Defining sequence elements that regulate *CYP450* gene expression

P.K. Rogan
Schools of Medicine and Computing & Engineering,
University of Missouri-Kansas City, 64108 USA

Summary

Many nuclear receptor transcription factors regulate genes by recognizing half sites separated by variable length sequences. Bipartite sites for HNF4α, CAR/RXRα, VDR/RXRα, and PXR/RXRα were modeled by information theory, a method that comprehensively captures sequence conservation and predicts binding site affinities. Promoters of Both known and previously unrecognized binding sites were found by scanning PXR/RXRα target genes with this model. Genome-wide scans with this model also predicted binding sites in the promoters of 19 novel target genes, which were subsequently validated *in vitro*. Expression studies of PXR-transfected HepG2 cells confirmed that rifampin treatment produced transcriptional responses in 13 of 17 predicted gene targets.

Introduction

Models of nucleic acid binding sites based on information theory accurately predict affinities of interactions with regulatory proteins because individual information content (R_i, in bits [bt]) of a binding site is related to the enthalpy of its interaction with the protein that recognizes it (Schneider 1997; Bi and Rogan 2005). The optimal sequence pattern, represented as an information weight matrix, is obtained by minimizing entropy (or maximizing information content) in either single block or bipartite sequence alignments (Bi and Rogan 2004). The computed R_i values of sites are related to experimentally-measured binding affinities (Schneider 1997). Many of transcription factor information weight matrices resemble consensus sequences due to experimental detection of strong sequence elements with the largest transcriptional effects. This bias can be mitigated with experimental (Vyhlidal, Rogan et al. 2004) and computational (Rogan et al. 2003) refinement procedures that produce models based on sites which span the natural range of binding affinities.

We present binding site models for promoter sequence elements recognized by human nuclear receptor factors that regulate CYP450 genes. We then used models of PXR/RXRα sites to predict and validate novel binding sites and gene targets for this factor.

Materials and Methods

Model development. Binding sites recognized by HNF4α (Sladek and Seidel, 2001), CAR/RXRα (cited in Handschin and Meyer 2003), PXR/RXRα (Vyhlidal, Rogan et al. 2004) or VDR/RXRα (cited in Bi and Rogan 2004) were used to derive bipartite information weight matrices by entropy minimization-based alignment (Bi and Rogan 2004).

Bipartite modeling and threshold determination. A bipartite information module [**M**] consists of left and right motif R_i values, and the associated gap surprisal function, g(d) defined as $-log\ (n(d) / n)$, where $n(d)$ is a function of the distance separating the motifs, d. The total information content, ($R_{i,total}$; Bi and Rogan, 2004) is:

$$R_{i,\,total} = R_i(left \mid d) + R_i(right \mid d) - g(d)$$

The minimum threshold value of $R_{i,total}$ was determined by embedding true sites in a noisy background sequence. $R_{i,total}$ is calculated for each site within the set of background DNA sequences, that are generated by Monte Carlo simulation based on a uniform multinomial distribution, p_0. The embedded sites derived from a matrix of known binding sites are assumed to have $R_{i,total}$ values normally distributed around $R_{sequence}$. A conditional distribution $P(S \mid R_i > 0, \mathbf{M})$ is determined for both chance and embedded bipartite sequences. Binding cutoff values (C_b for $R_{i,total}$, C_l and C_r for left and right-half site R_i values) are determined by minimizing Type I and II classification errors. True bipartite sites are subject to $R_{i,total} > C_b$ and R_i (left) $> C_l$ and R_i (right) $> C_r$.

Genome scanning. The human genome sequence (Build 31) was scanned with refined PXR/RXRα information weight matrices (Gadiraju et al 2003). Gene promoters predicted to contain binding sites were sorted by the respective R_i values of those sites. Candidate target genes were futher evaluated, based on the presence of at least one strong site or clusters of ≥2 sites with $R_i > R_{sequence}$.

Binding site assay. Expression plasmids containing the cDNA for hRXRα (pSG5-hRXRα) and hPXR (pSG5-hPXRΔATG) (Lehmann et al 1998) were transcribed and translated *in vitro* (Promega). Reference oligonucleotides [1 uM] were covalently linked to the surface of multi-well strips (Trimgen) and bound to synthetic PXR/RXRα in the presence or absence of a test oligo. The bound complex was detected with primary anti-RXRα antibody (Santa Cruz Biologicals) followed by secondary antibody conjugated to horseradish peroxidase on a Biomek 2000 workstation (Beckman Instruments). Oligo binding was validated in duplicate by demonstrating concentration-dependent competition with solid phase reference oligo for protein binding. Data were normalized

against "no-target" binding and "no-competitor" binding site controls and then averaged.

Quantitative PCR. HepG2 cells (ATCC) were stably transfected either with linearized pSG5-hPXRΔATG and pCI-neo (Promega**)** or linearized pCI-neo alone with FuGENE6 reagent (Roche), then characterized for PXR copy number and expression by quantitative RT-PCR. After antibiotic washout, pairs of independent cell lines were treated with either 10 uM rifampin (pPXR +pCI-neo) or DMSO (pCI-neo) and harvested at 0, 0.5, 2, 6, 24, 48 hours post-treatment. Intron-spanning primers were designed and optimized for each predicted and known gene target, which was evaluated for response to rifampin treatment. RNA isolated from these lines was analyzed by qRT-PCR in triplicate with Syber Green detection (Qiagen); $\Delta\Delta C_t$ values were computed relative to corresponding β-actin internal control expression for each time point (RNA levels of β-actin were unchanged in response to rifampin; not shown).

Results

Bipartite models of nuclear receptor binding sites. Many multimeric transcription factors recognize DNA sequence patterns by cooperatively binding to elements composed of half sites separated by a nonspecific sequence of variable length. The Bipartite pattern discovery algorithm searches for the optimal alignment, allowing for all half-site orientations (Bi and Rogan 2004). Models of VDR, PXR, CAR and HNF4α binding sites were built, based on published and, for PXR, experimentally refined binding sites. The respective sequence logos, which depict conservation and average information content (R_{seq}) are shown in Figure 1.

HNF4α binds as a homodimer, consistent with the symmetric pattern evident in the sequence logo (R_{seq} = 10.99 bt) . Information maxima are separated by one helical turn (10 nucleotides), consistent with recognition across the major groove. For VDR, the optimal pattern was composed of 7 nucleotide left- and right-half sites (7<[0,6]>7), with R_{seq} values of 6.76 and 6.57 bt, respectively separated by ≤ 6 nucleotide gap. The bipartite model R_{seq} value exceeds the best one-block motif by 2.74 bt, indicating that it more comprehensively depicts sequence conservation. For CAR, all half site orientations exhibit similar R_{seq} values (13.87 bt for direct repeat, 13.93 bt for reverse direct repeat, 13.81 bt for everted repeats and 13.96 bt for inverted repeats), suggesting that CAR may be capable of recognizing half-site pairs in any orientation. However, the 6<0,8>7 RDR model had the largest information increment and is the only model with an asymmetric pattern.

The optimal PXR motif was selected by maximizing incremental information, and by determining which model best fit the apparent binding site affinities. Because weak sites with low R_i values are not always detectable experimentally, the site boundaries, ie. widths, were determined by maximizing incremental information for a series of models. The best single-block model is 17-bp wide (Vyhlidal, Rogan et al. 2004), whereas the optimal bipartite model was 7<4>7. Minimum binding site strength thresholds were then defined by Monte Carlo simulation. The $R_{i,total}$ threshold was determined to be 8.3 bt, with the right half-site being more conserved than the left (4.8 vs. 3.7 bt). Like VDR, the PXR bipartite model captures significantly more sequence conservation than the single block model.

Promoter scanning. The promoters of several PXR target genes were scanned with the PXR bipartite model and the results were compared with previous single block model, which had been used to confirm previously identified PXR enhancers (PXREs) and identified Caspase 10 as a novel PXR target gene (Fig. 2; Vyhlidal, Rogan et al 2004). Sites found with this model both recapitulate results obtained with the single block model (Vyhlidal, Rogan et al. 2004) and include novel strong sites within the *CYP3A4*, *UGT1A6*, and *CASP10* promoters (Fig. 2, arrows).

Identification of sequences recognized by PXR/RXRα and potential transcriptional targets. Genome-wide promoter scans (Gadiraju et al. 2003) detected previously unrecognized binding sites in genes from the *CYP* family and other prospective PXR targets. Binding site predictions were tested (for sites with $R_i > R_{seq}$) using a multiwell protein-DNA binding site assay. The estimated affinities of PXR/RXRα for competitor and reference binding sites are matched based on selecting sequences with similar R_i values (increasing the likelihood that binding of the test oligonucleotide will be detectable). Using the proximal PXR enhancer of *CYP3A4* as a reference sequence, binding was verified for sites in the promoters of the *SCP2* (21 bt), *CRYZ* (20 bt), *PFKFB4* (19 bt), *SMN2/SMN1* (19 bt), *CYP3A7* (18 bt), *CUBN* (21 bt), *ANAPC5* (20 bt), *NUFIP1/ LSR7* (19 bt), *USP9X* (22 bt), *UGT1A3* (19 bt), *CBFA2T1* (19 bt), *DOCK4* (19 bt), *NAPB* (19 bt), *MAOB* (15 bt), M69296 (19 bt), *ALDH1A1* (14 bt), *SRP* (14 bt), *CPO* (11 bt), and the *PFKFB46* (13 bt) genes.

Changes in transcriptional response to the PXR-ligand rifampin were studied in HepG2 hepatocyte-derived cells that had been stably transfected with PXR. A time course expression study was carried out by quantitative reverse-transcription followed by PCR for most of these loci (Fig. 3). Of the 17 genes analyzed, the response was significant (>2 fold) for 7 genes,

was modest (1.4 to 2 fold) for 6 genes, and small or insignificant changes were evident for 4 of the genes (<1.4 fold). These *in silico* predictions enrich for genes regulated by PXR/RXRα. Using the same criteria, significant changes in transcription for 1564 of 12,626 Unigene clusters (12.4%) were seen in microarray expression studies of rifampin-treated HepG2 cells, 1169 (9.2%) had a modest response, and 9893 (78.4%) were essentially unchanged. The lack of response for some predicted PXR targets could be due to information model error, separation from other critical promoter elements, eg. coactivator binding sites, or the absence of expression of these genes in this cell line. Binding sites validated for *DOCK4* and *LSR7* >20 kb upstream from the transcription initiation sites of these genes, which may limit their impact on transcription. *CRYZ*, which encodes lens crystalline, has low level expression in liver, and *USP9X* is not expressed in HepG2 cells (Su et al. 2002).

Conclusions

Published and experimentally-refined binding sites were used to construct bipartite information models of nuclear receptor binding sites. These models tangibly improve upon single block information models in detecting predicted PXR enhancer elements in promoters of established and in novel gene targets. Expression studies confirm that this *in silico* approach can identify PXR targets which contain one or more strong binding sites in their promoters.

Acknowledgements

We gratefully acknowledge Dr. Erin Schuetz for providing microarray data on HepG2 expression. Research supported by the Katharine B. Richardson Foundation and PHS ES10855-02.

References

BI CP and ROGAN PK. Bipartite pattern discovery by entropy minimization-based multiple local alignment. *Nucl. Acids Res.,* **32**, 4979-4991, 2004.

BI CP and ROGAN PK. Information theory as a model for genomic sequences. In: **Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics**, NY, Wiley, 2005.

GADIRAJU S, VYHLIDAL CA, LEEDER JS, and ROGAN PK. Genome-wide prediction, display and refinement of binding sites with information theory-based models *BMC Bioinformatics*, **4**:38, 2003.

HANDSCHIN C and MEYER UA Induction of drug metabolism: The role of nuclear receptors. *Pharmacology Review*, **55,** 649-673, 2003.

LEHMANN, J. M., MCKEE, D. D., WATSON, M. A., WILLSON, T. M., MOORE, J. T., AND KLIEWER, S. A. *J. Clin. Investig.* **102,** 1016–1023, 1998.

SCHNEIDER TD, Information content of individual genetic sequences. *J. Theor. Biol.*, **189**, 427-441, 1997.

SLADEK FM and SEIDEL SD. Hepatocyte Nuclear Receptor 4α. Chapter 9, In: **Nuclear Receptors and Genetic Disease**, NY, Academic Press, 2001.

SU AI, COOKE MP, CHING KA, HAKAK Y, WALKER JR, WILTSHIRE T, ORTH AP, VEGA RG, SAPINOSO LM, MOQRICH A *et al*. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci USA* **99**, 4465-70, 2002.

VYHLIDAL CA, ROGAN PK, and LEEDER JS. Development and refinement of pregnane X receptor DNA binding site model using information theory: Insights into PXR mediated gene regulation *J. Biol. Chem*., **279**, 46779-46786, 2004.

Figure Legends

Figure 1. Sequence logos of (A) HNF4α 6<0,1>6; (B) VDR/RXRα 7<0,5>7; (C) PXR/RXRα 7<0,6>7; and (D) CAR/RXRα 7<0,8>7 bipartite information models.

Figure 2. Predicted strengths of 7<3>7 bipartite PXR/RXRα binding sites in target gene promoters. Panels show information scans of 10 kb sequence upstream of the (a) *CYP3A4* (b) *UGT1A3* (c), *UGT1A6* and (d) *CASP10* genes. Predicted novel strong PXR/RXRα binding sites are indicated with arrows. Dotted lines indicate R_i threshold levels for background sites. Vertical bars indicate R_i values for specific sequence locations. Inverted bars are sites on the antisense strand. The dotted line indicates the $R_{i,total}$ threshold value of 8.3 bt.

Figure 3. Gene expression changes in response to rifampin after >48 hrs for predicted PXR/RXRα gene targets. Genes selected for analysis contained at least one strong site (≥18 bt) and typically multiple promoter sites with $R_i > R_{seq}$. Changes were defined as significant: $\Delta\Delta C_t \geq 1.0$, slight: $0.5 \leq \Delta\Delta C_t \leq 1.0$, or unchanged: $0.5 \leq \Delta\Delta C_t$.

Fig 1

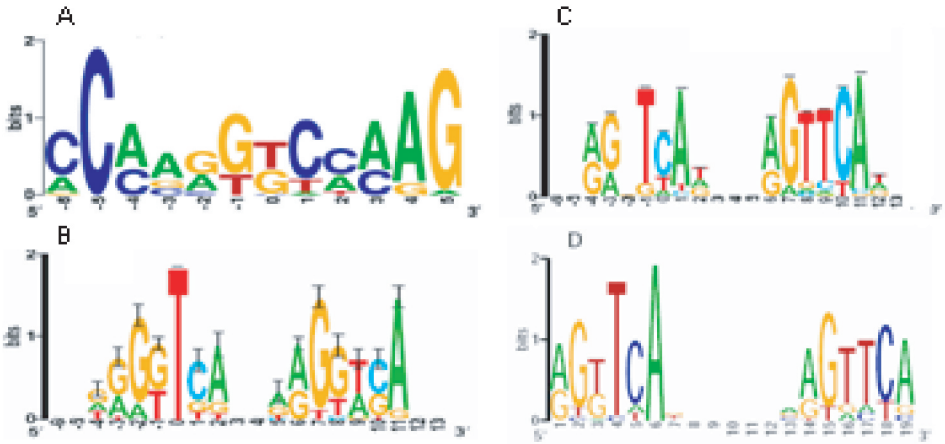

Fig 2

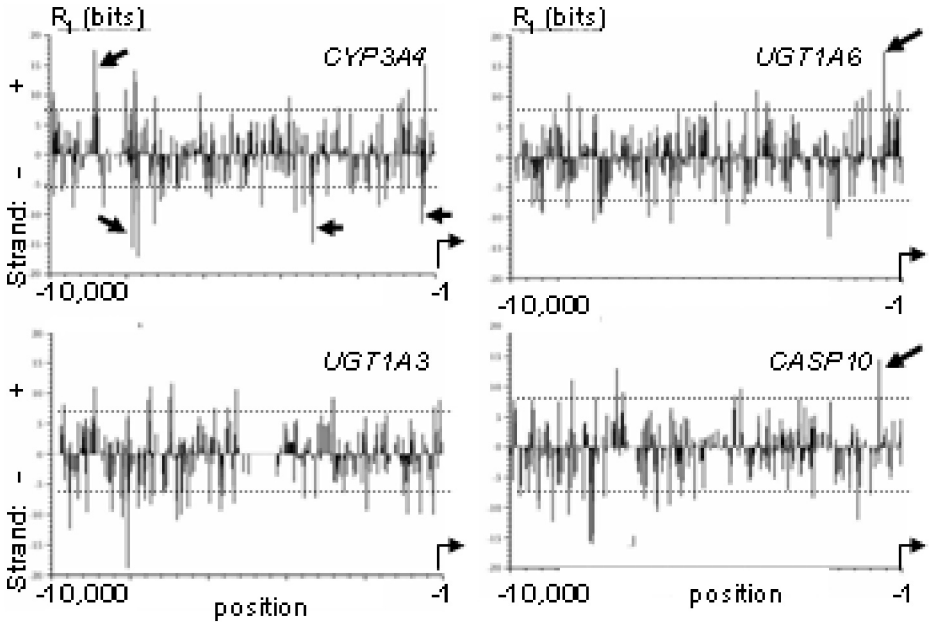

Fig 3

Gene	Response
CASP10	Induction
SRP	Induction
NOG	Induction
SMN1	Slight induction
SCP2	Slight induction
CLPN3	Repression
CUBN	Repression
NAPB	Repression
PFKFB4	Repression
ALDH1A1	Slight repression
ANAPC5	Slight repression
MAOB	Slight repression
NUFIP1	Slight repression
DOCK4	Unchanged
LSR7	Unchanged
USP9X	Unchanged
CRYZ	Unchanged