemicals. Of particular recent interest in pharmacogenomics are the xenobiotic- or drug-sensing receptors, constitutive androstane receptor, pregnane X receptor, also termed steroid and xenobiotic receptor or pregnenolone-activated receptor (PXR/SXR/PAR), and chicken xenobiotic receptor (CXR) (5, 6).

Cognate REs for NRs are repeats of single DNA hexamers in a distinct arrangement toward each other in terms of relative orientation and spacing. The hexamer half-sites have a canonical consensus sequence of AG(GlT)TCA. Half-site repeats are categorized into direct repeats (DRs) and palindromic inverted repeats (IRs) or everted repeats (ERs). The majority of NRs binds as homo- or heterodimers to these hexamer repeats, each dimer partner interacting with one of the hexamers (1). Variations in the half-site sequence, their relative orientation, and the length of the spacer sequence allow for a wide range of different REs, enabling specific binding of a given receptor dimer to multiple target genes with different affinities, as well as giving rise to integration of different signaling pathways by competition of different receptors at a given binding site (7).

Available laboratory techniques to identify likely REs in target genes are tedious and time consuming and include DNAse I-hypersensitivity assays, subfragmentation, and reporter gene experiments. Bioinformatics-based approaches could considerably facilitate these studies, but present methods have been poor predictors of NR REs because of a low degree of sequence conservation of these elements. Here, we describe a computer algorithm capable of predicting weakly conserved recognition sites for transcription factors that bind as dimers, such as NRs. Compared with present approaches, our algorithm provides more specificity in the recognition of putative NR REs, making the analysis of large stretches of genomic sequences feasible. The algorithm is based on weighted nucleotide distribution matrices and the combined analysis of both half-sites of the RE. We have named this data-mining tool NUBIScan for "nuclear receptor binding site scanner."

Abbreviations: CAT, Chloramphenicol acetyltransferase; CXR, chicken xenobiotic receptor; CYP, cytochrome P450; DR, direct repeat; ER, everted repeat; GE, glutethimide; GLUT2, glucose transporter 2 gene; IR, inverted repeat; MDR1, multidrug resistance gene 1; NR, nuclear receptor; PB, phenobarbital; PBRU, phenobarbital response unit; PXR, pregnane X receptor; RE, response element; XREM, xenobiotic-responsive enhancer module.

To assess the validity of predictions by this algorithm, we have analyzed the sequences of three genes known to be regulated by NRs and compared our findings to experimental evidence. Moreover, we have used NUBIScan as a tool to identify new functional NR REs in the chicken cytochrome P450 (CYP) 2H1 and 3A37 genes. We find that the predictions obtained with the NUBIScan algorithm correlate well with experimental findings and provide an efficient way to select putative regulatory regions in large genomic sequences for experimental analysis.

RESULTS AND DISCUSSION

Testing the Algorithm with Known REs

To assess the validity of predictions by the NUBIScan algorithm, known REs for NRs discovered by classical methods were used as test cases. Results are summarized in Table 1.

First, we examined the 5′-flanking region of the human multidrug resistance gene 1 (MDR1). MDR1 codes for P-glycoprotein, an efflux pump responsible for actively moving hydrophobic compounds out of cells. It is considered to be a first defense mechanism against potential toxic substances and was initially discovered in cancer cells as a resistance mechanism against cytostatic drugs. P-glycoprotein is found in intestine, liver, kidney, and the blood-brain barrier. Intestinal MDR1 expression was shown to be inducible by an array of compounds similar to those mediating induction of the human CYP3A4 gene via PXR/RXR (8). More recently, Geick *et al.* (9) have published a PXR/RXR RE in the 5′-flanking region, mediating drug induction of the human MDR1 gene.

For our analysis, we have selected 28′630 bp from a bacterial artificial chromosome clone sequence (GenBank accession no. AC002457), corresponding to the sequence between the end of the upstream gene's coding sequence and the beginning of the MDR1 coding sequence.

Within this sequence, we scanned for high-scoring matches to DR4, DR3, and ER6 motifs, using the matrix described in Fig. 1. These motifs are known to be recognized by PXR/RXR and have been found in the MDR1 RE. The functional DR4(I) element (9) had the best DR4 match with a Z score of 10.23 (Fig. 2A). An ER6 motif, which binds PXR/RXR heterodimers *in vitro*, was predicted as the second-best ER6. Two further DR4 elements, DR4(II) and DR4(III), which could not be shown to be functional, are found on the 17th rank with identical Z scores of 6.94. Also, a DR3 element found in the sequence but shown to be ineffective was found at third rank with a Z score of 7.59 (Fig. 2A).

Next, we investigated the flanking region of the human CYP3A7 gene (10). In analogy with the closely related CYP3A4, the regulation of CYP3A7 was found to be mediated by PXR chiefly through a distal xenobiotic-responsive enhancer module (XREM) at −8 kb and to a small extent through a previously published PXR binding site at –153 bp in the proximal promoter (10, 11). The XREM is made up of an ER6 and a DR3 element in close association, which were termed

Table 1. Predicted NR REs in Regulatory Regions of Various Genes

Gene	Length[a]	Receptor	Site Type	Predicted RE		Rank[c]	Z Score	Reporter[d]	TA[e]	EMSA[f]	Mutation[g]
				Position	Name[b]						
MDR1	28.6 kb	PXR/RXR	DR4	−8544	DR4(I)	1	10.23	F	ND	F	F
			ER6	−8521	ER6	2	7.07	NF	ND	F	F
			DR3	−8544	DR3	3	7.59	NF	ND	NF	NF
CYP3A7	11.2 kb	PXR/RXR	DR3	−7748	dNR1	1	8.76	F	ND	ND	ND
			ER6	−7718	dNR2	2	7.42	F	ND	ND	ND
				−221	pNR	3	7.38	Weak	F	F	F
GLUT2	5.2 kb	PPARγ/RXR	DR1	+68	PPARE	1	7.19	ND	F	F	F
CYP3A37	3.4 kb	CXR/RXR	DR4	−1082	159-bp PBRU	1	8.52	F	F	F	F
CYP2H1	4.8 kb	CXR/RXR	DR4	−1636	264-bp PBRU	1	8.92	F	F	F	F
				−4198	ND	2	8.14	NF	ND	ND	ND
				−4974	240-bp PBRU	3	7.02	F	F	F	F
				−1906	ND	4	6.74	NF	ND	ND	ND
				−1519	264-bp PBRU	5	6.56	F	F	F	F

F, Functional; NF, not functional; ND, not determined.
[a] Length of analyzed sequence.
[b] According to original publication.
[c] In a list of predictions sorted descending by Z score.
[d] Activity of element in reporter gene assays.
[e] Activity of element in transactivation assays.
[f] Ability of element to bind the receptor dimer in EMSAs.
[g] Stated as functional if mutation of the element was shown to abolish either activity in reporter gene assays or transactivation assays or binding in EMSAs.

A

Halfsite	Sequence	Genes	Receptors	References
1	AGTACA	CYP450 2b10 CYP450 2B2 CYP450 2B6	CAR/RXR	(25-27)
2	AGTTGA	CYP450 2b10 CYP450 2B2	CAR/RXR	(25, 26)
3	GTGTCA	CYP450 2b10 CYP450 2B2	CAR/RXR	(25, 26)
4	GGTTCA	CYP450 2B6 Osteopontin	CAR/RXR VDR/RXR	(25) (28)
5	GGGTCA	CYP450 2B6 MoMLV	CAR/RXR TR/RXR	(25) (28)
6	AGGTCC	MoMLV	TR/RXR	(28)
7	AGGACA	ACOA	PPAR/RXR	(29)
8	AGGGCA	CYP4A6(Z) ApoAI	PPAR/RXR RAR/RXR	(30) (28)
9	AGAACA	-	Canonical GR	(28)
10	AGTTCA	-	Canonical NR halfsite (version 1)	(28)
11	AGGTCA	-	Canonical NR halfsite (version 2)	(28)

B

i	*1*	*2*	*3*	*4*	*5*	*6*
p_i(A)	0.72	0	0.09	0.27	0	0.90
p_i(C)	0	0	0	0	0.90	0.09
p_i(G)	0.27	0.90	0.54	0.09	0.09	0
p_i(T)	0	0.09	0.36	0.63	0	0
W_i	57.73	78.02	33.89	37.96	78.02	78.02

Fig. 1. Contributing Sequences and the Derived Matrix

In panel A, the sequences that were used in the construction of the weighted stochastic matrix are shown. In total, 11 experimentally verified half-site sequences were taken from the literature. Although five sequences are representative of NRs in general (half-sites 6, 7, and 9–11), six half-site sequences were chosen that are implicated in drug-mediated transcriptional regulation. In this way, balance between general validity of the model and propensity to recognize REs implied in drug-mediated induction was attempted. In panel B, the matrix calculated from the 11 half-site sequences of panel A is shown. p_i(A) through p_i(T), Probability of nucleotide at position *i* of the matrix. W_i, Weight of position *i*, *i.e.* degree of conservation at position *i* of the matrix.

dNR1 and dNR2. In the proximal promoter, another ER6, termed pNR, is located.

We analyzed 11.2 kb of the CYP3A7 promoter (GenBank accession no. AF329900) for DR3 and ER6 repeats (Fig. 2B). The functional dNR1 site was found as the best match (Z score 8.67). In contrast, the inactive dNR3 site, located 400 bp downstream from the XREM, attained a Z score of only 3.75. Searching for ER6-type REs, the dNR2 site was the second-best match, followed by the pNR site. Notably, the best-scoring ER6 element is located only 100 bp upstream from the dNR2 site, and it is possible, but untested so far, that this element contributes to the activation mediated by the (extended) XREM.

Because the underlying nucleotide distribution matrix defines what type of REs are recognized by the algorithm, it can be adapted to recognize REs for various receptors. To demonstrate this possibility of extension and adaptation, we have compiled a set of specific matrices for the α- or γ-subtypes of the PPAR from experimental data by Juge-Aubry *et al.* (12), weighted according to the ability of the hexamer sequences to bind either receptor isoform *in vitro*. As a test case for these matrices, we chose the rat glucose transporter 2 gene (GLUT2/SLC2A2). GLUT2 is a low-affinity facilitative glucose transporter, present in pancreatic β-cells, hepatocytes, as well as intestine and kidney epithelial cells. A role for GLUT2 as glucose sensor in pancreatic β-cells stimulating insulin release has been proposed: homozygous GLUT2 null mice show hyperglycemia and relative hypoinsulinemia (13). Furthermore, tro-

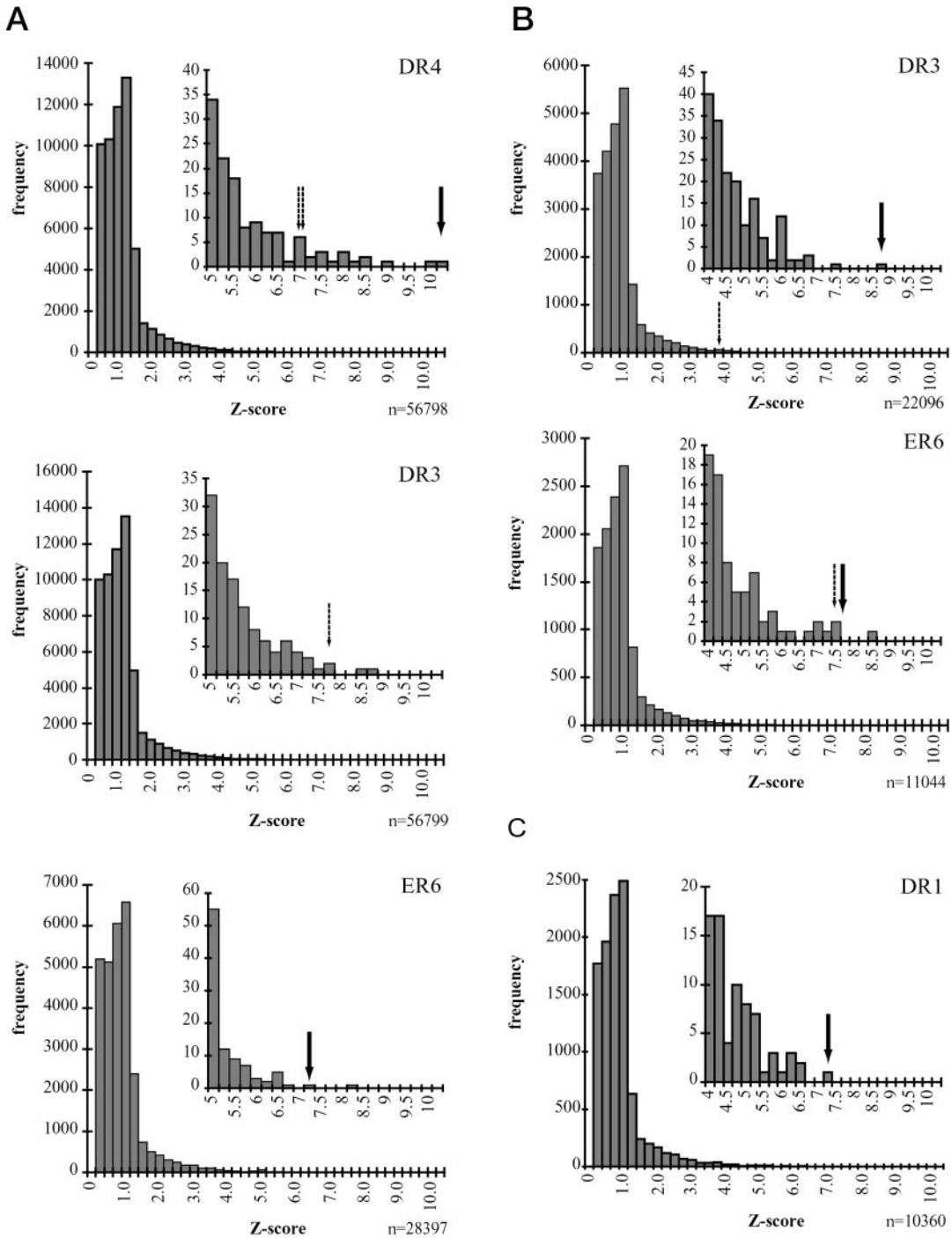

Fig. 2. Distribution of Scores of Predicted NR REs in the Flanking Regions of the Human MDR1 and CYP3A7 Genes and the Rat GLUT2 Gene

Bold arrows depict functional NR half-sites; *dashed arrows* mark half-sites initially suspected, but experimentally disproven in the original publications. A, 28,630 bp of MDR1 5′-flanking region sequence were analyzed with NUBIScan for matches to DR4, DR3, and ER6 NR binding sites. Threshold was set to 0, and the frequency of resulting Z scores was plotted. B, Analysis of 11.2-kb 5′-flanking region of the human CYP3A7 gene for occurrences of DR4 and ER6 type sites. N, Number of investigated possible REs. C, Analysis of the rat GLUT1 gene flanking region (5192 bp) for PPARγ sites (DR1). For this analysis, a weighted matrix was constructed from experimental data by Juge-Aubry *et al.* (12). Data were plotted as in panels A and B.

glitazone, an antidiabetic agent that increases glucose tolerance and insulin sensitivity, has previously been shown to be an activator of PPARγ and to increase GLUT2 protein levels (14, 15). Recently, a functional PPARγ RE was identified in the 5′-part of the GLUT2 gene (16).

We screened 5192 bp of the 5′-part of the GLUT2 gene (GenBank accession number: L28126) for DR1 elements, the cognate REs for PPAR. The functional PPARγ RE was the best match with a Z score of 7.19 when using a matrix specific for PPARγ (Fig. 2C), and with a Z score of 7.13 when using a matrix geared toward PPARα (not shown). This difference in match quality may be reflected in the activation potential of these two receptors: whereas PPARγ activated the RE efficiently in transactivation assays, PPARα activated the element to a much lesser extent (16).

These exemplary applications show that prediction of functional NR REs is possible in sequences as long as 30 kb. It must be emphasized that the Z score of a match is a measure of the resemblance to known functional sites but does not prove function, which depends on additional variables.

The Algorithm as a Predictor of Binding Sites

After proving that the NR REs predicted by our algorithm correlate well with functional binding sites in characterized regulatory regions, we applied NUBIScan to previously uncharacterized regulatory regions and tested the predictions.

We searched for REs in the recently characterized chicken CYP3A37 gene, the expression of which is increased by prototypical inducers of the CYP3A gene family *in ovo* and in LMH chicken hepatoma cells (17). From a chicken genomic library, we isolated a cosmid clone containing all CYP3A37 exons and flanking regions. We have subcloned and sequenced 3.1 kb of upstream sequence.[1] Because CYP3A37 is induced by a range of drugs and steroids that were shown to activate the chicken xenobiotic receptor CXR (6, 17), we surmised that CXR is the receptor mediating induction of this gene. Therefore, we scanned the upstream sequence with NUBIScan for DR4-type CXR REs, using the matrix described in Fig. 1. The best-scoring match (Z score, 8.72) was located at −1082 bp (Table 1). Concurrent with *in silico* analysis, the upstream region was subfractionated into smaller pieces, which were cloned into the pBLCAT5 chloramphenicol acetyl transferase (CAT) reporter gene vector. These constructs were tested for inducibility in LMH cells. The site conferring drug inducibility coincided with the predicted DR4-type binding site (Fig. 3A). Based on these results, we have defined a phenobarbital response unit (PBRU), able to confer drug induction, which is 159 bp in length, encompassing the predicted binding site (Fig. 3B). Site-directed mutagenesis of one of the predicted half-sites abolished drug induction completely. Mutation also abolished activation by CXR in transactivation assays (Fig. 3C) and prevented binding of CXR/RXR to the site in EMSAs (Fig. 4).

[1] These sequence data have been submitted to the DDBJ/EMBL/GenBank databases under accession no. AF486653.

Finally, we have applied our algorithm to the analysis of a drug-inducible 4.8-kb *Bam*HI fragment (GenBank accession no. AF236668) located upstream of the chicken CYP2H1 gene. From previous experiments, we could correlate the predictions of the algorithm with a comprehensive set of experimental data (18). Two PBRUs in this sequence have been shown to be activated by CXR, binding to cognate DR4 elements within the PBRUs (6, 19).

A search for DR4 repeats using the previously used matrix (see Fig. 1) yielded five predicted REs with a Z score higher than 6.5 (Table 1). The best-scoring result (Z score 8.92) is the DR4 element essential for drug-induced activation of the 264 bp, as shown by site-directed mutagenesis (18). While the second- and fourth-best results concern fragments unresponsive in reporter gene assays, the third-best match colocalizes with the DR4 described as necessary for CXR binding and drug-mediated induction in the 240-bp PBRU (19). The fifth predicted RE lies within the previously described 264-bp PBRU (Fig. 5A) and may explain the weak residual induction previously observed after mutation of the other DR4 element in this PBRU. To study the relative function of the two REs in more detail, we created mutants of the 264-bp PBRU. In addition to the previously described mutant of the 264-bp PBRU, where the NR1 site is mutated (18), we mutated both half-sites of the NR2 site in wild-type and ΔNR1 constructs, as depicted in Fig. 5B. We tested these constructs in reporter gene assays in LMH cells. Inducibility was measured after treating the cells for 24 h with 400 μM phenobarbital (PB), 500 μM glutethimide (GE), or vehicle alone (0.1% dimethylsulfoxide) (Fig. 5B). Mutation of the upstream DR4 element (ΔNR1) profoundly reduced drug inducibility conferred by the element (from 74.5-fold to 6.5-fold for GE) but did not abolish it completely. Only additional mutation of the downstream DR4 (ΔNR1ΔNR2) completely abolished induction, suggesting functional roles for both DR4 elements in the 264-bp PBRU. Interestingly, mutation of only the downstream RE (ΔNR2) led to very similar results as observed with mutation of the upstream element (ΔNR1). Inducibility was greatly reduced (from 74.5-fold to 4.3-fold for GE). This attributes important roles in induction to both REs in the 264-bp PBRU.

The interaction between CXR and the two DR4 elements was studied in transactivation assays in CV-1 monkey kidney cells. Cells were transfected with CXR expression plasmid in addition to the reporter constructs and exposed to 10 μM clotrimazole and other drugs as indicated. The effects of the mutations on transactivation by CXR were identical with the reporter gene results in LMH cells (Fig. 5C).

In EMSAs, we then determined whether CXR/RXR can still bind to the mutant PBRUs (Fig. 6). Radiolabeled wild-type probe could be efficiently displaced by cold wild-type competitor, but not with the ΔNR1 mutant. ΔNR2 competitor also displaced the wild-type probe quite efficiently, whereas the combined

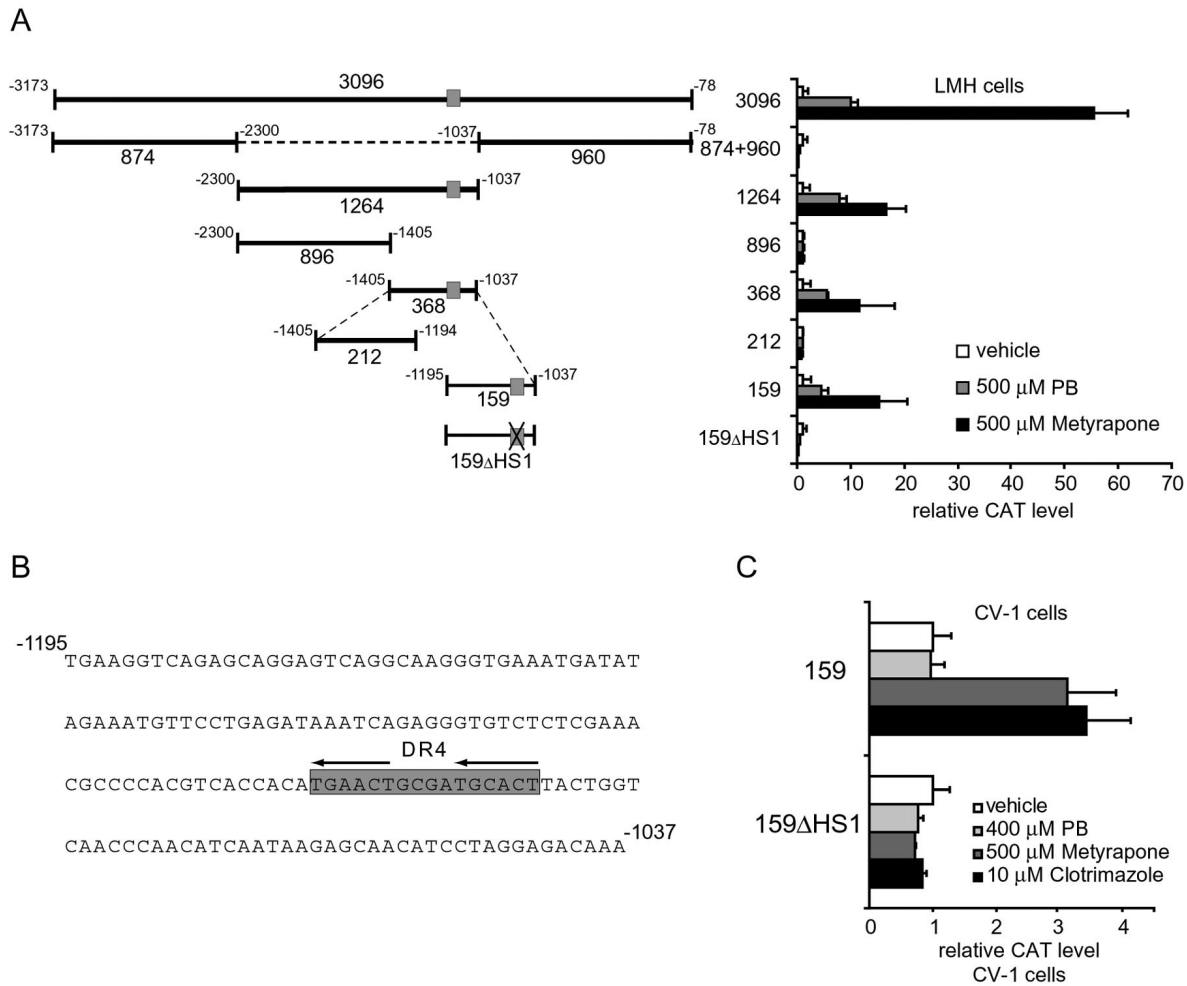

Fig. 3. Isolation of a Drug-Responsive Element from the CYP3A37 Gene

A, Subfragmentation of the drug-responsive 3.1-kb element. The *gray box* indicates the predicted DR4 site. Fragments were transiently transfected into LMH cells, and induction was measured after 24 h induction with the indicated compounds. In the 159ΔHS1 construct, the DR-4 element was destroyed by site-directed mutagenesis. B, Sequence of the 159-bp PBRU. The *gray box* indicates the DR4-type NR binding site, consisting of two hexamer half-sites designated by *arrows*. C, Wild-type and mutant 159-bp PBRUs were cotransfected with a CXR expression plasmid into *a priori* non-drug-responsive CV-1 cells. After transfection, cells were induced for 24 h with the indicated compounds. Data shown are the average of three independent experiments. *Error bars* represent 1 SD.

ΔNR1ΔNR2 mutant did not compete with the probe. When wild-type and mutant PBRUs were radiolabeled and used as probe, strong binding of CXR/RXR heterodimer was observed to the wild-type probe and to the ΔNR2 probe. ΔNR1 could still weakly bind the receptors, and with the ΔNR1ΔNR2 mutant, no binding at all was observed.

The NR1 and NR2 sites exhibit different affinities for CXR/RXR *in vitro*, NR2 interacting much more weakly with the receptors. However, both sites seem to contribute equally to the induction of the 264-bp PBRU, as witnessed in the experiments with single-site mutant constructs. Based on these data, we hypothesize that the observed synergistic effect is not likely to be caused at the level of DNA interaction, but happens later on, *e.g*. in the recruitment of coactivators. With the identification of a second DR4 element in the chicken CYP2H1 PBRU—in analogy to the PBRUs found in mouse, rat, and man (20)—the notion of evolutionary conserved drug-response signaling pathways put forward in Ref. 21 is corroborated.

In conclusion, application of the NUBIScan algorithm to this well characterized regulatory region thus has verified the usefulness of the algorithm in determining likely sites of regulation by NRs. Not unexpectedly, not all predicted sites mapped to distinctly drug-inducible fragments of the CYP2H1 flanking region. Recognition and transactivation by NRs depends on more than the core binding site, such as three-dimensional DNA structure, accessory binding sites, and surrounding sequences, factors not accounted for in the present model. Nevertheless, among the five best matches, the majority of the predictions were positive, and this refined *in*

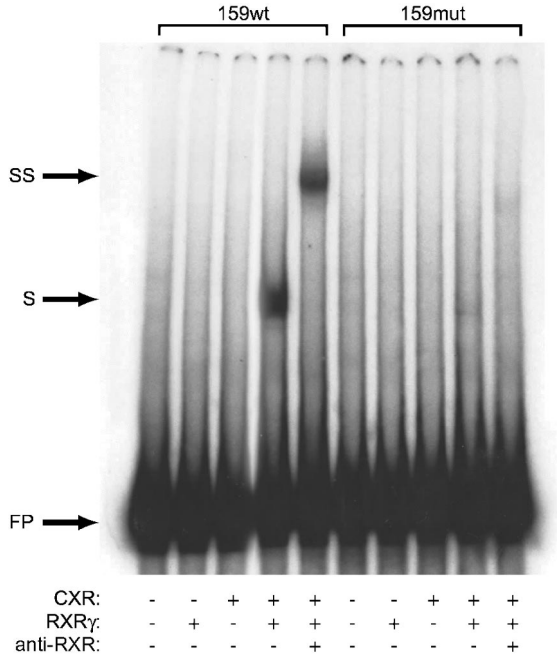

Fig. 4. CXR/RXR Binding to the CYP3A37 159-bp PBRU *in Vitro* Is Dependent on the Presence of the DR4 Motif

A wild-type 159-bp fragment or 159-bp fragment, where one of the two half-sites of the DR4 motif was altered to a *Not*I restriction site, was radiolabeled and incubated with *in vitro* transcribed/translated CXR, chicken RXRγ, or both. Binding of the heterodimer to the probe results in a band shift only with the wild-type fragment. When anti-RXR antibody was added, this complex was supershifted. *Arrows* designate free probe (FP), the complex of CXR, RXRγ, and probe (S), and the complex of CXR, RXRγ, anti-RXR antibody, and probe (SS).

silico analysis has led to the discovery of a further RE in the 264-bp PBRU, which had eluded previous studies and computer analyses.

Advantages and Limitations of the NUBIScan Algorithm

The present algorithm clearly provides more specificity than an approach based on matching consensus patterns. Already in generating a consensus pattern, much of the information about nucleotide frequency contained in an aligned set of RE sequences is discarded. When searching with a consensus, then, again information is lost, because only a match/no match decision can be made. In contrast, an approach using comparisons to a nucleotide frequency matrix conserves frequency information from the set of REs and furthermore, leads to differentiated quality scores for each match and thus retains more specificity (22). Presently available matrix-based approaches for the prediction of transcription factor-binding sites, such as the MatInspector algorithm (23), are designed to cover a large number of different protein-DNA interactions and do not consider all aspects of the internal structure of a particular binding site. The here described algorithm proved to be considerably more efficient and specific in the detection of functional NR REs.

Generally, detection of NR REs is hampered by the low degree of sequence conservation among REs, leading to low predictive power of the algorithm. As a result, the researcher is often overwhelmed with a wealth of putative REs, the quality scores of which are very similar. Increasing the specificity of predictions alleviates this problem.

Our approach focuses on a prediction of specific combinations of two low-affinity binding sites, such as NR binding sites. As a new concept, we consider each half-site of the RE as a motif and then search for a specific combination of these motifs. The net result of this approach of combining two matches to a matrix is that the mean score of all matches in a sequence drops dramatically, because the score of weak matches is decreased by the combination step, whereas the score of good matches (close to 1) remains nearly the same.

Using this combined scoring approach, a good half-site can compensate to some extent for a weaker half-site, an observation that was also made in experimental analysis of NR binding sites (*e.g*. in Ref. 18).

It is clear, however, that an *in silico* approach can never replace biological evidence for the functionality of a predicted RE. Transcriptional activation is a dynamic, multiprotein process, which depends on further factors such as chromatin structure; the sequences surrounding the NR binding site are important for its function. Because the algorithm focuses solely on the hexamer cores, predicted REs thus may not be functional in a native context. On the other hand, NRs have been shown to act also indirectly by cooperation with other proteins without contacting the DNA themselves. This level of regulation will be missed by our algorithm; it can only predict REs that involve direct receptor/DNA interaction.

Nevertheless, active REs were consistently found among the top-scoring predictions, and an analysis with NUBIScan is a time-efficient first approach to highlighting regions of interest for further experiments.

The estimate of false positive or false negative predictions is strongly dependent on the cut-off chosen for the Z score. With a threshold of $>$6.5 Z scores, we have not yet detected a false negative result, *i.e.* a site to which a receptor binds that is not detected. With the same Z score, we have seen false positives in the CYP2H1 flanking region, as discussed above.

Compared with three common approaches, visual inspection, pattern searches, and searches using the MatInspector algorithm (23), NUBIScan performs favorably: visual inspection of the sequence is restricted to short sequences. A pattern search approach based on degenerate consensus patterns will result in too many hits without any means to quantitatively distinguish good and poor matches. The MatInspector algorithm uses a database of weighted nucleotide matrices. Unfortunately, within the current release of the database, there are very

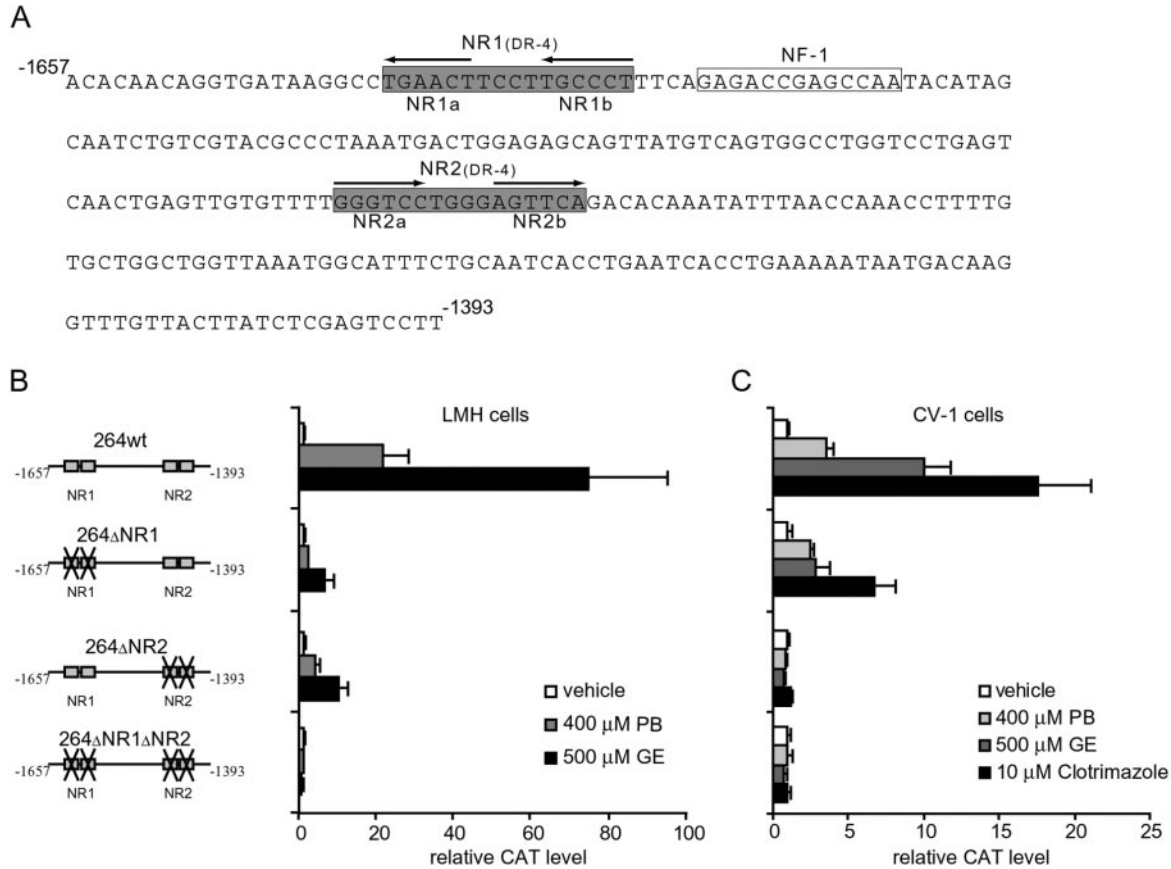

Fig. 5. Mutational Analysis of the Two Functional CXR Binding Sites in the 264-bp PBRU

A, Sequence of the 264-bp CYP2H1 PBRU. Indicated as *gray boxes* are the two DR4-type NR binding sites, each consisting of two hexamer half-sites designated by *arrows*. A putative NF-1 binding site is *boxed*. B, Effect of mutations on the inducibility of the 264-bp PBRU: the proximal or distal DR4 element or both were destroyed by site-directed mutagenesis. The inducibility of wild-type and mutation constructs was tested in LMH cells in transient transfection experiments. C, The same constructs were also cotransfected with CXR expression plasmid into CV-1 cells. After transfection, cells were induced for 24 h with 400 μM PB, 500 μM GE, or 10 μM clotrimazole. Data shown are the average of three independent experiments. *Error bars* represent 1 SD.

few matrices for NR REs. Moreover, analysis of large genomic sequences (>1000 bp) with MatInspector for single transcription factors is not recommended by the authors, as the predictive power of the algorithm is decreased by random matches that exceed the set threshold. This is prevented in NUBIScan by the multiplicative half-site combination step.

With this gain of specificity, achieved through the modular approach of combining distinct, but related, sites rather than using a larger, more variable region spanning both half-sites, screening large sets of genomic sequences for particular binding sites becomes feasible, where in our experience other programs fail to produce reliable results due to low specificity of the attained matches.

CONCLUSIONS

Responses to hormones and to environmental challenges are often mediated via enhancer elements that can be many kilobases away from the transcriptional start site and thus are difficult to detect and analyze. In this large-sequence context, "eyeballing" of the sequence is impractical or impossible, and supplementation by a computational approach is highly desirable. We have successfully applied the here described algorithm to such large genomic sequences. The algorithm can be adapted to specific recognition tasks by using a different matrix, on the condition that the RE consists of two distinct parts, as found with transcription factors binding as dimers.

Although the algorithm obviously cannot replace experimental verification of the proposed sites of regulation, it selects the set of sites most promising for further investigation. Sequence information of genes and their surrounding sequence is increasingly available for many genomes. A researcher interested in the regulation of a specific gene can apply the algorithm to mine this growing information base. This screening method for target genes of a particular regulatory pathway complements high-throughput approaches, such as gene expression arrays.

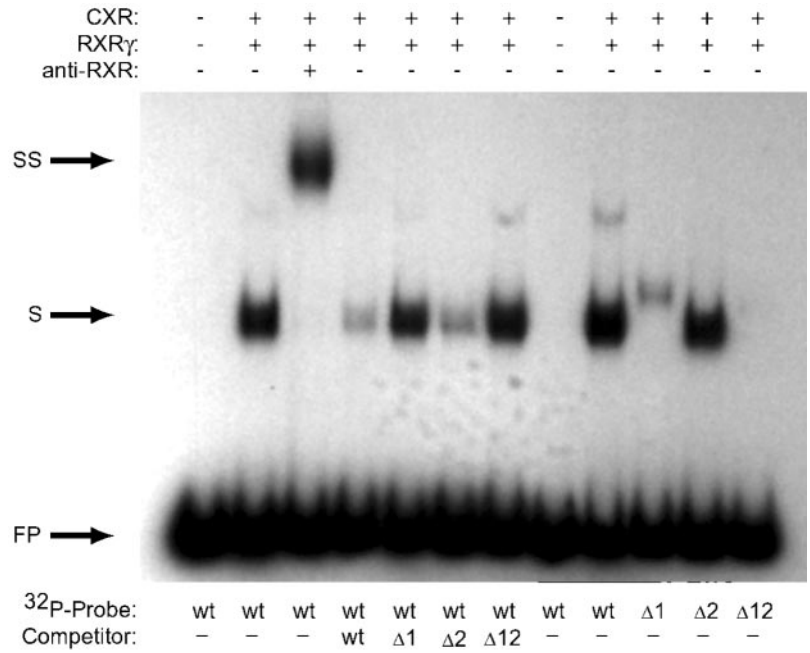

Fig. 6. Mutation of the NR Binding Sites Affects CXR/RXR Binding to the CYP2H1 264-bp PBRU *in Vitro*

Lanes 1–3, Wild-type 264-bp PBRU (wt) was radiolabeled and incubated with *in vitro* transcribed/translated CXR and RXRγ (lane 2) and additionally with anti-RXR antibody (lane 3). Lanes 4–7, Binding of CXR/RXR to the labeled wild type was competed with 20-fold molar excess of wild-type fragment or fragments mutated in the NR1 (Δ1), NR2 (Δ2), or both NR binding sites (Δ12). In lanes 9–12, wild-type and mutant fragments were incubated with CXR/RXR and binding of the receptors to the fragments was assayed. *Arrows* designate free probe (FP), the complex of CXR, RXRγ, and probe (S), and the complex of CXR, RXRγ, anti-RXR antibody, and probe (SS).

MATERIALS AND METHODS

Reagents

Unless specified otherwise, all chemical reagents were obtained from Fluka/Sigma-Aldrich Corp. (Buchs, Switzerland) and were of the highest grade available. Cell culture reagents were obtained from Life Technologies, Inc. (distributed by Invitrogen AG, Basel, Switzerland) unless stated otherwise. Plasmid preparations were done with the QIAGEN (Basel, Switzerland) system. The pBLCAT5 cloning vector was generously provided by Dr. G. Schütz (German Cancer Research Center, Heidelberg, Germany) and was described previously (24).

Matrix Calculation

Initially, a nucleotide distribution matrix is constructed from a set of aligned sequences, *e.g*. from a group of NR half-sites. Subsequently, a position weight is calculated for each position of the matrix representing conservation of bases at that position

$$W_i = \frac{100}{\ln 4}\left(\sum_{b=A\to T} \underbrace{p_i(b)*\ln p_i(b)}_{0,\ \text{if } p_i(b)=0} + \ln 4\right)$$

Equation 1: definition of matrix position weight according to information content. W_i, weight at position i; b, base; $p_i(b)$, frequency of base b at position i (taken from stochastic matrix).

This position weight is arbitrarily scaled between 0 and 100, where 100 signifies a position with full sequence conservation throughout the training set and 0 signifies no conservation. Due to the positional weights of the matrix, more conserved nucleotides are considered more important in the analysis of a query sequence.

Scanning a Query Sequence for Potential NR Binding Sites

To localize specific NR binding sites, NUBIScan searches for two occurrences of the matrix within a specified distance, having a defined relative orientation (DR, ER, or IR). These parameters are defined before execution of the algorithm. The comparison is a two-step process: initially, single scores, *i.e.* quality scores for all possible matches on both strands of the query sequence to the matrix, are calculated. Match quality scores are expressed as ratio to the best attainable match.

$$S_{halfsite}(j) = \frac{\sum_{i=1}^{n} W_i p_i(b_{j+i})}{\sum_{i=1}^{n} W_i p_i(\max)}$$

Equation 2: calculation of the half-site score at position j of the query sequence. n, Length of matrix; W_i, matrix weight at position i; b_{j+i}, base at offset i from starting point of present calculation (j); $p_i(b_{j+i})$, nucleotide score; $p_i(\max)$, maximal nucleotide frequency found in the matrix at position i.

Finally, matrix matches are combined by multiplication to give the score for the desired arrangement and spacing of half-sites. Thus, $S_{Full}(j) = S_{half\text{-}site}(j) * S_{half\text{-}site}(j+n)$ for position j in the query sequence and a DR with n spacer nucleotides. For ERs and IRs, scores from the sense and antisense strand are combined analogously. Positions whose score is over a predefined threshold value are included in a list of found matches. This threshold can be set either as a final score or as a Z score (*i.e.* a number of standard deviations from the mean score of all possible sites in the sequence).

The core algorithm has been programmed in C as a command-line tool for Microsoft Corp. Windows and Unix plat-

forms. A web interface for the algorithm was programmed in Perl.

The algorithm is shown in detail in the supplemental data to this paper published on The Endocrine Society's Journals Online web site, http://mend.endojournals.org/. In addition, this algorithm is the core of a web interface that functions as a central hub, providing background information and user instructions as well as access to the programs. This web interface is available to researchers from nonprofit and academic institutions at the following URL: http://www.nubiscan.unibas.ch. It allows users to create their own matrix files from sets of sequences and to scan query sequences with their own matrices or with provided general-purpose matrices.

Users can define a search strategy for their query sequence, so that a sequence can be scanned for different arrangements of half-sites at the same time. In this way, variations in the admissible half-site spacing for a given receptor can be accommodated. For each single search, a threshold can be set either as absolute score or as Z score. As a starting point, we recommend setting the threshold at six to seven Z scores, and to examine the resulting predictions subsequently in rank order.

Cloning and Mutagenesis of REs

CAT reporter gene plasmids containing subfragments of the CYP3A37 5′-flanking region were constructed by PCR and subsequent restriction digestion (see Fig. 3). The 3096-bp fragment was amplified by PCR using 5′-ATC GGA TCC AGC TGG GTG TAG GGT CCA T-3′ and 5′-ATC GGA TCC ACT GGC CTC ATG TCC CGA-3′ as sense and antisense primers, creating a *Bam*HI site on both ends to facilitate cloning.

From the 3096-bp fragment, the 1264-bp fragment was excised using *Eco*NI and *Nsi*I, religation resulting in the 874 bp + 960 bp construct. The 1264-bp fragment was further cut with *Aat*II, resulting in the 896-bp and 368-bp fragment. Finally, digestion with *Bsp*MI yielded the 212-bp and 159-bp fragments. The DR4 site of the 159-bp PBRU was mutated into a *Not*I restriction site by PCR using standard overlap techniques as described previously (18). Thus, the DR4 was altered from the wild-type 5′-**TGAACT**GCGA**TGCACT**-3′ sequence to 5′-**gcggcc**GCGA**TGCACT**-3 (altered bases are in *lowercase letters*, the hexamer cores are represented in *boldface*). CAT reporter gene plasmids containing the wild-type 264-bp CYP2H1 PBRU or the 264-bp CYP2H1 PBRU with the NR1 half-sites mutated were described (18). Briefly, the NR1 site was changed from 5′-**GAACT**TCCT**TGCCCT**-3′ to 5′-**ccgcgg**TCCT**gatatc**-3′, introducing *Sac*II and *Eco*RV restriction sites. Mutants of the second NR site were done with PCR overlap techniques. By changing the wild-type 5′-**GGGTCC**TGGG**AGTTCA**-3′ sequence to 5′-**tctaga**TGGG-**ctcgag**-3′, the half-sites were mutated into an *Xba*I and a *Xho*I restriction site.

Cell Culture and Transient Transfection, EMSA

LMH chicken hepatoma cells were cultivated as previously described (6). One day before transfection, the culture medium was replaced with William's E medium containing glutamine and penicillin/streptomycin, supplemented with 10% delipidated, charcoal-stripped FCS (Sigma, St. Louis, MO). This change was observed to greatly increase the extent of drug-mediated induction, presumably by the absence of inhibitory substances from the serum. Cells were transfected with 1 μg of reporter vector and 100 ng of pRSV-β-Galactosidase (kindly provided by Dr. A. Kralli, Biozentrum, University of Basel, Switzerland) vector per well, using FuGene6 transfection reagent (Roche Molecular Biochemicals, Rotkreuz, Switzerland) according to the manufacturer's instructions. Four hours after transfection, an appropriate dilution of the inducer compounds was added to the wells, and cells were incubated for an additional 24 h before preparation of cell extracts to assay for reporter gene and β-galactosidase expression. Transactivation assays in CV-1 monkey kidney cells, reporter gene assays, and EMSAs were performed as described before (6). Expression plasmids for CXR and chicken RXRγ have been described (6). The monoclonal antimouse RXR rabbit antibody used for supershifts was kindly provided by Dr. P. Chambon (Institut de Génétique et de Biologie Moléculaire et Cellulaire, Université Louis Pasteur, Illkirch, France).

Acknowledgments

M.P. would like to thank Joseph Pelrine, guru-level developer, for priceless discussions about the design of efficient and elegant computer algorithms.

Received December 26, 2001. Accepted February 13, 2002.

Address all correspondence and requests for reprints to: Michael Podvinec, Division of Pharmacology/Neurobiology, Biozentrum of the University of Basel, Klingelbergstrasse 70, CH-4056 Basel, Switzerland. E-mail: michael.podvinec@unibas.ch.

Address requests for reprints to: Urs A. Meyer, Division of Pharmacology/Neurobiology, Biozentrum of the University of Basel, Klingelbergstrasse 70, CH-4056 Basel, Switzerland. E-mail: urs-a.meyer@unibas.ch.

This work was supported by the Swiss National Science Foundation.

REFERENCES

1. Mangelsdorf DJ, Thummel C, Beato M, Herrlich P, Schutz G, Umesono K, Blumberg B, Kastner P, Mark M, Chambon P 1995 The nuclear receptor superfamily: the second decade. Cell 83:835–839
2. Steinmetz AC, Renaud JP, Moras D 2001 Binding of ligands and activation of transcription by nuclear receptors. Annu Rev Biophys Biomol Struct 30:329–359
3. McKenna NJ, Xu J, Nawaz Z, Tsai SY, Tsai MJ, O'Malley BW 1999 Nuclear receptor coactivators: multiple enzymes, multiple complexes, multiple functions. J Steroid Biochem Mol Biol 69:3–12
4. Collingwood TN, Urnov FD, Wolffe AP 1999 Nuclear receptors: coactivators, corepressors and chromatin remodeling in the control of transcription. J Mol Endocrinol 23:255–275
5. Waxman DJ 1999 P450 gene induction by structurally diverse xenochemicals: central role of nuclear receptors CAR, PXR, and PPAR. Arch Biochem Biophys 369:11–23
6. Handschin C, Podvinec M, Meyer UA 2000 CXR, a chicken xenobiotic-sensing orphan nuclear receptor, is related to both mammalian pregnane X receptor (PXR) and constitutive androstane receptor (CAR). Proc Natl Acad Sci USA 97:10769–10774
7. Xie W, Barwick JL, Simon CM, Pierce AM, Safe S, Blumberg B, Guzelian PS, Evans RM 2000 Reciprocal activation of xenobiotic response genes by nuclear receptors SXR/PXR and CAR. Genes Dev 14:3014–3023
8. Schuetz EG, Beck WT, Schuetz JD 1996 Modulators and substrates of P-glycoprotein and cytochrome P4503A coordinately up-regulate these proteins in human colon carcinoma cells. Mol Pharmacol 49:311–318
9. Geick A, Eichelbaum M, Burk O 2001 Nuclear receptor response elements mediate induction of intestinal MDR1 by rifampin. J Biol Chem 276:14581–14587

10. Bertilsson G, Berkenstam A, Blomquist P 2001 Functionally conserved xenobiotic responsive enhancer in cytochrome P450 3A7. Biochem Biophys Res Commun 280: 139–144
11. Pascussi JM, Jounaidi Y, Drocourt L, Domergue J, Balabaud C, Maurel P, Vilarem MJ 1999 Evidence for the presence of a functional pregnane X receptor response element in the CYP3A7 promoter gene. Biochem Biophys Res Commun 260:377–381
12. Juge-Aubry C, Pernin A, Favez T, Burger AG, Wahli W, Meier CA, Desvergne B 1997 DNA binding properties of peroxisome proliferator-activated receptor subtypes on various natural peroxisome proliferator response elements. Importance of the 5′-flanking region. J Biol Chem 272:25252–25259
13. Guillam MT, Hummler E, Schaerer E, Yeh JI, Birnbaum MJ, Beermann F, Schmidt A, Deriaz N, Thorens B, Wu JY 1997 Early diabetes and abnormal postnatal pancreatic islet development in mice lacking Glut-2. Nat Genet 17: 327–330
14. Lehmann JM, Moore LB, Smith-Oliver TA, Wilkison WO, Willson TM, Kliewer SA 1995 An antidiabetic thiazolidinedione is a high affinity ligand for peroxisome proliferator-activated receptor γ (PPAR γ). J Biol Chem 270: 12953–12956
15. Higa M, Zhou YT, Ravazzola M, Baetens D, Orci L, Unger RH 1999 Troglitazone prevents mitochondrial alterations, β cell destruction, and diabetes in obese prediabetic rats. Proc Natl Acad Sci USA 96:11513–11518
16. Kim HI, Kim JW, Kim SH, Cha JY, Kim KS, Ahn YH 2000 Identification and functional characterization of the peroxisomal proliferator response element in rat GLUT2 promoter. Diabetes 49:1517–1524
17. Ourlin JC, Baader M, Fraser D, Halpert JR, Meyer UA 2000 Cloning and functional expression of a first inducible avian cytochrome P450 of the CYP3A subfamily (CYP3A37). Arch Biochem Biophys 373:375–384
18. Handschin C, Meyer UA 2000 A conserved nuclear receptor consensus sequence (DR-4) mediates transcriptional activation of the chicken CYP2H1 gene by phenobarbital in a hepatoma cell line. J Biol Chem 275: 13362–13369
19. Handschin C, Podvinec M, Looser R, Amherd R, Meyer UA 2001 Multiple enhancer units mediate drug-induction of CYP2H1 by the xenobiotic-sensing orphan nuclear receptor CXR. Mol Pharmacol 60:681–689
20. Sueyoshi T, Negishi M 2001 Phenobarbital response elements of cytochrome P450 genes and nuclear receptors. Annu Rev Pharmacol Toxicol 41:123–143
21. Handschin C, Podvinec M, Stockli J, Hoffmann K, Meyer UA 2001 Conservation of signaling pathways of xenobiotic-sensing orphan nuclear receptors, chicken xenobiotic receptor, constitutive androstane receptor, and pregnane X receptor, from birds to humans. Mol Endocrinol 15:1571–1585
22. Lavorgna G, Boncinelli E, Wagner A, Werner T 1998 Detection of potential target genes in silico? Trends Genet 14:375–376
23. Quandt K, Frech K, Karas H, Wingender E, Werner T 1995 MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res 23:4878–4884
24. Boshart M, Kluppel M, Schmidt A, Schutz G, Luckow B 1992 Reporter constructs with low background activity utilizing the cat gene. Gene 110:129–130
25. Sueyoshi T, Kawamoto T, Zelko I, Honkakoski P, Negishi M 1999 The repressed nuclear receptor CAR responds to phenobarbital in activating the human CYP2B6 gene. J Biol Chem 274:6043–6046
26. Honkakoski P, Negishi M 1997 Characterization of a phenobarbital-responsive enhancer module in mouse P450 Cyp2b10 gene. J Biol Chem 272:14943–14949
27. Trottier E, Belzil A, Stoltz C, Anderson A 1995 Localization of a phenobarbital-responsive element (PBRE) in the 5′-flanking region of the rat CYP2B2 gene. Gene 158: 263–268
28. Stunnenberg HG 1993 Mechanisms of transactivation by retinoic acid receptors. Bioessays 15:309–315
29. Dreyer C, Krey G, Keller H, Givel F, Helftenbein G, Wahli W 1992 Control of the peroxisomal β-oxidation pathway by a novel family of nuclear hormone receptors. Cell 68:879–887
30. Muerhoff AS, Griffin KJ, Johnson EF 1992 The peroxisome proliferator-activated receptor mediates the induction of CYP4A6, a cytochrome P450 fatty acid ω-hydroxylase, by clofibric acid. J Biol Chem 267:19051–19053