Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
Molecular Pharmacology Fast Forward. Published on September 1, 2004 as doi:10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

MOL Manuscript # 1677R

A novel computational approach for the prediction of networked transcription factors of Ah-receptor regulated genes.

Alexander Kel , Susanne Reymann, Volker Matys, Paul Nettesheim, Edgar Wingender and Jürgen Borlak

BIOBASE GmbH, Halchtersche Str. 33, D-38304 Wolfenbüttel, Germany (A.K., V.M., E.W.); Institute of Cytology and Genetics, 630090, pr. Lavrentyev-10, Novosibirsk, Russia (A.K.); Fraunhofer Institute (Fh-ITEM) of Toxicology and Experimental Medicine, Center for Drug Research and Medical Biotechnology, Nikolai-Fuchs-Str. 1, D-30625 Hannover, Germany (S.R., J.B.); National Institute of Environmental Health Sciences (NIEHS), Research Triangle Park, NC 27709, USA (P.N.); Center of Pharmacology and Toxicology, Medical School of Hannover, Carl-Neuberg-Str. 1, 30625 Hannover, Germany (J.B.)

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Copyright 2004 by the American Society for Pharmacology and Experimental Therapeutics.

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

MOL Manuscript # 1677R

Running title: In silico genomic footprinting

Corresponding author:

Univ.-Prof. Dr. Jürgen Borlak

Fraunhofer Institute of Toxicology and Experimental Medicine

Center for Drug Research and Medical Biotechnology

Nikolai-Fuchs-Str. 1

D-30625 Hannover

Phone: +49 (0) 5 11/ 53 50 -559

Fax: +49 (0) 5 11/ 53 50 -573

e-mail: borlak@item.fraunhofer.de

Number of text pages: 35

Number of tables: 8

Number of figures: 8

Number of references: 30

Number of words in the Abstract: 188

Number of words in the Introduction: 654

Number of words in the Discussion: 1569

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Abstract

A novel computational method based on a *genetic algorithm* was developed to study composite structure of promoters of co-expressed genes. Our method enabled an identification of combinations of multiple transcription factor binding sites regulating the concerted expression of genes. In this paper we study genes whose expression is regulated by a ligand-activated transcription factor AhR (aryl hydrocarbon receptor) that mediates responses to a variety of toxins. AhR mediated change in expression of Ahr target genes was measured by oligo microarrays and by RT-PCR in human and rat hepatocytes. Promoters and long distance regulatory regions (>10 kb) of AhR responsive genes were analysed by the genetic algorithm and a variety of other computational methods. Rules were established on the local oligonucleotide context in the flanks of the Ah-receptor binding sites, on the occurrence of clusters of AhR recognition elements and on the presence in the promoters of specific combinations of multiple binding sites for the transcription factors co-operating in the AhR regulatory network. Our rules were applied to search for yet unknown Ah-receptor target genes. Experimental evidence is presented to demonstrate high fidelity of this novel *in silico* approach.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Introduction

Regulation of gene expression is accomplished through binding of transcription factors (TFs) to distinct regions of DNA (TF binding sites), and, after anchoring at these sites, transmission of the regulatory signal to the basal transcription complex. Some of these TFs are specific for a particular tissue, a definite stage of development, or a given extracellular signal, but most transcription factors are involved in gene regulation under a rather wide spectrum of cellular conditions. It is clear by now that combinations of transcription factors and not single transcription factors drive gene transcription and define its specificity.

The aryl hydrocarbon receptor (AhR) is a well understood ligand-activated transcription factor that mediates responses to a variety of toxins. Among them are halogenated aromatic toxins such as 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD), polynuclear aromatic hydrocarbons, combustion products, and numerous phytochemicals such as flavonoids and indole-3-carbinol (I3C). The nuclear AhR complex is a heterodimer composed of the AhR and AhR nuclear translocator (Arnt) proteins. Binding of this complex to so-called dioxin responsive elements (DREs) leads to activation of a large number of AhR-responsive genes including phase 1 (e.g. CYP1A1) and phase 2 drug metabolizing enzymes and also genes coding for cellular differentiation, metabolism and apoptosis amongst others (Safe and Wormke, 2003).

In depth studies of the last years revealed many details about the mechanisms of action of the AhR transcription factor and promoter activation of target genes including CYP1A1 (Whitlock 1999). Protein-DNA interactions of AhR were analysed and expression of target genes were studied. Despite our advanced understanding of AhR receptor biology, the precise mechanism of activation of target genes remains uncertain. There are a number of genes known

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

to be regulated (activated or repressed) by AhR in combination with other transcription factors, including AP-1and ER (Safe and Wormke, 2003) but the general combinatoric rules to enable robust prediction of gene expression are lacking. Furthermore, there is a need to identify networks of cooperating transcription factors. Thus, computational approaches for the prediction of the transcriptional network of AhR target genes are needed.

In the last few years a number of computational approaches were developed addressing the problem of combinatorial regulation of transcription by networks of transcription factors. Specific TF binding site combinations were used for an identification of muscle-specific promoters (Frech et al., 1998; Wassermann and Fickett, 1998), promoters of liver-enriched genes (Tronche et al., 1997), of yeast genes (Brazma et al.,1997), of immune-specific genes (Boehlk et al., 2000; Fessele et al., 2001; Kel 1993), and promoters of genes regulated during cell cycle (Kel et al., 2003). In the database TRANSCompel® (Kel-Margoulis et al., 2002) known composite regulatory elements and specific combinations of pairs or triplets of TF binding sites that are located in close proximity to each other were collected. This provides information on synergistic or antagonistic effects on gene regulation by different TFs. Nonetheless, information on TF networks is limited and descriptions of such networks are often highly speculative. With the introduction of global gene expression arrays into the armory of biomedical research investigators have been enabled to analyze the expression patterns of hundreds or thousands of genes altered in parallel in response to physiological as well as pathological stimuli. To gain insight into the mechanisms governing these changes we need to acquire a better understanding of the regulatory proteins as well as the regulatory regions of the genes involved. To facilitate the expansion of this knowledge base one suitable approach is to develop genetic algorithms for the identification and characterization of promoters and other regulatory elements. Including such

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

genetic algorithms we have developed a computational strategy which we used to analyze promoters of differentially expressed genes regulated by the aryl hydrocarbon receptor (AhR). Appplying this strategy, we identified composite modules that include AhR binding sites arranged in clusters and accompanied by a number of co-localized specific DNA motifs including HNFs, CEBPs and others. The revealed composite modules were used to identify novel gene targets based on whole genome searches.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Materials and Methods

Tissue and cultures.

Isolation of human and rat hepatocytes was done as described previously (Borlak et al., 2002; Borlak et al., 2003). Male Sprague Dawley rats weighing 200-220 g were obtained from Charles River (Germany). Anesthesia was done with Ketamin/Xylazin. Following midline incision the liver was perfused via the portal vein with 200 ml Ca^{2+}-free Krebs-Ringer buffer and subsequently with 100 ml containing Ca^{2+} and collagenase at 37°C. The liver capsule was gently removed and liver tissue was washed as described above. Primary human and rat hepatocytes were cultured enclosed between two layers of collagen as described previously (Dunn et al., 1991).

RNA isolation and production of copy RNA

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Total RNA was isolated using QIAGEN´s Rneasy total RNA isolation kit according to the manufacture´s recommendation. 10µg of total RNA was used for the synthesis of double-stranded cDNA with Superscript II RT. Primer extension was with HPLC-purified T7-$(dT)_{24}$ (GenSet SA) as a primer. After clean up the double-stranded cDNA was used for the synthesis of biotin-labeled cRNA (ENZO® BioArray High Yield RNA Transcript Labeling Kit, Affymetrix). cRNA was purified with Rneasy spin colums from Qiagen and cleaved into fragments of 35-200bases by metal-induced hydrolysis.

Microarray experiments

This was done according to the manufactor´s recommendations as detailed in Singh-Gasson et al. (1999) with n=3 repeats. Further, we employed n=15 cDNA arrays containing 302 well-known genes that code for detoxification, cell proliferation, tumor development, heat shock response, signal transduction, apoptosis, and cell cycle regulation (Borlak et al., 2004).

Thermocycler RT-PCR

For PCR amplification of cDNA, a 25µl reaction mixture was prepared containing 10x polymerase reaction buffer, 1.5mM $MgCl_2$, 0.4 nM dNTPs (Roche, Mannheim, Germany), 400nM concentration of the 3´and 5´specific primers (synthesized by Gibco, Karlsruhe, Germany), 1U Taq-polymerase and 1µl of cDNA.

PCR reactions were carried out in a thermal cycler (T3, Biometra, Germany) using primer specific melting, annealing and extension cycling conditions. DNA contamination was checked

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

for by direct amplification of RNA extracts prior to conversion of RNA to cDNA and was excluded. PCR-reactions were done within the linear range of amplification. PCR products were separated using an agarose gel, stained with ethidium bromide and photographed on a transilluminator. A semiquantitative measurement was performed using the program NIH Image V.1.62.

Sequence retrieval.

118 (68 human and 50 rat) differentially expressed genes were used as training sets (see Table 1). For 41 human and 37 rat genes the level of expression was determined experimentally.

The Ensembl database and UCSC Genome browser was used to extract the 5' regions of the AhR regulated genes and to build the training sets of sequences. For the majority of the genes and based on the Ensembl annotation of the first exons 5' regions could be retrieved. The beginning of the first exon was considered as a tentative TSS (transcription start site). However, for some genes identification of TSS was difficult. In these cases several possible 5' regions were considered. For some of genes given in Table 1 the annotation of promoters in genomes was not reliable (based on mapping of 5' incomplete ESTs or based on an *in silico* gene prediction). In such cases we did not include these genes into the training sets. Finally, 96 5' regions of these genes were included into the training set (81 human and 16 rat). Four sets of sub-sequences were prepared: (-500/+100), (-1000/+100), (-2000/+2000) and (-20000 –gene –+20000). We refer to these sequences as G600, G1100, G4000 and G20gene20kb sets.

Another training subset contained 5' upstream regions for 58 genes (42 human and 16 rat) for which gene expression was quantified (see Table 1). We used the logarithms of the

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

relative expression values for further studies of correlation between promoter structure and the relative expression values. Furthermore, several sequence sets were extracted from the human genome database and used as controls. One set contained 200 promoter sequences from human genes in chromosome 21 (PR). Another set contained far upstream sequences of human genes (based on RefSeq annotation) from chromosome 21 (from –50kb to –40kb) (UPG10kb set). Another set was used for an analysis of exons 3 sequences of human genes (EXON3 set). No functional TF binding sites are within exon 3, thus these sequences can be used as a TF "site free" background. A further sequence set containing randomly selected intergenic sequences of chromosome 22 (INTERG set) were also used and considered in our analysis.

Databases.

Two databases for prediction of gene expression were used (BIOBASE GmbH, Wolfenbüttel, Germany). TRANSFAC® is a database that collects information about gene regulation in eukaryotes based on binding of transcription factors to their target sites. We used TRANSFAC® Professional rel.6.4 and TRANSCompel® (rel.6.4) which contains known composite regulatory elements in mammalian genes.

Weight matrix method for recognition of AhR sites.

The most widely used methods for recognition of transcription factor binding sites is the application of positional weight matrices (PWMs) (Quandt et al., 1995; Whitlock, 1999). TRANSFAC® is the largest collection of weight matrices for eukaryotic transcription factors

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

(Wingender et al., 2001) (www.biobase.de; www.gene-regulation.com). In this database three different weight matrices for AhR sites were stored prior to this study (acc. numbers: M00139, M00235 and M00237). In the current work a novel weight matrix was constructed on the basis of 25 known, experimentally verified, genomic AhR-binding sites (TRANSFAC® acc. number: M00778, id: V$AHR_Q5). All these matrices were used for searching potential AhR sites in genomic sequences. For this search we employed the MATCH™ algorithm calculating scores for the matches by applying the so-called information vector (Kel et al., 2003).

Software tools for searching TF binding sites.

We used the TRANSPLORER® software package (BIOBASE GmbH, Wolfenbüttel, Germany) to identify potential TF binding sites. This software uses a position weight matrix (PWM) collection and the most up-to-date library of matrices derived from the TRANSFAC® Professional database. A TF binding site search can be made using all or a particular subset of matrices from this library. Such matrix subsets with defined cut-off values for every matrix are called “profiles” in TRANSPLORER®.

We used a profile that included matrices for different transcription factors of vertebrate organisms (TRANSFAC® rel. 6.4) known to be involved in gene regulation in liver, in immune cells and in the cell cycle. The cut-offs for the matrices are set to a very low value to minimize the rate of false negative predictions.

Method for revealing short sequence motifs in the flanking regions of TF sites. Local context.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

To analyze the flanking regions of AhR sites, we applied an algorithm that was developed recently (Kel et al., 2001; Kel et al., 1999). The composition of over- and underrepresented short oligonucleotides (=local context) is investigated in the flanking regions of functional AhR sites (50bp to the left and right of the 11 bp long AhR sites). This local context is taken into account while searching genomes for potential AhR sites.

The applied algorithm is based on a comparison of two sets of sequences of equal length L: a training set Y consisting of the functional AhR sites including their flanking regions and a negative control set N consisting of the sequences from "non-promoter" regions of the human genome. To compile the N set we computed a MATCH™ search in the INTERG sets using the AhR matrix (acc. no. M00778; score value $q > q_{cut\text{-}off}$=0.98, it gives approximately 1 match per 5kb of genomic sequence, as it was tested on the chromosome 22). 100 of randomly selected matches together with their flanks were placed into the control set N. Thus, set N consists of sequences that contain a motif fitting to the AhR matrix, however, because of its position in the genome most of these "sites" should not be functional. This way, by comparison of sets Y and N we can reveal features that characterize the sequence environment (context) of functional AhR sites. As contextual features, we consider the frequency of occurrence of short motifs $\lambda = (a_1 a_2 .. a_k)$ $(a \in \{A,T,G,C,W,S,R,Y,M,K,B,V,H,D,N\}$ [1]) of the length $k \leq 4$ in a window $w=[t_1,t_2]$ $(0<t_2<t_1<L-k+1)$.

In our previous work (Kel et al., 2001; Kel et al. 1993) we described a statistical method that permits an identification of motifs λ and the windows w that are characterized by a significant difference of their frequencies $f(\lambda,w,S)$ in the sequences S from the sets Y and N. The

[1] We are using the following one letter code for different combinations of alternative nucleotides: W-(A/T, (read A or T)); R-(A/G); M-(A/C); K-(T/G); Y-(T/C); S-(G/C); B-(T/G/C);V-(A/G/C); H-(A/T/C); D-(A/T/G); N-(A/T/G/C).

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

motifs found are used then for creating a context analyzer that is able to perform an additional filtering of the potential sites revealed by the weight matrix method.

The context analyzer is developed by means of linear discriminant analysis. The motifs selected at the previous step (λ_1, λ_2,…, λ_m) are used for construction of a linear classification function discriminating sets Y and N. So, for every sequence X we calculate the score of context:

$$d = \beta + \sum_{i=0}^{m} \alpha_i \times f(\lambda_i, w_i, X) \qquad (1)$$

where α_i and β are the coefficients of the discriminating function.

A genetic algorithm to determine composite regulatory modules (CM).

We define a composite module CM as a set of TF weight matrices with given matrix cut-offs and other parameters which is associated with a specific functional type of gene regulatory regions. We have developed a new computational method to determine CMs in a set of promoter (or other regulatory) sequences of regulated genes. This method is based on a *genetic algorithm* (a prototype of this method is incorporated in the tool package ClusterScan (Kel-Margoulis et al., 2002). The *CMs* are characterized by the following parameters*: K*, the number of PWMs in the module (typically 6 to 12), cut-off values $q_{cut-off}^{(k)}$, relative importance values $\phi^{(k)}$ and max. number of best matches $\kappa^{(k)}$ that are assigned to every weight matrix *k* (*k=1,K*) in the *CM*. These *K* matrices are selected by the program from a library of all considered matrices. We use different profiles including the profile vertebrate_minFN62.prf, which includes 410 matrices for different transcription factors of vertebrate organisms (TRANSFAC® rel. 6.4). Some matrices

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

are organized in pairs. Such pairs are designed to model composite elements witch play a very important role providing synergistic or antagonistic regulation of transcription through binding of different transcription factors. (Kel et al., 1995). A parameter *R* is defined that puts a limit on a distance between matches of these matrix pairs (at least one pair of matches should be found fitting this limit). When all these parameter settings can be defined a "composite module score" (CM score) is given for any sequence X using the following equation:

$$F_{CM}(X) = \sum_{k=1,K} \phi^{(k)} \times \sum_{i=1}^{\kappa^{(k)}} q_i^{(k)}(X) \qquad (2)$$

where $q_i^{(k)}(X)$ are the $\kappa^{(k)}$ best scoring sites found in the sequence *X* by the matrix *(k)*. Then the *genetic algorithm* is used to determine the specific parameters of *CMs* for a particular set of promoters. The general description of the *genetic algorithm* is reported elsewhere (Kel et al., 2001).

We define the goal function *G* as a weighted sum of false negative and false positive errors and determine the statistical significance (t-test) over several random iterations of bootstrap procedure by splitting the initial sequence sets into a training and a testing subset. In addition, we test the normal distribution of the *F* function over the set of positive and negative sequences. Such thorough way of calculating the goal function using bootstrap test and normality check allows to assess the usability of the obtained solutions for classification of individual sequences. Due to the small size of the sequence set we decided to use the complete set for the training and apply the thorough analysis of distribution rather then to sacrifice a part of the set for independent testing. Nonetheless, the final biological and truly independent test of the method is done by the complete genome screening and by comparing results with microarray data (see Results).

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

The program CMFinder (cmf) takes as an input two sets of sequences (the set which is analysed and a background set) and a set of weight matrices for transcription factors. For defined parameters *K* and *R* and over a number of iterations, the program optimizes the set of matrices selected, their number, their cut-offs, the relative importance and the maximum number of best matches. The user can vary parameters *K* and *R* and compare results of the program. The output of the program is a profile ready to run by MATCH™ or TRANSPLORER®.

Method to determine CMs correlated with gene expression.

The method is based on a modification of the *genetic algorithm* described in the previous section. The program CMFinder now searches for a combination of TF matrices that minimizes a goal function G that is a weighted sum of the least square deviation between logarithms of gene expression values and the score function F_{CM} described in the previous section and a rank correlation coefficient

$$\alpha \times \sum_{genes} (LR - F_{CM})^2 + \beta \times \rho_{rank}(LR, F_{CM}) \rightarrow \min \qquad (3)$$

here $LR = \log_2(ex1/ex2)$ which is the log-ratio of the expression values of a particular gene in the experiment versus control, e.g. a signal log-ratio.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Results

Construction of a novel AhR weight matrix.

We initially analysed experimental findings on AhR bindings sites in promoters of targeted genes. From the 35 sites reported in 17 genes (human, mouse and rat) 25 independent sites were selected after exclusion of close homologs (see Table 1). A novel algorithm was used to identify an optimal length of the matrix (Kel et al., 2003) which comprised 11 nucleotides. This matrix was stored in TRANSFAC® (acc no. M00778) and used for further analysis. Fig. 1 depicts the constructed AhR weight matrix and the corresponding consensus sequence. Cut-off $q_{cut\text{-}off}$ =0.966 is the maximal cut-off that allows recognition of all 25 sites of the training set. This cut-off was used in further analysis.

The core of consensus sequence of the new matrix is "TNGCGTG", which fits exactly to the 7 nucleotide length consensus reported earlier (Watson & Hankinson, 1992; Bacsi et al., 1995). So, with our analysis which was based on much broader collection of known AhR sites we confirmed the well known consensus. In addition to that, we constructed the weight matrix that provide information on the nucleotide preferences in a broader, 11 nucleotide, window around this core consensus sequence.

Modulation of gene expression after AhR activation

Microarray and RT-PCR experiments enabled an assessment of gene expression and the results are given in the Table 1. We selected 114 genes (64 human and 50 rat) that changed their

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

expression upon AhR activation. Among the 114 "changed" genes we reconfirmed some of the well known targets of AhR with their known binding sites, such as CYP1A1, CYP1B1, Ugt1a1, c-fos amongst others. We also confirmed some of the recently reported AhR targets including Bcl-2, Bcl-xl, BAD and cyclooxygenase 1 as shown in the Table 1.

Analysis of short sequence motifs in the flanking regions of AhR sites (local context).

To analyze the flanking regions of AhR sites we computed the *Local context* as described in the Methods section. The analyzed sequences contained a proven AhR site in the center (11bp long, aligned according to the best fit to the matrix) and 50bp flanks left and right from the site (the DNA strand orientation with regard to the TSS site was not considered). We identified 13 characteristics in these sequences (Table 2). 9 of them are overrepresented in comparison to the negative control set of sequences (positive characteristics) and 4 underrepresented (negative characteristics). The negative control set was prepared as described in the Method section by selecting from the INTERG sequences those subsequences of 111 nucleotide length in which the 11 central positions match with the AhR weight matrix (score cut-off $q = 0.966$).

As shown in Table 2, the selected motifs are 3-4 nucleotides long and contain many ambiguous letters (from the IUPAC 15 letter code). The corresponding windows in which the selected motifs were found to be overrepresented/underrepresented are rather different in their length and location (second and third column in Table 2). Some of the windows are located on the left flank, some on the right and some of them overlap in the central positions.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

On the basis of these 13 characteristics (m = 13) we constructed the linear classification function according to formula (1) discriminating sets of real AhR sites from the false positives in INTERG sequences. We refer to this function as AhR *score of local context d*.

The advantage of the *score of local context d* as an additional filter to identify potential AhR sites is demonstrated in Fig. 2. As a first step we identified the potential AhR sites by using the weight matrix with a definite cut-off value $q_{cut\text{-}off}$. Then, we applied the *d score* to an extended region surrounding each potential AhR site. The requirement $d > 0$ ($d_{\text{cut-off}}$=0.0) is used as an additional criterion for recognition of AhR sites. In Fig. 2 the distribution of *d* values for the set of real AhR sites (red) and for the set of "false" AhR-like motifs in the intergenic regions (blue) are shown. These two distributions are clearly distinct from each other and by setting the cut-off value to *d=0.0* we can filter out almost 80% of the "false" sites, while loosing only 14% of the true positive sites. Thus, this method allows us to increase the correctness of our predictions of AhR sites in genomic sequences.

Searching for AhR sites in genomic sequences using score of context.

We applied our high precision method for recognition of AhR binding sites to the genomic sequences of genes that were shown to be regulated (up or down) after activation of AhR. It is known that functional AhR sites may be situated in the promoter regions of the AhR target genes, in their introns as well as in far upstream and downstream non-coding regions around the genes. Therefore, we analysed the G20gene20kb set (see Methods) that includes 20kb regions upstream and downstream of the genes. This set contains 40 sequences of the total length 1,855,876bp.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

We searched for AhR sites on either strand using the consecutive application of AhR site weight matrix and the score of context. 195 sites with a frequency of 0.000105072 sites per bp (or approximately 1 site per 10,000bp) were identified with a matrix score cut-off *s=0.98* and the context score cut-off *d=0.0*. Sequences of chromosome 21 with a total length of 34,011,570 bp were used for data curation. The same cut-offs were applied (s=*0.98* and d=0.0). The total number of identified AhR sites was 1880 which corresponds to 0.00005528 sites per bp (or approximately 1 site per 18,000bp). Consequently, the frequency of AhR sites in the AhR-regulated promoters is almost 2 times higher when compared with overall sequences of the entire chromosome 21.

To check whether the revealed sites have any preferences in their location relative to the start of transcription (TSS position) we calculated the frequencies of predicted AhR sites in different regions around TSS of AhR-regulated genes. Strikingly, for some windows of limited length around the TSS the frequency of AhR sites is much higher than in the other regions of regulated genes as well as in the whole chromosome 21 (see Fig. 3). For instance, within the window: [-1000 TSS +1000] (-1000 upstream and +1000 downstream from the start of transcription) we observed AhR sites with a frequency of 0.0005125 sites per bp (or approximately 1 site per 2000bp), which is almost 10 times higher than in the whole chromosome 21.

Although the frequency of AhR sites is very high near the start of transcription, it is not the only place where binding sites can be found. They can be found in the promoter regions of AhR-regulates genes (in 5' proximity to the start of transcription), inside the first non-coding exon, in introns as well as in the far 5' and 3' regions of the genes. A very distinct feature of the location of AhR sites is a preponderance to form clusters. These clusters may contain from k = 4-

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

5 up to 10-15 sites and the extent of the clusters varies from l = 50bp up to 1000bp (see examples in the Fig. 4).

We computed the probability of a given (k,l) cluster to appear just by chance in a sequence of length L given the total number of sites K in this sequence. The probability is calculated according to the algorithm described previously (Kel et al., 2001). We found statistically significant AhR-site clusters in about 40% of AhR-regulated genes. Among them there are the following genes: ALDH3A1(human) - 6 sites inside 66bp, from +1540 to +1606; BAD(human) - 6 sites inside 338bp, from -7519 to –7181; NR3C1 (human) - 6 sites inside 765bp, from –893 to –128) and many others.

Composite modules revealed in promoters of AhR-regulated genes.

In order to reveal novel composite modules the local nucleotide sequence of promoter regions of AhR-regulated genes was analysed. Several subregions around TSSs were studied, e.g. (-2000/+2000), (-1000/+100) and (-500/+100) and we applied different parameters for searches with the CompositeModuleFinder. We compared the structure of AhR-regulated promoters to a set of promoters randomly taken from the human genome. An extended table with the results of this analysis is available on request. Among the most discriminative matrices were those recognizing HNF1, AhR, GR, OCT and C/EBP (Table 3). Interestingly, the AhR matrices selected by the algorithm were in the set of promoters (-1000/+100) only. Shorter and longer promoter sequences were not selected by the algorithm. This indicates that not only the presence of a binding site is important, but also its relative position within the promoter.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Among the most prominent matrix pairs that were selected by the algorithm are: HNF1/Sp1 and AP-1/NF-1 for the max. distance R=100bp; E2F/NF-1, AhR/Myb, HNF3/NFY, HNF6/NF-kappaB, and Sp1/Myb for the max. distance R=50bp and HNF-1/GR for the max. distance R=40bp. Importantly, not all matrices found individually were also found in pairs. Thus, a composite module for promoters could be defined.

In Table 4 a combination of individual matrices and matrix pairs for a set of promoters of AhR regulated genes (-2000/+2000) are listed. One matrix (YY1) and two pairs (E2F/ROR and AhR/CREB) were selected by the algorithm, since their frequencies in the AhR responsive promoters were lower than in others. All other matrices and pairs of matrices were selected to define an AhR-associated composite regulatory module CM_{AhR} in the promoters investigated.

As given in Fig. 5, the distribution of the composite module score F_{CM} is distinctly different (t-test value is 12.12, p value = $1.7\ 10^{-28}$) and demonstrates robust recognition of AhR-responsive promoters. A summary of sites and site pairs from CM_{AhR} found in AhR-regulated genes is given in Table 5.

The composite structure of an AhR-regulated promoter, e.g. the human ALDH3A1 gene is given in Fig. 6. Identified sites, site pairs and the AhR site cluster are all distributed over the whole region of 4kb around the TSS. Comparison of this region between different species reveals that only the proximal promoter is conserved. There were no detectable homologies between human and mouse/rat in the other parts of this 4kb region. It seems that the sites predicted in those parts are specific for the human gene only. In previous experimental studies several AhR sites as well as NF-1 and GR sites were detected in different 5' upstream regions of rat and mouse ALDH3A1 gene (Vasiliou et al., 1999). Here we confirmed that the proximal promoter of this gene has a typical AhR – regulation associated structure.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Prediction of gene expression based on the composite modules.

Next, we addressed the differences in the structure of the regulatory regions between up- and down-regulated genes. The dynamics of regulation is actually quite different. Therefore, we aimed to identify transcription factors which influence the dynamics of up- or down-regulation in response to AhR protein-DNA interaction.

We thus compared two sets of promoters: for the up-regulated genes ExpressUP and for down-regulated genes ExpressDOWN (taking the regions of promoters from –2000 to +2000). The relative expression level was considered as additional information for this analysis.

With the help of the CompositeModuleFinderExpress, we determined a set of matrices that correlates with gene expression in the two sets of regulated genes (Table 6). Patterns with positive “impact” values ϕ are over-represented in promoters of up-regulated genes (shown in bold). Patterns with negative "impact" values are over-represented in promoters of down-regulated genes (underlined).

A liniar function is computed (see Method section, equation 2) that predicts the level of up- or down-regulation of a gene based on the sequence of its promoter. This function takes into account “impact” values for selected patterns given in Table 6. We obtained a significant correlation when the logarithm of the experimentally determined gene expression levels were plotted against predicted ones (Fig. 7).

Identification of genomic target for the aryl-hydrocarbon receptor (AhR).

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

The study of promoters of AhR-regulated genes and the hierarchy of AhR sites enabled us to develop a definition of a novel genetic algorithm. The rules of this algorithm are given below and were applied to search for new AhR targets within the human genome:

1. Promoters of 1100bp in length around the tentative TSS (based on available RefSeq annotation of genes) were retrieved. Only genes with annotated 5'-UTRs longer than 150bp were considered. This enhances the reliability of the selected TSSs, as shorter 5'-UTRs annotations frequently result from non-full length mRNAs. By this approach 9246 promoters were retrieved from the human genome.

2. We searched for potential AhR binding sites using the newly generated matrix (cut-off = 0.966). Initally, 9982 sites were found of which 3999 sites passed through the local context filter.

3. For every gene that contains at least one potential AhR site within a region of 20kb upstream and downstream of the TSS clusters with k = 2 to 7 covering the regions from L=6 to 800 bp were computed. The probability for clusters to appear by chance was calculated as well. Finally, we selected only genes that have at least one cluster of AhR sites with a probability less then 0.1. With this approach 864 genes were selected. The full list of genes is given as supplementary data, which is available on request.

4. Composite modules (CMs) were identified with the CMFinder program. So far, 194 potential target genes were identified that matched both composite modules.

5. A list of 71 predicted target genes with the highest prediction scores are present in Table 7. Furthermore, expression of predicted genes was verified experimentally by RT-PCR and/or microarray-studies. The results are given in Table 7 and we demonstrate approximately 75 % concordance between predicted regulation and experimentally verified change in expression.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Discussion

We developed a computational method to investigate promoters of genes that are regulated by the Ah-receptor. Our approach allows specific features of regulated promoters to be identified and to utilize this knowledge for the identification of novel target genes in various genomes. The developed method is based on a concept of hierarchical composite structure of eukaryotic gene regulatory regions.

The concept of the composite regulatory modules (CM).

Functionally related genes involved in the same molecular-genetic, biochemical, or physiological process are often regulated coordinately by specific combinations of transcription factors. Dynamic function-specific complexes of many different transcription factors, so called enhanceosomes (Merika and Thanos, 2001) are formed at gene promoters and enhancers driving gene expression in specific manners. At the level of DNA, the blueprints for assembling of such variable TF complexes on promoter regions may be seen as specific combinations of TF binding sites located in close proximity to each other. We call such structures “composite regulatory modules”. The various types of composite modules can be classified according to their ability to regulate gene expression. There may be several different types of CMs located in the regulatory region of one gene, that can be spaced (e.g. liver specific and muscle specific enhancers of one gene) or overlapping. We also consider the hierarchy of the CMs based on the number of different TFs involved and the length of the sequence covered. Composite elements (CEs) consisting of two/three closely located sites

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

belong to the lowest hierarchical level of CMs. CMs of a higher level may include more sites as well as CMs of a lower level such as CEs.

Recently, a number of approaches identifying composite motifs were described: BioProspector (Liu et al., 2001), Co-Bind (GuhaThakurta and Stormo et al., 2001), MITRA (Eskin and Pevzner, 2002), dyad search (van Helden et al., 2000).

These programs help to discover new regulatory sites for yet unknown transcription factors, but an "*ab initio*" motif finding method is limited by the length of sequences and may not be suitable for the analysis of long regulatory regions of genes of human and other higher eukaryotic organisms. A valuable source to identify transcription factor binding sites is the TRANSFAC® database (http://www.biobase.de), (Wingender et al. 2001).

We developed a new method for detecting *de novo* composite modules using information from TRANSFAC® and the weight matrices for many different transcription factor families. It is a rather complex computational task to find combinations of weight matrices that are characteristic for a particular group of genes. Many variables are included in the model, such as the number of matrices in a module, the maximal number of matches considered, the distances between matches, cut-off values for the matrices and their relative impact on the composite score. To find an optimal structure of the model we applied a GA (genetic algorithm) approach which was shown to be efficient in finding solutions. One of the reasons to use a genetic algorithm was the need to perform simultaneous fitting of the modeled parameters and to reduce the number of parameters. Because of the large number of possible matrix combinations it is impractical to use canonical multidimensional regression (linear or logit) without prior reduction of the number of parameters. Of course, when selecting a heuristic algorithm for optimization one must be aware that it finds only local optima and solutions can vary in different runs of the

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

program. The selection of a proper goal function is very important for this approach. We constructed a robust goal function using some ideas of "utility calculations" (Kel et al., 1993). The requirement for normality and bootstrap procedures (see Method section) provides the correct environment to find composite modules that describe the whole set of promoters and not just a particular subset.

The success of applying such tools greatly depends on how well the general model of the promoter we are optimizing actually fits to reality. Due to the limitations of data here we consider relatively simple models only. To improve the model in future we will consider such additional parameters as mutual distances between each site in the CM, their orientations to each other and location relative to the start of transcription. All these and many other parameters could be easily incorporated in the model and optimized by the help of the genetic algorithm.

The algorithm which we applied identified several factors including, SP1, NF-kappaB and C/EBP. Interestingly, the AhR matrix alone was insufficient for the description of AhR-regulated promoters. Thus, specific combinations with other factors provided the best discrimination for detecting AhR-regulated genes.

We also performed computational tests with the cmf program on virtual data. We generated a set of random sequences by shuffling randomly nucleotides in every sequence of the G4000 set. Then we implanted several binding sites of certain transcription factors by placing them at random. We used the cmf program to compare the sequence set with the implanted sites and those selected at random. We varied the minimal scores of the sites and their frequencies to find the borders. The result of this simulation (see Table 8) shows that the cmf program was able to determine correctly all TF matrices used for site implantation, although the more matrices were used for implantation the more difficult it became for the program to reveal the sites

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

correctly. Importantly, matrices were found to behave differently in the test. For instance, the matrices for C/EBP and OCT were easily identified whereas recognition of the matrix for AhR factors was more difficult, and the AP-1 matrix was most difficult to identify. Obviously, similarity between different matrices in the library is a confounding factor. Unfortunately, the algorithm can not identify different members belonging to the same TF family. For example, in the case of C/EBP identification of a particular member, e.g. C/EBP-alpha, beta or gamma was impossible. Nevertheless, specific combinations of transcription factors are indicative of a particular regulatory mechanism.

Another test performed was the “knocking-out” of those matrices most important for AhR promoter activation, e.g. distortion of 6 matrices: AhR, C/EBP, OCT, AP-1, NF1 and HNF1 by exchanging values for nucleotides A versus T and C versus G. After such distortion (without changing the C/G content), the program was run again with complete library where these 6 matrices were “knock-outed”. No significant differentiation could be identified between AhR-regulated promoters and other promoters. This proved the specificity of our algorithm and indicated that the matrices we identified were indeed important for recognition of the AhR-regulated genes.

Our composite module contained matrices for AhR, PPAR, HNF-6, STAT, ROR and ETS. It is interesting that three different AhR matrices were included by the algorithm in the CM. This suggests that sites for AhR in the promoters play an important role in the up or down regulation of these genes. V$AHR_01 got the maximal impact value. It seems to be specific for several cytochrome P450 genes and influences their expression in response to AhR. V$AHRARNT_01 was constructed on the basis of data from SELEX experiments whereas V$AHR_Q5 was constructed in the current study on the basis of genomic sites. These two

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

matrices have very different impact values: the first matrix has a positive, and the second matrix a negative impact value. Comparison of the structure of these two matrices shows that they are very similar in the core, but differ in some nucleotides at the flanks. For example, in the position 10 in the matrix V$AHR_Q5 (see Fig. 1), the most prominent nucleotide is G, whereas in the matrix V$AHRARNT_01 (see TRANSFAC®) nucleotide G in the corresponding position is absolutely "forbidden". This could influence binding of some other factors such as repressors in the vicinity of AhR sites.

The combination of matrices and matrix pairs included in the CMs revealed a network of transcription factors governing the AhR dependent gene regulation. Fig. 8 depicts this network. Some of the genes of the transcription factors included in the CM, such as C/EBP-alpha, HNF1-beta, HNF3, HNF4-gamma, OCT-1, are regulated by AhR as shown previously (Borlak et al., 2002, Borlak et al., 2003). Analysis of their promoters using the CM points to an intriguing network of transcription factors with several feedback loops and a hierarchy of regulatory signals. This network of transcription factors can also explain the regulation of several genes that are not direct targets of AhR binding. Their regulation can be mediated through other TFs whose expression is regulated directly by AhR (see Fig. 8).

On the genome scale to search for potential new AhR target genes we applied very stringent set of criteria to avoid high rate of false positives. In addition to the high score of the CM match, we require AhR sites with the proper oligonuclotide context to be found in the promoters as well as presence of clusters of AhR sites. With such stringent criteria we definitely miss the indirect targets - genes whose AhR-dependent regulation is mediated through other TFs. As a result, a high proportion of true positive prediction is to be expected as it is confirmed by comparison with the results of microarray studies.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

In conclusion, a computational method was developed for an *ab initio* identification of AhR regulated genes. The applied algorithm revealed a specific motif within a distinct promoter region near the AhR sites. This finding enabled us to identify clusters of regulatory proteins. Importantly, the genetic algorithm predicts the level of transcript expression and may be used for the identification of genes regulated by AhR.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Acknowledgments

The authors are indebted to Olga Kel-Margoulis (BIOBASE GmbH, Germany) for fruitful discussion of the results, to Tatjana Ivanova Institut (Cytology and Genetics, Novosibirsk) for testing of the genetic algorithm approach, to Birgit Lewicki-Potapov (BIOBASE GmbH, Germany) and Rüdiger Zemlin (Fraunhofer Institute, Hannover, Germany) for the help in preparing sequence sets, and to Asili Barre-Dirrie for the help in collection of AhR-binding sites.

This work was mainly funded by BIOBASE GmbH (Wolfenbüttel, Germany). Parts of this work was supported by Siberian Branch of Russian Academy of Sciences, by grant of Volkswagen-Stiftung (I/75941) and by a grant of the Lower Saxony ministry of culture and science to Jürgen Borlak.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

References

Boehlk S, Fessele S, Mojaat A, Miyamoto NG, Werner T, Nelson EL, Schlondorff D and Nelson PJ (2000) ATF and Jun transcription factors, acting through an Ets/CRE promoter module, mediate lipopolysaccharide inducibility of the chemokine RANTES in monocytic Mono Mac 6 cells. *Eur J Immunol* 30:1102-1112.

Bacsi SG, Reisz-Porszasz S and Hankinson O (1995) Orientation of the heterodimeric aryl hydrocarbon (dioxin) receptor complex on its asymmetric DNA recognition sequence. *Mol Pharmacol* 47:432-438.

Borlak J, Dangers M and Thum T (2002) Aroclor 1254 modulates gene expression of nuclear transcription factors: implications for albumin gene transcription and protein synthesis in rat hepatocyte cultures. *Toxicol Appl Pharmacol* 181:79-88.

Borlak J, Hock A, Hansen T and Richter E (2003) DNA adducts in cultures of polychlorinated biphenyl-treated human hepatocytes. *Toxicol Appl Pharmacol* 188:81-91.

Borlak J, Drewes J, Hofmann K and Bosio A (2004) Toxicogenomics applied to in vitro toxicology: A cDNA array based gene expression and protein activity study in human hepatocyte cultures upon treatment with Aroclor 1254. (submitted)

Brazma A, Vilo J and Ukkonen E (1997) Finding Transcription Factor Binding Site Combinations in Yeast Genome, in Computer Science and Biology (Frishman D and Mewes HW

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

eds) pp 57-59, Proceedings of the German Conference on Bioinformatics GCB '97 D, Martinsried.

Dunn JC, Tompkins RG and Yarmush ML (1991) Long-term in vitro function of adult hepatocytes in a collagen sandwich configuration. *Biotechnol Prog* 7:237-245.

Eskin E and Pevzner PA (2002) Finding composite regulatory patterns in DNA sequences. *Bioinformatics* 18:Suppl. 1, S354-S363.

Fessele S, Boehlk S, Mojaat A, Miyamoto NG, Werner T, Nelson EL, Schlondorff D and Nelson PJ (2001) Molecular and in silico characterization of a promoter module and C/EBP element that mediate LPS-induced RANTES/CCL5 expression in monocytic cells. *FASEB J* 15:577-579.

Frech K, Quandt K and Werner T (1998) Muscle actin genes: a first step towards computational classification of tissue specific promoters. *In Silico Biol* 1:29-38.

Guha Thakurta D and Stormo GD (2001) Identifying target sites for cooperatively binding factors. *Bioinformatics* 17:608-621.

Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV and Wingender E (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res* 31:3576-3579.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Kel AE, Kel-Margoulis OV, Farnham PJ, Bartley SM, Wingender E and Zhang MQ (2001) Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors. *J Mol Biol* 309:99-120.

Kel AE, Ponomarenko MP, Likhachev EA, Orlov YL, Ischenko IV, Milanesi L and Kolchanov NA (1993) SITEVIDEO: A computer System for Functional site Analysis and Recognition. Investigation of the human splice sites. *Comput Applic Biosci* 9:617-627.

Kel A, Kel-Margoulis O, Babenko V and Wingender E (1999) Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. *J Mol Biol* 288:353-376.

Kel OV, Romaschenko AG, Kel AE, Wingender E and Kolchanov NA (1995) A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Res* 23:4097-4103.

Kel-Margoulis O, Kel AE, Reuter I, Deineko IV and Wingender E (2002) TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res* 30:332-334.

Kel-Margoulis OV, Ivanova TG, Wingender E and Kel AE (2002) Automatic annotation of genomic regulatory sequences by searching for composite clusters. *Pac Symp Biocomput* 187-198.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Liu X, Brutlag DL and Liu JS (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput* 127-138.

Merika M and Thanos D (2001) Enhanceosomes. *Curr Opin Genet Dev* 11:205-208.

Niehof M, Streetz K, Rakemann T, Bischoff SC, Manns MP, Horn F and Trautwein C (2001) Interleukin-6-induced tethering of STAT3 to the LAP/C/EBPbeta promoter suggests a new mechanism of transcriptional regulation by STAT3. *J Biol Chem* 276:9016-9027.

Quandt K, Frech K, Karas H, Wingender E and 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.

Safe S and Wormke M (2003) Inhibitory aryl hydrocarbon receptor-estrogen receptor alpha cross-talk and mechanisms of action. *Chem Res Toxicol 16*:807-816.

Singh-Gasson S, Green RD, Yue Y, Nelson C, Blattner F, Sussman MR and Cerrina F (1999) Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat Biotechnol* 17:974-8.

Tronche F, Ringeisen F, Blumenfeld M, Yaniv M and Pontoglio M (1997) Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *J Mol Biol* 266:231-245.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

van Helden J, Rios AF and Collado-Vides J (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res* 28:1808-1818.

Vasiliou V, Reuter SF, Williams S, Puga A and Nebert DW (1999) Mouse cytosolic class 3 aldehyde dehydrogenase (Aldh3a1): gene structure and regulation of constitutive and dioxin-inducible expression. *Pharmacogen* 9:569-580.

Wasserman WW and Fickett JW (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol* 278:167-181.

Watson AJ and Hankinson O (1992) Dioxin- and Ah receptor-dependent protein binding to xenobiotic responsive elements and G-rich DNA studied by in vivo footprinting. *J Biol Chem* 267:6874-6878.

Werner T (1999) Models for prediction and recognition of eukaryotic promoters. *Mamm Genome* 10:168-175.

Whitlock JP (1999) Induction of cytochrome P4501A1. *Annu Rev Pharmacol Toxicol* 39:103-125.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Wingender E, Chen X, Fricke E, Geffers R, Hehl R, Liebich I, Krull M, Matys V, Michael H, Ohnhäuser R, Prüß M, Schacherer F, Thiele S and Urbach S (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res* 29:281-283.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Table Legends:

Table 1. Expression levels of different genes after treatment of human or rat hepatocytes with Aroclor 1254, and the LocusLink ID (identifier) of these genes.

Table 2. Over and underrepresented oligonucleotides found in the flanking regions of AhR sites. The first 9 characteristics are “positive”, the last four are “negative”. The sequences considered were of the length 111bp (11bp AhR site in the center and 50bp flanks).

Table 3. List of factors the matrices of which were selected by the algorithm in 5’ regions of different length.

Table 4. Composite module CM_{AhR} constructed for the (-2000 +2000) set of promoters.

Table 5. Summary of sites, site pairs found in promoters (-2000 +2000) and clusters found in 5’ regions (-20kb +20kb) of AhR-regulated genes and other genes for transcription factors involved.

Table 6. Composite module E001 that correlates with gene expression in the G4000ExpressUP and G4000ExpressDOWN sets.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Table 7. 71 potential target genes for AhR regulation selected by the highest values of different prediction scores.

Table 8. Results of testing CMFinder on simulated data.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Figure Legends:

Fig. 1. AhR weight matrix (TRANSFAC® acc. number: M00778, id: V$AHR_Q5). In each of the 11 positions the weights for the four nucleotides are given.

Fig. 2. Discrimination between real AhR sites with 50bp flanks from the false positives in genomic sequences. Values of the *score of local context d* are shown on the abscissa. The numbers of observations (No of obs) are given on the ordinate.

Fig. 3. Frequency of AhR sites found by the program in different genomic sequences. Frequency in the whole 21 chromosome is shown by the yellow bar. Blue bars represent the frequency of the sites in different parts of the AhR-regulated genes.

Fig. 4. Some examples of the distributions of the found AhR sites in AhR regulated sequences. For each cluster the probability to appear by chance was calculated. Positions of the start of transcription (according to the RefSeq annotation of the first exon) are shown by arrows. The scale shows positions relative to the start of transcription. Short lines above the scale show the locations of predicted AhR binding sites.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Fig. 5. Discrimination between (-2000 +2000) promoters (black bars) and background promoters from chromosome 21 (PR:-2000 +2000) (stripy bars) by the composite module CM_{AhR}. The composite module score F_{CM} is given on the abscissa the numbers of observations (No of obs) are given on the ordinate.

Fig. 6. Analysis of potential TF binding sites in the 5' region of the human ALDH3A1 gene. Locations of the first exon of the transcripts are shown in yellow. a) Map of potential TF binding sites found in the region of the gene from –2000 to +2000. Potential composite elements AhR/Myb, AhR/NF-1, Myb/SP-1 and NF-1/AP-1 are marked. A cluster of AhR sites in +1500 region is also marked. b) Phylogenetic footprinting of a region near start of transcription (TSS) for 3 species (human, mouse and rat). Potential CE AhR/Myb is conserved among these three species. The experimentally known AhR site in the mouse gene is underlined. The predicted sites for HNF-3B, C/EBP and USF are also conserved for three species. NF-1 and second MyB (both shown by a single-dashed line) can be identified in human and mouse sequences, but not in rat.

Fig. 7. Correlation between real relative expression values ($\log_2$ of the relative expression) and predicted ones. Prediction is done on the basis of the selected combinations of potential TF binding sites in the promoter regions of the genes. Different colours represent different promoters where AhR sites were found. Red – by two AhR matrices and blue by only one.

Fig. 8. A scheme of proposed network of transcription factors involved in regulation of gene expression under induction by Ah-receptor.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Table 1. Expression levels of different genes after treatment of human or rat hepatocytes with Aroclor 1254, and the LocusLink ID (identifier) of these genes.

Gene	Relative expression level (human)	LocusLink-ID (human)	Relative expression level (rat)	LocusLink-ID (rat)
ERB-B	1.5	2064	3	24337
ERB-A	0.5	7067	3	81812
c-jun	0.6	3725	5	24516
LXR-alpha	0.5	10062	1.8	58852
CYP1A1	400	1543	20	24296
CYP1A2	26	1544	20	24297
CYP1B1	10	1545	40	25426
Ugt1a1	3	54658	10	24861
CAR-beta	0.6	9970	-	-
PPAR-gamma	0.8	5468	1.01	25664
COX-2	0.7	4513	15	26198
ARNT-1	0.7	405	5	25242
CEBP-alpha	3	1050	3	24252
RAR-alpha	0.7	5914	1.5	24705
RXR-alpha	2	6256	1	25271
Gluco-corticoid receptor	2	2908	4	24413
Aldh3a1	60	218	40	25375
pRB	1.5	5925	2	24708
E2F(1)	1	1869	0.8	NW_047658 (RefSeq)
Estrogen receptor-alpha	0.2	2101	0.5	24890
Fos	2	2353	10	314322
SREBP-1c	0.5	836	-	-
MAD2	0.5	4085	-	-
Bax	2.5	581	2.5	24887
Connexin26	1.5	2706	4	-
Connexin32	1.5	2705	4	29584
PCNA	0.9	5111	0.5	25737
CDK2	0.8	1017	15	246381
CDK4	0.6	1019	0.9	94201
Raf-1	3	5894	3	24703
p53	0.8	7157	5	24842
p27	0.8	1027	1.5	83571
p21	0.8	1026	1.5	114851
MDR1	2	19	4	313210
n-ras	0.9	4893	2	24605
c-src	1	1445	20	83805
CyclinD1	0.5	595	4	58919

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Gene	**Relative expression level (human)**	**LocusLink-ID (human)**	**Relative expression level (rat)**	**LocusLink-ID (rat)**
c-myc	0.8	4609	1	24577
Histon deacetylase	0.8	8841	1.5	84578
ICAM1	0.6	3383	2	25464
NQO1	2	1728	6	24314
CYP3A1	-	-	60	286929
CYP3A4	1.7	1576	-	-
CYP3A5	2.6	1577	-	-
CYP3A7	5.3	1551	-	-
Ugt1a3	2.6	54659	1.12	-
Ugt1a4	3.1	54657	0.71	-
Ugt1a5	2.5	54579	0.83	-
Ugt2b4	1.5	7363	-	-
Ugt2b7	2.1	7364	-	-
Ugt2b15	2	7366	-	-
Ugt2b17	1.5	7367	-	-
PXR	0.44	8856	-	-
GSTa1	3	2938	-	-
GSTa2	3	2939	-	-
GSTa3	3	2940	2.02	24423
Cyclooxy-genase 1	R [a]	5742	-	-
Microsomal epoxide hydrolase	2.3	2052	-	-
Alcohol dehydro-genase iso 1	2.6	216	-	-
Bcl-2	R [a]	596	0.7	24224
Bcl-xl	R [a]	598	-	-
BAD	R [a]	572	0.76	64639
Glutathione peroxidase	0.37	2876	0.88	24404
NF-kappaB	0.7	4790	0.67	81736
TGFbeta 1	0.48	7040	1.39	59086
MET	0.14	4233	0.99	-
Apolipo protein 1	0.53	335	-	-
Ubiquitin	4.7	7314	-	-
Hsp27	2	3315	1.36	24471
Hsp70	1.2	3303	0.74	294254
Hsp90	3	3320	0.89	299331
CYP2B1	-	-	12.67	29295

R [a] = reported in the literature to be regulated by PCBs

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Table 2. Over and underrepresented oligonucleotides found in the flanking regions of AhR sites. The first 9 characteristics are "positive", the last four are "negative". The sequences considered were of the length 111bp (11bp AhR site in the center and 50bp flanks).

Olig[a]	from	to[b]	in/out[c]	Utility[d]	FreqY[e]	FreqN[f]	FreqY/FreqN[g]
RDVB	0	15	0	0.60	3.333333	2.313131	1.441048
CNYK	36	71	0	0.84	3.333333	2.121212	1.571429
DYSY	33	87	0	0.68	8.037037	5.717172	1.405771
YRMG	9	57	0	0.61	2.000000	0.838384	2.385542
SVWY	30	42	0	0.56	0.962963	0.252525	3.813334
SBDY	33	39	0	0.73	1.037037	0.171717	6.039216
WHRH	51	63	0	0.69	1.000000	0.222222	4.500000
THDM	48	66	0	0.69	0.888889	0.232323	3.826087
DYVC	58	68	1	0.77	9.407408	6.282828	1.497321
WANW	12	87	0	0.61	1.370370	4.262626	0.321485
WBNR	48	52	0	0.60	0.000000	0.555556	0.000000
HBWG	48	52	0	0.60	0.000000	0.505050	0.000000
CVD	61	62	0	0.60	0.074074	0.565657	0.130952

[a] Olig = the oligonucleotide over- or underrepresented in the sequences (written in the ambiguous one letter code).

[b] from/to = positions of the window.

[c] in/out = oligos are counted in the window (0) or outside the window (1).

[d] Utility = Utility value U (-1 < U < 1) is an indicator of significance of the difference between two distributions of frequencies freqY and freqN. Utility is calculated on the bases of a number of statistical criteria including tests of mean difference, distribution overlapping, normal-likeness and bootstrap tests.

[e] FreqY = frequency of the oligonucleotide in the AhR site sequences (average number of oligonucleotides in the window).

[f] FreqN = frequency of the olig in the background sequences.

[g] FreqY/FreqN = relative frequency in Y vs. N.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Table 3. List of factors the matrices of which were selected by the algorithm in 5' regions of different length.

(-500+100)	(-1000+100)	(-2000+2000)
E2F	NFY	E2F
HNF1	HNF1	HNF1
USF	C/EBP	OCT
NF-kappaB	SP1	GR
YY1	YY1	YY1
	AHR	HNF6

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Table 4. Composite module CM_{AhR} constructed for the (-2000 +2000) set of promoters.

Factor or pair of factors (distance)	ϕ (see equation (4))
E2F	0.105086
OCT1	0.084289
GR	0.077050
YY1	-0.169821
IRF/SRY(50)	0.213636
HNF3/SRY(50)	0.164787
AP1/NF1(100)	0.149481
SP1/MYB(50)	0.138358
GR/HNF1(40)	0.137571
AHR/MYB(50)	0.124308
HNF1/SP1(100)	0.110810
E2F/ROR(100)	-0.115593
AHR/CREB(100)	-0.086788

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Table 5. Summary of sites, site pairs found in promoters (-2000 +2000) and clusters found in 5' regions (-20kb +20kb) of AhR-regulated genes and other genes for transcription factors involved.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Gene	AhR+ context	C/EBP	NF-kappaB	HNF3	E2F	OCT1	GR	HNF4	HNF1/ SP1 (100)	AP1/NF1 (100)	AhR/ Myb (50)	SP1/ Myb (50)	GR/ HNF1 (40)	IRF/ SRY (50)	AhR clusters
Human															
CYP1A1	+								+	+				+	+
LXR						+		+						+	+
CYP1B1	+				+	+	+	+							+
UGT1A1					+				+					+	
CAR-beta	+			+				+		+	+				
COX-2						+		+				+			
CYP3A4						+						+	+	+	
CYP3A5						+			+			+	+	+	
CYP3A7						+						+	+		+
UGT1A3	+					+	+		+	+	+				
UGT1A4						+	+				+		+		
UGT2B4				+		+				+					
C-JUN	+														+
ESR-alpha		+		+		+						+			+
FOS	+	+				+				+				+	
SREBP1		+										+			+
N-RAS		+				+				+					
C-SRC	+											+		+	
UGT2B7				+		+			+	+		+		+	+
UGT2B15						+				+				+	
CAR				+			+		+			+			
ARNT		+				+					+			+	+
GSTA2		+		+		+									
GSTA3		+		+		+		+							
C/EBP-alpha	+										+				
RAR-alpha						+				+				+	
RXR-alpha	+							+			+		+		+
EPHX1	+					+				+			+		
ADH1A		+				+		+	+	+		+		+	
ALDH3A1	+			+		+				+	+	+			+
PRB	+					+	+				+		+	+	
E2F1					+	+				+					+
MAD2	+	+		+		+	+			+	+				+
BCL-2	+						+							+	+
GPX1										+		+			
GJB2	+			+		+								+	
GJB1	+					+			+	+					
PCMA	+	+		+	+	+									+
C-RAS	+							+							
NFKB1	+					+								+	
P53				+		+		+							+
P27	+					+						+			+
P21	+									+					+
UGT1A5						+	+		+	+		+	+	+	
NR3C1 (GR)	+					+								+	+
RAF-1	+			+								+			
UGT2B17	+					+				+		+		+	
NQO1	+	+				+	+			+	+	+			
Hsp90		+		+		+		+						+	

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Gene	AhR+ context	C/EBP	NF-kappaB	HNF3	E2F	OCT1	GR	HNF4	HNF1/ SP1 (100)	AP1/NF1 (100)	AhR/ Myb (50)	SP1/ Myb (50)	GR/ HNF1 (40)	IRF/ SRY (50)	AhR clusters
Human															
Hsp27	+	+		+						+			+		+
UBC	+				+										
Ubiquitin C	+				+									+	
ICAM1	+					+						+			+
ApoA1	+					+		+			+			+	
met_proto-	+	+									+				+
c-myc	+					+				+					
TGFb1	+												+		
CyclinD1	+	+	+		+	+					+			+	+
ERB-A	+	+				+					+				
ERB-B		+				+					+			+	
MDR1						+							+	+	
HDAC3	+			+		+				+	+				+
PXR_1		+		+		+							+		
PTGS1_1	+	+				+					+			+	
BCL-XL_1	+	+				+						+			+
BAX_1	+			+		+			+						
BAD	+	+	+	+							+	+			+
CDK2_1												+			
CDK4_1	+			+		+				+	+				
PPAR_gam		+				+							+		
CYP1A2				+						+		+			+
HNF1-beta						+						+	+	+	
OCT1		+		+		+							+	+	
Rat															
Raf-1		+		+		+				+					+
Pcna		+		+	+	+			+	+				+	+
Cyp1A1		+		+		+			+	+		+		+	+
Cyp1B1	+			+	+	+		+		+					+
Ptgs2	+	+				+				+			+		+
Arnt1		+		+		+					+			+	+
Cebp-a	+	+				+				+	+				
Rar-a		+		+		+							+		+
Rxr-a	+	+		+		+		+						+	+
GR1		+		+	+	+		+	+	+		+			+
Aldh3A1	+					+									+
ICAM1	+		+			+		+	+					+	
n-ras						+								+	
P27	+	+		+		+			+	+		+		+	+
P53	+	+	+	+		+		+		+					
Other genes for transcription factors															
Human															
HNF1-beta	+	+				+						+			+
HNF4-	+	+	+	+		+		+	+			+	+	+	
OCT1	+			+		+			+		+	+		+	+
HNF3-alpha	+					+		+		+		+	+	+	
HNF3-beta	+			+		+			+			+			
Rat															
HNF1-beta						+						+	+	+	
OCT1		.		.		.							.	.	

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Table 6. Composite module E001 that correlates with gene expression in the G4000ExpressUP and G4000ExpressDOWN sets.

ϕ (eq.(2))	core cut-off	cut-off	matrix acc.	matrix ID	κ (eq.(2))
0.480000	**0.755000**	**0.692000**	**M00528**	**V$PPARG_03**	**4**
3.340000	**1.000000**	**0.933000**	**M00139**	**V$AHR_01**	**4**
1.260000	**1.000000**	**0.926000**	**M00235**	**V$AHRARNT_01**	**2**
3.080000	**1.000000**	**0.927000**	**M00639**	**V$HNF6_Q6**	**2**
1.340000	**0.861000**	**0.904000**	**M00340**	**V$ETS2_B**	**2**
-1.040000	0.939000	0.993000	M00492	V$STAT1_02	1
-0.480000	1.000000	0.967000	M00778	V$AHR_Q5*)	2
-0.740000	1.000000	0.916000	M00156	V$RORA1_01	1

*) This matrix is a new AhR matrix constructed in the current study.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Table 7. 71 potential target genes for AhR regulation selected by the highest values of different prediction scores.

Experimentally tested[a]	Confirmed (yes/no)	Gene	CM score (S671)	CMexpress score (expr-s4-8)	LokusLink-Acc.-No.	Chromosomal-Lokalisation
		ASH2L	0,448983	5,348242	9070	chrom=8p11.2
		DED	0,459072	5,329628	26574	chrom=17
		GAS41	0,804663	4,179732	8089	chrom=12
		ZFR	0,312676	4,034527	51663	chrrom=5p15.2
+	yes	SLC3A2	0,699764	3,984838	6520	chrom=11q12-q22
		CLGN	0,601957	3,939733	1047	chrom=4
		DKFZp434D177	0,403696	3,906978	84224	chrom=1
		FOXM1	0,88978	3,887026	2305	chrom=12p13
+	no	SNRP70	0,70229	3,601697	6625	chrom=19q13.3-qter
		CNTNAP1	0,310554	3,436642	8506	chrom=17q21
+	yes	M6PR	0,363536	3,427701	4074	chrom=12
+	yes	FLJ20533	0,456768	3,320127	54968	chrom=8
		DAP	0,684227	3,145814	1611	chrom=5p15.2
		X102	0,528255	3,02357	63969	chrom=15
		POLA2	0,385477	3,019933	23649	chrom=11
		SLC38A4	0,342434	3,009857	55089	chrom=12q13
		TES	0,326831	2,953249	26136	chrom=7q31.2
		EIF2C2	0,383712	2,942913	27161	chrom=8q24
		FBXO4	0,629685	2,942402	26272	chrom=5p12
		KIAA0144	0,396641	2,895884	9898	chrom=1
+	yes	LOC90806	0,397701	2,88055	90806	chrom=1
		NPR3	0,313687	2,848264	4883	chrom=5p14-p13
		IFNAR1	0,490227	2,772589	3454	chrom=21q22.1
		LOC51134	0,478122	2,582767	51134	chrom=12
		LOC57116	0,448834	2,437427	57116	chrom=1
+	no	DPM1	0,311498	2,374712	8813	chrom=20q13.10
		FLJ22626	0,491389	2,319131	79576	chrom=X
		STC2	0,31078	2,226394	8614	chrom=5
		DKFZP564D116	0,312825	2,224058	26123	chrom=10
		DKFZP586M0122	0,318268	2,158908	25885	chrom=2
		TACSTD2	0,715578	2,132282	4070	chrom=1p32
		RNF13	0,368514	2,117528	11342	chrom=3
		DKFZp434F1819	0,320963	1,977744	84209	chrom=17
		LOC91344	0,369804	1,97066	91344	chrom=12
		MAPRE2	0,440917	1,95473	10982	chrom=18
		EXTL2	0,362347	1,948633	2135	chrom=1p11-p12
		EVC	0,536331	1,943895	2121	chrom=4p16
		MST4	0,841598	1,923181	51765	chrom=X
		PEMT	0,30361	1,913363	10400	chrom=17p11.2
+	yes	SYPL	0,306396	1,89363	6856	chrom=7
		PTPN4	0,396457	1,827322	5775	chrom=2
+	yes[b]	IGFBP1	0,878297	1,825774	3484	chrom=7p13-p12
		AMPD1	0,318563	1,804475	270	chrom=1p13
		MGC10753	0,324073	1,797132	51322	chrom=10
+	no	GAD1	0,303744	1,751589	2571	chrom=2q31
+	yes	FLJ10539	0,338895	1,744145	55164	chrom=3
		MRP63	0,701134	1,648883	78988	chrom=13p11.1-q11
+	yes	SLC1A5	0,386735	1,552027	6510	chrom=19q13.3
		LOC92345	0,310469	1,541319	92345	chrom=4
		HSA249128	0,674854	1,448967	54765	chrom=11
		LOC90480	0,394572	1,428735	90480	chrom=19
		LOC112936	0,512638	1,41761	112936	chrom=11
		KIAA0870	0,320461	1,408811	22898	chrom=8
		TUBGCP3	0,321219	1,406862	10426	chrom=13q34
		LOC90488	0,341424	1,397628	90488	chrom=12
		ERG-1	0,465449	1,384104	50624	chrom=10
		PIP5K1A	0,320143	1,341131	8394	chrom=1q22-q24
		FLJ11937	0,407384	1,18203	64780	chrom=6
		PTP4A1	0,400378	1,131907	7803	chrom=6q12
		MGC3329	0,784041	1,117763	79066	chrom=17
+	yes	NR1H3	0,352805	1,081247	10062	chrom=6
		REV3L	0,62078	1,074046	5980	chrom=6q22

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

		KIF1C	0,330711	1,073282	10749	chrom=17p
		CKN1	0,301125	1,038971	1161	chrom=5q12.1
		PKIA	0,405713	1,010471	5569	chrom=8
		FLJ10154	0,331306	-1,142423	55082	chrom=13
		TXNL	0,634063	-1,205094	9352	chrom=18q21.1-18q21.32
+	yes	FGFR1	0,711906	-1,222143	2260	chrom=8p12
		MAPK8IP2	0,338084	-1,302075	23542	chrom=22q13
		SDBCAG84	0,454627	-2,334489	51614	chrom=20pter-q12

[a] by Microarray-Technology

[b] experimentally tested: IGBP1 = unchanged, IGBP2 = repressed, IGBP5 = induced

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Table 8. Results of testing CMFinder on simulated data.

# of matrices	Frequency of implanting of the CM (% of sequences)			
	50	70	100	implanted
2	+/- (**AhR**, GATA)	+	+	**AhR, AP-1**
4	+/- (**AhR, C/EBP, OCT,** HNF3)	+	+	**AhR, OCT, C/EBP, AP-1**
6	-/+ (**AhR, C/EBP**, HNF4, HNF3A, ROR)	+/- (**AhR, C/EBP**, **AP-1**, GATA, ROR, **HNF1**, COUP)	+/- (**AhR, OCT, C/EBP**, ROR, **NF1**, **HNF1**)	**AhR, OCT, C/EBP, AP-1, NF1, HNF1**

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

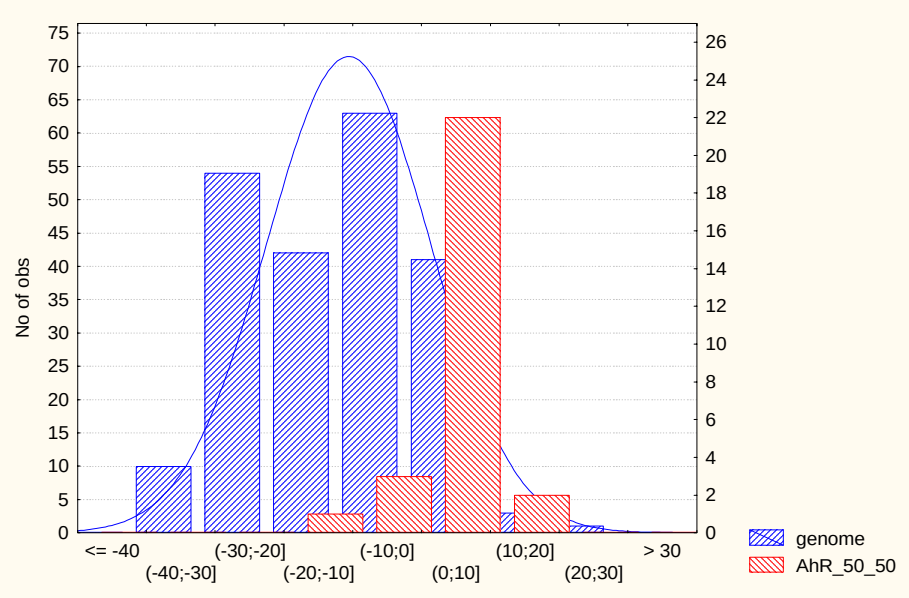
Figure 1

A	2	1	4	0	0	0	0	0	9	4	7
C	11	2	6	0	25	0	0	0	9	7	7
G	6	5	5	25	0	25	0	25	2	11	5
T	6	17	10	0	0	0	25	0	5	3	6
Consensus	**N**	**T**	**N**	**G**	**C**	**G**	**T**	**G**	**N**	**N**	**N**

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

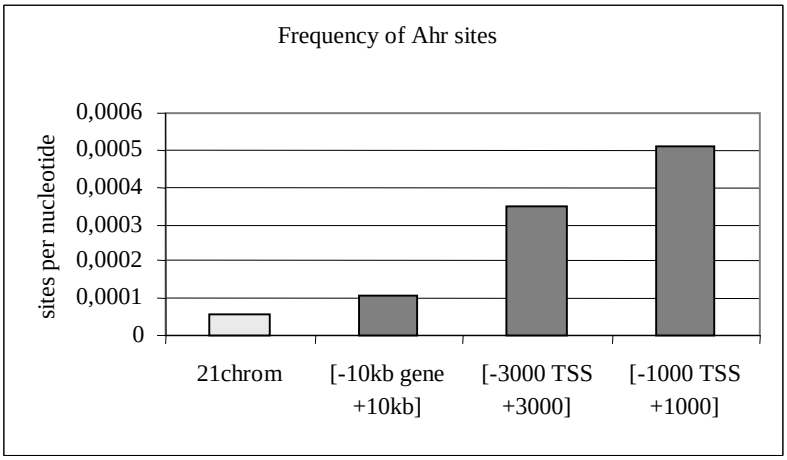
Figure 2



Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Figure 3

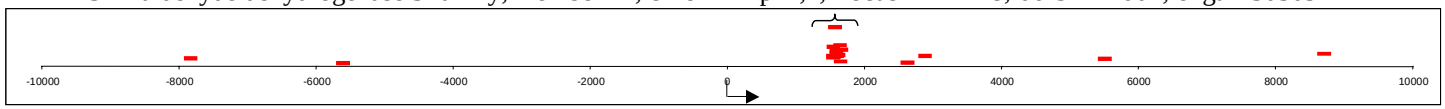


Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

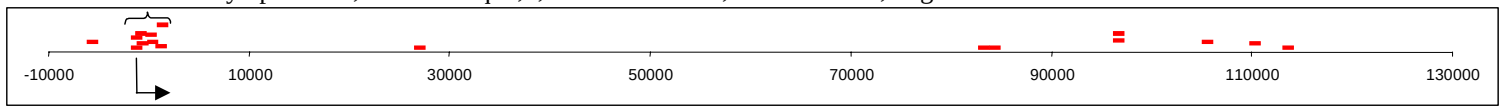
Figure 4

Potential Ahr binding sites found in the human genes (-10kb+gene+10kb)

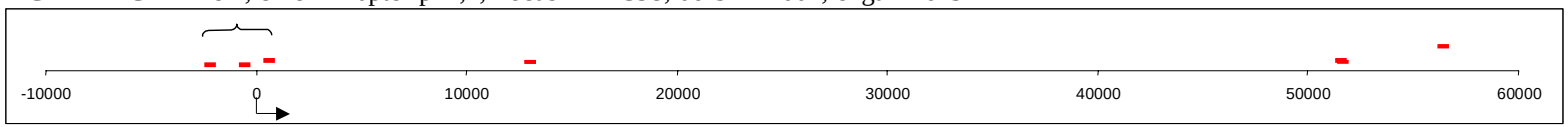
ALDH3A1 aldehyde dehydrogenase 3 family, memberA1; chrom=17p11,2; LocusLink=218; 06-SEP-2002;length=30309



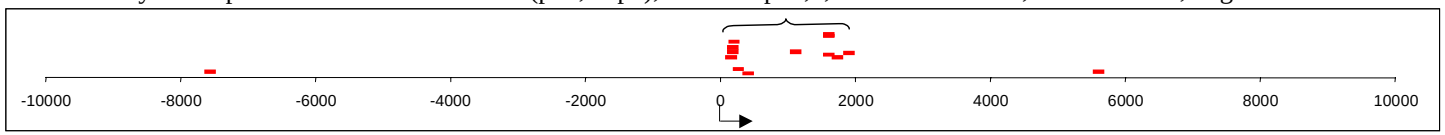
BCL2 B-cell CLL/lymphoma 2; chrom=18q21,3; LocusLink=596; 06-SEP-2002;length=148720



BCL2L1 BCL2-like 1; chrom=20pter-p12,1; LocusLink=598; 06-SEP-2002;length=76492



CDKN1A cyclin-dependent kinase inhibitor 1A (p21, Cip1); chrom=6p21,1; LocusLink=1026; 06-SEP-2002;length=29586

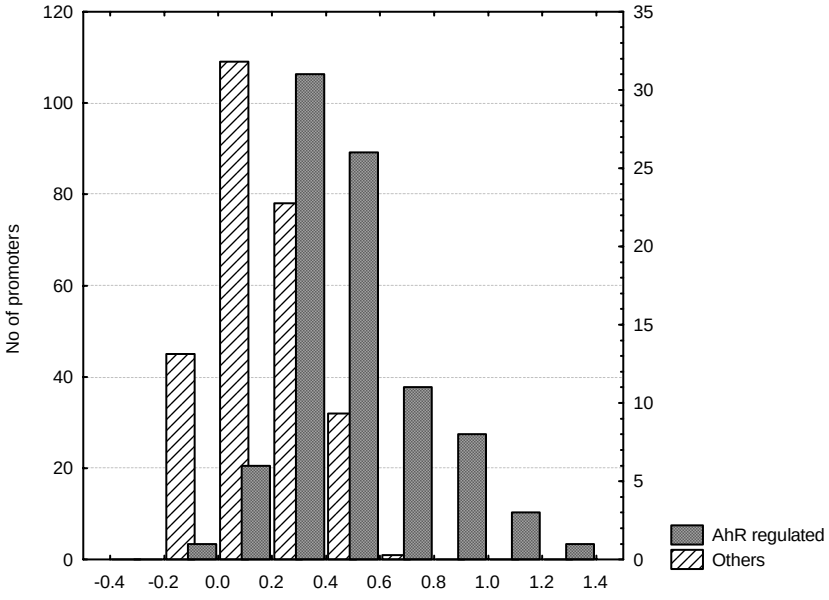


Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

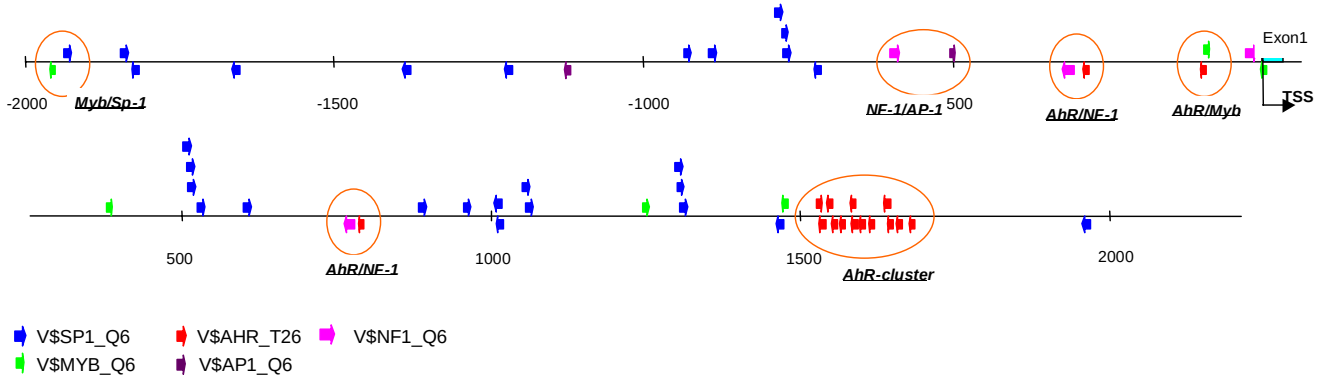
Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Figure 5



Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Figure 6a

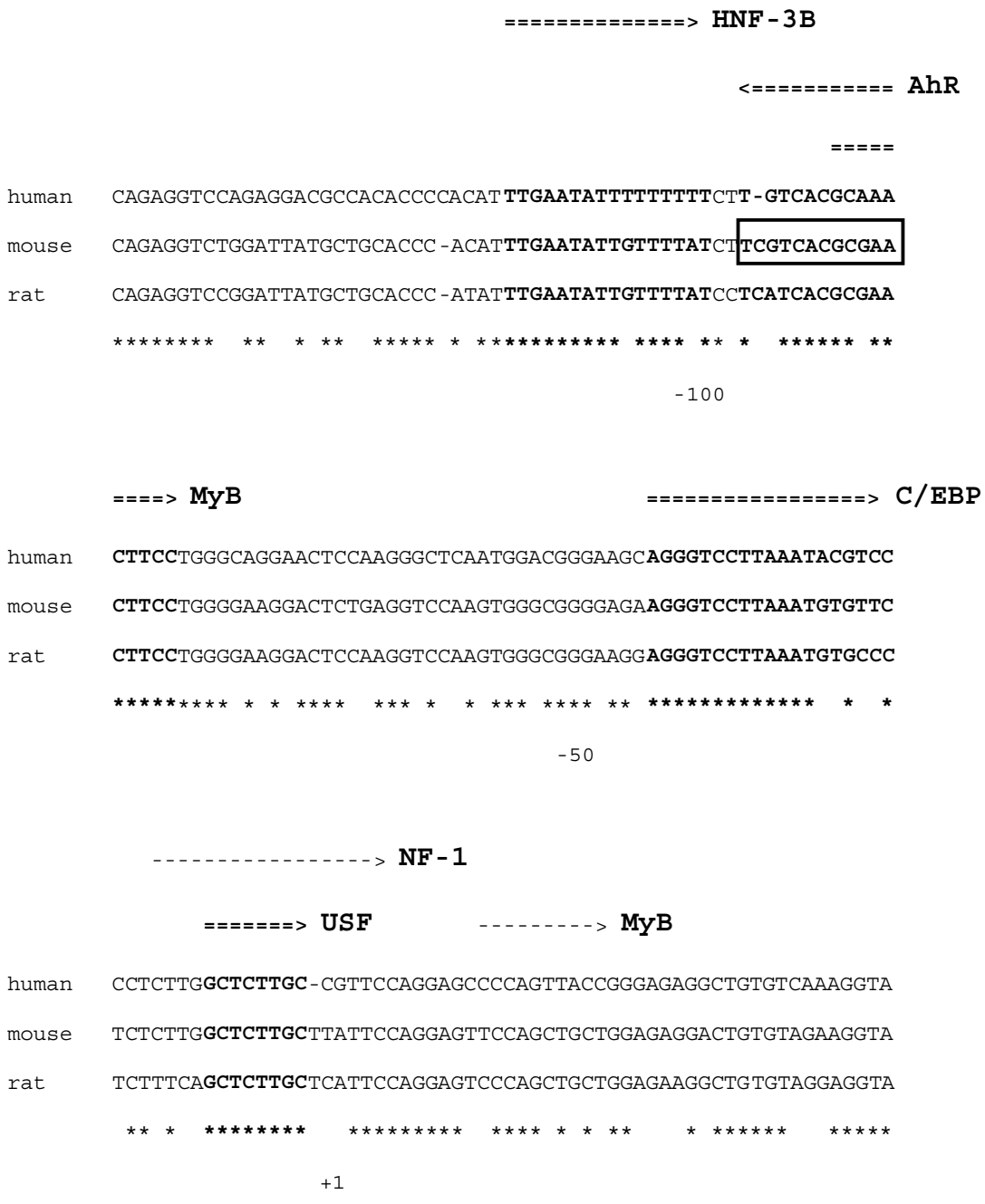


Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Figure 6b



Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

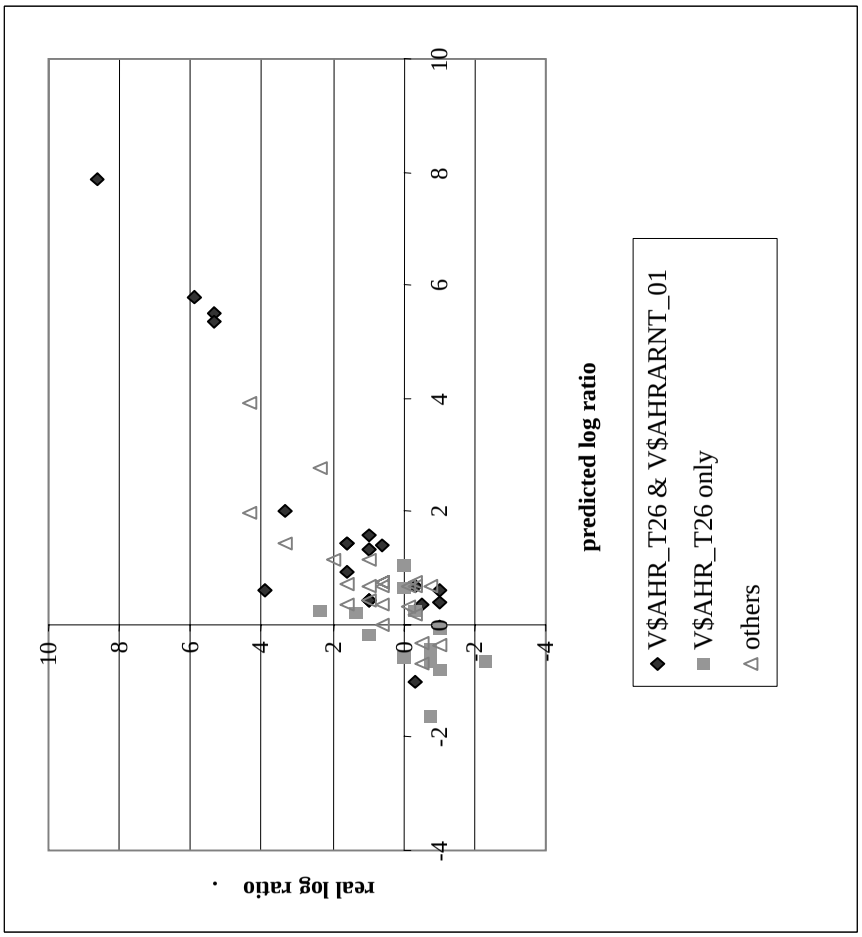
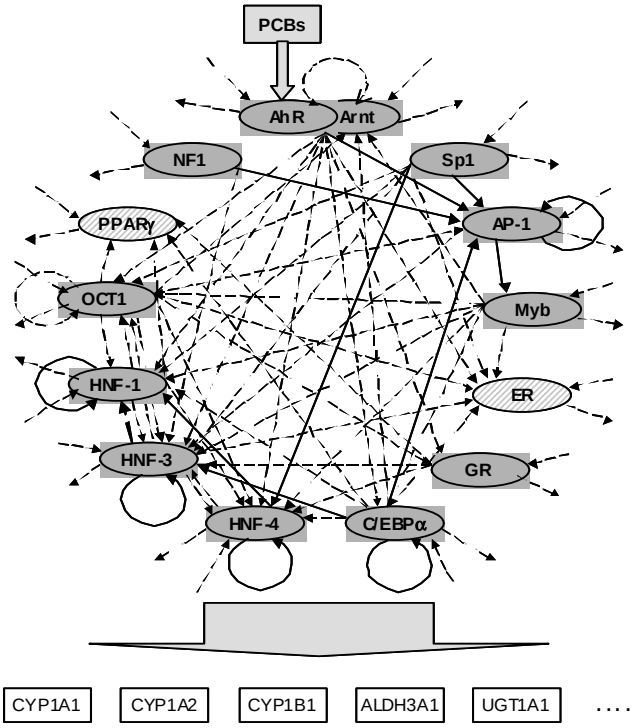


Figure 7

Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024

Molecular Pharmacology Fast Forward. Published on September 1, 2004 as DOI: 10.1124/mol.104.001677
This article has not been copyedited and formatted. The final version may differ from this version.

Figure 8



Downloaded from molpharm.aspetjournals.org at ASPET Journals on April 19, 2024