A Predictive Model for Regulatory Sequences Directing Liver-Specific Transcription

William Krivan[1] and Wyeth W. Wasserman[2]

Bioinformatics Unit, Center for Genomics and Bioinformatics, Karolinska Institutet, 17177 Stockholm, Sweden

The identification and interpretation of the regulatory signals within the human genome remain among the greatest goals and most difficult challenges in genome analysis. The ability to predict the temporal and spatial control of transcription is likely to require a combination of methods to address the contribution of sequence-specific signals, protein–protein interactions and chromatin structure. We present here a new procedure to identify clusters of transcription factor binding sites characteristic of sequence modules experimentally verified to direct transcription selectively to liver cells. This algorithm is sufficiently specific to identify known regulatory sequences in genes selectively expressed in liver, promising acceleration of experimental promoter analysis. In combination with phylogenetic footprinting, this improvement in the specificity of predictions is sufficient to motivate a scan of the human genome. Potential regulatory modules were identified in orthologous human and rodent genomic sequences containing both known and uncharacterized genes.

[Supplementary data and the submission of sequences for analysis are available at http://www.cgb.ki.se/krivan/liver/liver.html.]

Unprecedented insights into the mechanisms of gene regulation are promised from the compilation of eukaryotic genomes and the subsequent application of sequence analysis algorithms. Present-day bioinformatics, however, is largely restricted to phenomenological approaches. Given the limited knowledge about molecular interactions between regulatory proteins and DNA, computational biologists face a daunting challenge in studying transcription—the physics of gene regulation is not yet understood (Claverie 2000).

Experimental progress has been made in understanding biochemical mechanisms governing transcription, including enumeration of the components and characterization of their interactions (Roeder 1996). Recent advances have addressed the complex architectures of the regulatory regions (enhancers and repressors) that mediate alterations of the transcription rate (Blackwood and Kadonaga 1998). Detailed molecular analyses of genes from multicellular eukaryotes have established that regulatory regions can be viewed as modules of multiple transcription factor (TF) binding sites that act in combination to confer expression patterns upon genes (Yuh et al. 1998; Loots et al. 2000).

We present here a procedure to identify regulatory regions on the basis of characteristics of sequence modules experimentally verified to direct transcription selectively to liver cells. The method is extended to a genomic scale by incorporating cross-species sequence comparisons [phylogenetic footprinting (Fickett and Wasserman 2000; Wasserman et al. 2000)] between the available human and rodent genomic sequences. The combination of these two methods accurately identifies a set of experimentally documented, as well as potential, regulatory modules.

RESULTS

Collection of Experimental Data

To develop a model for transcriptional regulation in liver, it is necessary to compile a representative collection of experimentally defined liver-selective regulatory modules. By selective we refer to genes that are either expressed exclusively in liver (specific) or in a small number of tissues including liver.

Figure 1 illustrates the regulatory regions that we used as the foundation of our model, all of which are verified experimentally to direct gene transcription in liver cells. As with a previously described skeletal muscle regulatory region collection (Wasserman and Fickett 1998), it was observed that most extensively studied liver regulatory modules are shorter than 200 bp in length. Typically, regulatory regions contain more than one experimentally verified TF-binding site. An important observation from the collection of experimental data is the following: liver-specific gene expression is governed by the combined action of a small set of TFs, primarily HNF-1, HNF-3, HNF-4, and C/EBP. This conclusion is consistent with established biological knowledge (Ktistaki and Talianidis 1997; Tronche et al. 1997).

It is instructive to have a closer look at the different roles of the critical TFs during the development of the liver; HNF-3 is strongly linked to transcription of genes expressed early in liver development, whereas HNF-1 and HNF-4 are linked to expression of genes in the mature liver (Darlington 1999; Locker 2000). As the genes in our collection were selected for expression in the mature liver (e.g., in differentiated hepatocytes), it is anticipated that HNF-1 and HNF-4 may have a prominent role in the transcriptional regulation of these genes. This necessarily orients the module detection algorithm toward the detection of regulatory regions of genes that are activated late in liver development.

[1]Present address: ZymoGenetics Inc., 1201 Eastlake Avenue East, Seattle, WA 98102.
[2]Corresponding author.
E-MAIL wyeth.wasserman@cgb.ki.se; FAX 46 8 337412.
Article published on-line before print: *Genome Res.*, 10.1101/gr. 180601. Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.180601.

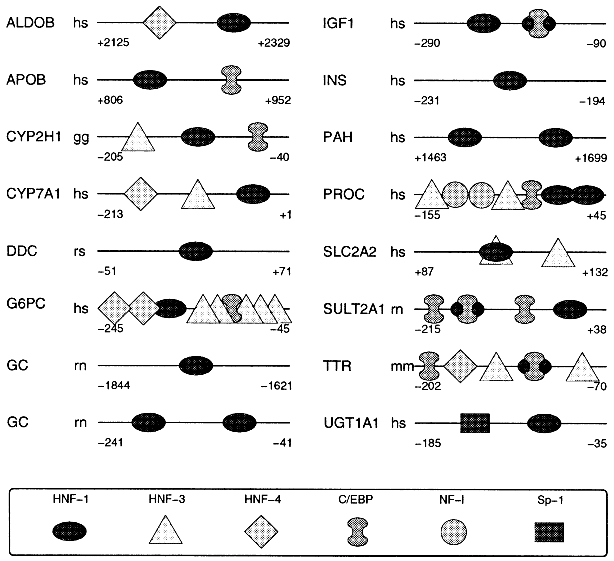

Figure 1 Clusters of TF-binding sites that direct liver-specific transcription. The critical factors are HNF-1, HNF-3, HNF-4, and C/EBP. For human genes, the names denote the approved HUGO symbols, otherwise they are given by the corresponding name of the human ortholog, if available. The numbers denote positions relative to the TSS.

Position Weight Matrices

Before addressing the problem of describing clusters of binding sites that constitute regulatory regions, it is necessary to develop models for the binding of the individual critical TF. For this purpose, we recorded experimentally determined TF-binding sites for the principal transcription factors. To build robust and descriptive binding profiles, we obtained sites from genes with diverse expression profiles. Although a portion of the collected sites was present in the regulatory regions from the collection described above, most experimentally defined sites were from a larger set of genes with different patterns of expression (for instance, genes expressed early in liver development).

From the binding-site collection, we generated position-weight matrices [PWMs (Fickett 1996a), sometimes also called motifs or profiles]. In the best implementations, such models have been shown to produce quantitative predictions analogous to DNA-binding energies (Stormo and Fields 1998), which are highly predictive of in vitro protein–DNA interactions (Tronche et al. 1997). To generate the models, known TF-binding sites are first aligned to generate a count matrix. The subsequent step consists of a log-odds rescaling of the count matrices and is described below. There are two principal difficulties one has to face when constructing matrices, compiling a sufficiently large number of binding sites and low-binding specificity of some transcription factors. For the four factors under study, a sufficient number of binding sites has been verified experimentally. However, the binding sites for C/EBP are diverse in sequence, posing a significant challenge for multiple sequence-alignment algorithms. Whereas a published count matrix was used initially to model the binding of C/EBP (Johnson and Williams 1994), we meanwhile identified several tools capable of producing alignments of the collected sites consistent with the published descriptions of the binding specificity. Sequence logos for the individual matrices are presented in Figure 2.

Toward a Predictive Model

Although PWMs adequately describe the binding specificity of the individual transcription factors, the predictions produced with the models contain a large number of false sites lacking any experimental evidence; the typical frequency of predicted sites in genomic sequence is between 1/(10,000 bp) and 1/(100 bp), depending on the selected score threshold. This phenomenon is understood in terms of the well-established relation between the information content I of a TF-binding site and the predicted site frequency ν in the genome (Schneider et al. 1986), given by $I = \log_2(1/\nu)$. Because of the high rate of false positives, predictions based solely on PWMs are of little or no practical use for the identification of binding sites with functional roles in vivo. The fact that the binding specificity of individual TFs is too low to provide the

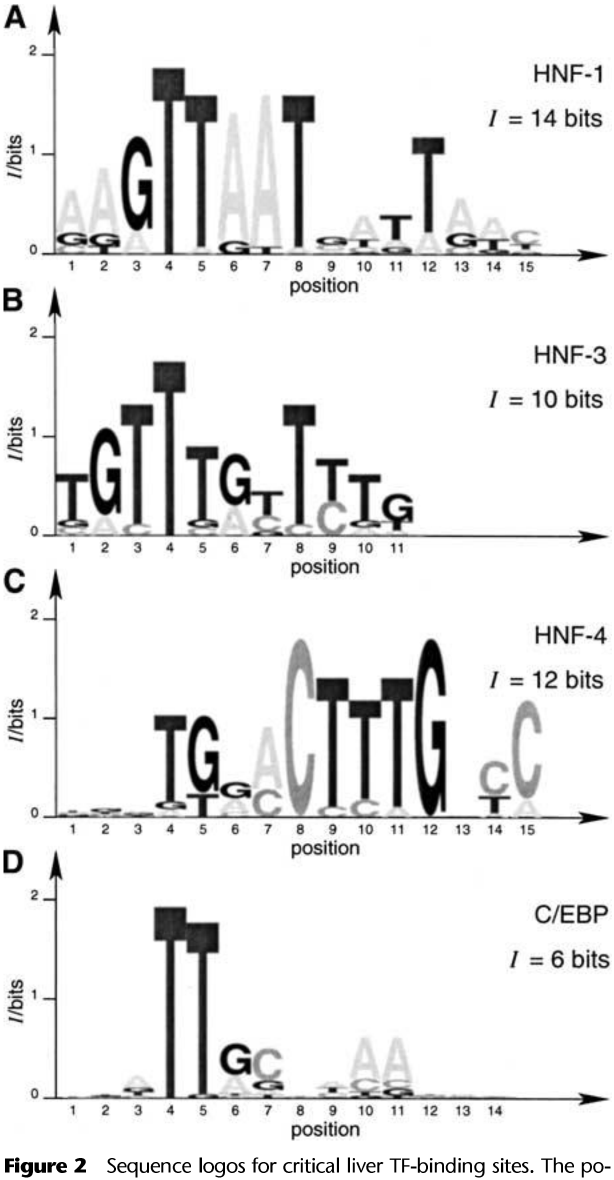

Figure 2 Sequence logos for critical liver TF-binding sites. The position-specific information content is plotted in bits along the ordinate. The low total information content of HNF-3 and C/EBP reflects the low binding specificity of these two factors.

basis for the modeling of regulatory mechanisms is consistent with the biochemical observation that groups of TFs rather than single factors are required for the function of regulatory regions. Thus, one can try to construct a model with a considerably lower rate of false positives by combining the scores of the individual PWMs in a well-defined way.

Logistic Regression Analysis

We have chosen multivariate logistic regression analysis (LRA) as the mathematical approach to combining the scores of the individual PWMs, following the methods described previously for a model of skeletal muscle regulatory regions (Wasserman and Fickett 1998). Logistic regression is a probabilistic method that is used to classify objects into two distinct sets (Hosmer and Lemeshow 1989). The two distinct sets are represented by a binary outcome variable y. The variable y, in turn, is modeled by a function $\pi \equiv \pi(x)$, in which x is a vector of independent variables. In our case, the objects are segments of nucleotide sequence and the two sets are given by segments that are liver-specific regulatory regions (positives) and those that are not (negatives). Each of the components of the four-dimensional vector x is associated with a PWM for an individual TF; the numerical value of each vector component is given by the maximum score produced by the PWM analysis of each subsegment within the 200-bp DNA segment constituting the object. The positive training set consisted of the 16 nonorthologous regulatory regions shown in Figure 1. The negative training set consisted of 1804 randomly selected human genomic sequences [as reported previously (Wasserman and Fickett 1998)]. The logistic regression coefficients generated for the liver model are given in Table 1. Note that the P values for the coefficients associated with HNF-3 and C/EBP are unconvincing, and the numerical values of the coefficients themselves are about a factor of six smaller than those for HNF-1 and HNF-4. Consequently, the contributions of HNF-3 and C/EBP to the current model are minimal. For biological clarity, they were retained in the model, although a purely statistical argument could be proffered for their exclusion.

Assessing the Model

Several properties of the model must be explored. First, a suitable score threshold must be selected for the classification of sequences into positive and negative categories. Second, the performance of the model must be defined in terms of specificity and sensitivity. Finally, the robustness of the model must be examined, particularly with regard to issues of circularity.

The definition of a suitable score threshold is based on balancing performance specificity and sensitivity. We identified a threshold for the liver model to allow, on average, one predicted region per gene, with the length of a gene defined as 35,000 bp. Specifically, a threshold of 0.25 was determined, which produces a prediction frequency of one module per 35,000 bp. We screened the liver-specific training set and observed that 62% (10 of 16) of the positive training sequences were correctly classified above the threshold. To assess the validity of this observation, a cross-validation/jack-knife procedure was performed. In these analyses, each positive LRA training sequence was removed from the training data and scored by the output model. All 10 sequences that were correctly classified when included in the training set were still correctly classified when removed from the training set. To address the issue of circularity in the construction of the weight matrices, we performed TF-binding site jackknife tests, leaving each experimentally verified binding site from the regulatory regions out of the PWM model construction. Compared with the LRA training jackknife test described above, the more complete jackknife results showed slightly higher variation between the different models. For the regions that contained HNF-1 or HNF-4 sites, the number of correctly classified training regions was 9 ± 1. This finding indicates that the LRA modeling process is sensitive to changes in the PWMs, or, from a different viewpoint, that the modeling of the binding of the individual transcription factors would benefit from a greater number of experimentally verified sites. However, as indicated earlier, only a small fraction of the binding sites used to generate the PWMs were from the regions used to train the LRA model. To determine the sensitivity of the LRA model to the circular use of TFBSs from the target regions, complete jackknife analyses were performed in which all transcription factors from each regulatory region were removed from PWM construction and the region itself removed from the LRA training process. Two regulatory regions were no longer classified correctly in the comprehensive analysis, CYP7A1 and SULT2A1 (both due to alterations in the HNF-1 profile). Thus, half of the regulatory regions were detected in all studies without any circularity conflicts.

To ascertain the stability of the model and suggest a minimal number of training sequences for future studies, the numerical values of the LRA coefficients were determined as a function of the number of positive training sequences. Results are shown in Figure 3. The coefficients for the intercept, HNF-1, and HNF-4 are shown in logarithmic plots in A, B, and D. A non-logarithmic representation was chosen for the more variable coefficients of HNF-3 and C/EBP in C and E. Inclusion of 11 or more randomly selected positive training sequences produced stable coefficients for all matrix models (on the basis of coefficient deviation from 15 trials for each training set size).

Table 1. The Logistic Regression Coefficients of the Liver Model

Variable	Coefficient	Wald Chi-square value	P-value
Intercept	−23.23	39.10	1.25×10^{-9}
HNF-1	0.64	38.05	6.89×10^{-10}
HNF-3	0.10	0.47	0.49
HNF-4	0.58	18.61	1.61×10^{-5}
C/EBP	0.12	0.25	0.62

Phylogenetic Footprinting

The selection of clusters of binding sites with the LRA model allows a reduction of the prediction rate from one TF-binding site per 250 bp to one liver module per 35,000 bp. Despite this improvement by two orders of magnitude, with regard to scanning large pieces of genomic sequence, the results are not sufficiently specific to motivate human inspection. For a search of the human genome, the expected number of positives is on the order of tens of thousands. Therefore, it is highly desirable to complement the use of the liver module model (LMM) with an independent method that is consistent with known biology and likely to improve prediction specificity. The LMM is based on one characteristic of regulatory regions, namely the clustering of patterns representing TF-binding sites. Another property of regulatory regions, ob-

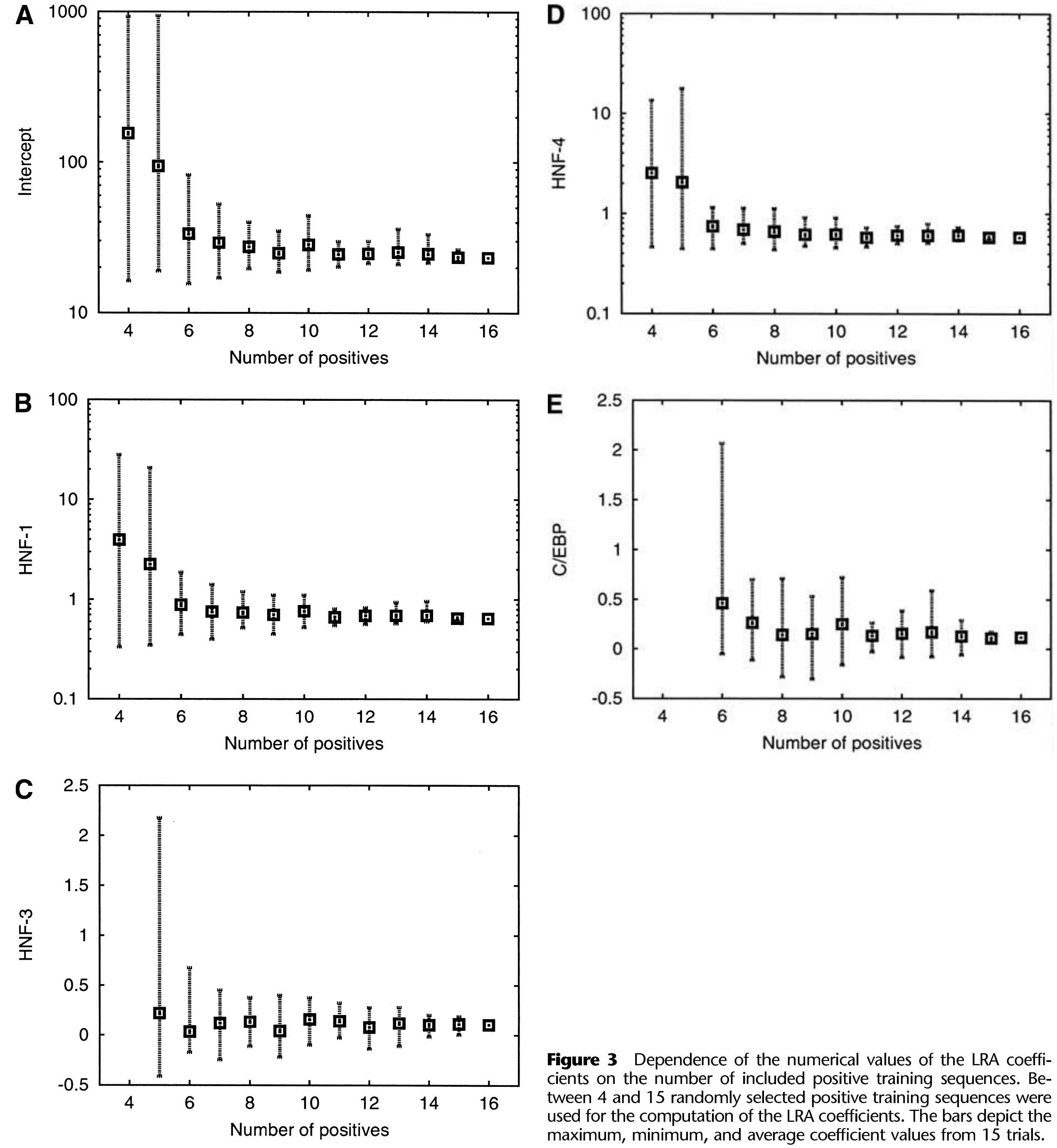

Figure 3 Dependence of the numerical values of the LRA coefficients on the number of included positive training sequences. Between 4 and 15 randomly selected positive training sequences were used for the computation of the LRA coefficients. The bars depict the maximum, minimum, and average coefficient values from 15 trials.

served by multiple groups, is their tendency to be more strongly conserved over the course of evolution than regions lacking sequence-specific function. This differential conservation can be quantified to reveal patterns of conservation that have been called phylogenetic footprints (for review, see Fickett and Wasserman 2000). A conservation-based filter provides an independent means of selecting predicted regulatory modules likely to have sequence-specific functions.

To confirm the complementarity of the LMM and conservation measures, orthologous pairs of liver genes were analyzed (Fig. 4). In all cases, the liver model correctly classifies the documented liver regulatory regions. Furthermore, the regulatory regions exhibit a high level of sequence conservation. Figure 4A and B address sequences contained in the training set. Figure 4A shows CYP7A1 (cholesterol 7α-hydroxylase), in which there is a documented regulatory region upstream of the first exon at ~2000 (Cooper et al. 1997), and additional annotated exons are located in the vicinity of positions 3700 and 4800. Figure 4B shows IGF1 (insulin-like growth factor 1), in which there is a significant sequence similarity in the documented regulatory region (Nolten et al. 1995) at ~1500 and in the adjacent exon downstream. The LMM score peaks in the documented regulatory region. Further upstream, at position 700, the LMM and conservation scores suggest the presence of a regulatory region. To our knowledge, there exists no documented regulatory region in that portion of the promoter.

Screening of Genomic Sequence

To enable a computational screen for regulatory regions on a genome scale, a procedure was implemented to identify regulatory regions conserved between orthologous human and rodent genomic sequences. A sequence-specific conservation is observed for between 15% and 25% of the human genome in comparison with rodents (Jareborg et al. 1999; Stojanovic et al. 1999), the use of phylogenetic footprinting will not eliminate false-positive predictions. However, the elimination of nonconserved regulatory regions promised increased specificity in the set of predictions. To assess the contribution of sequence conservation, a scan was conducted of long genomic sequences.

On the basis of a systematic procedure (Fig. 5), 147 unique potential human liver regulatory regions were identified. Fourteen regions were adjacent to described genes (Table 2), including four from the LMM training set (Table 2A: CYP7A1, G6PC, IGF1, PAH). Of the remaining 12 regulatory regions from the original training set (1) four sequences were never detected by the LMM (as indicated above, INS, SLC2A2, PROC, and TTR), (2) there were no orthologous sequences for three (ALDOB, APOB, and CYP2H1), and (3) five human–rodent regulatory region pairs are not sufficiently similar for detection at the applied blast similarity threshold. The regions and associated *E* values are DDC (*E* = 0.008), the two GC regions (*E* = 0.5, 0.03), SULT2A1 (*E* > 1), and UGT1A1 (*E* > 1). Sequence conservation of regulatory regions is not universal, although the function may be retained. For instance, although UGT1A1 has highly conserved HNF-1 sites, there are pronounced differences between the mouse and human promoters. Certain polymorphisms in the human regulatory region are the cause for Gilbert's syndrome (Bernard et al. 1999), a relatively common liver disorder, suggesting a reduced pressure for sequence conservation in this region.

Four detected modules (Table 2B) are documented as liver regulatory regions in the literature, but were not in our original training collection [FGA (Hu et al. 1995), HNF1 (Kaisaki et al. 1997), IGFBP1 (Suwanichkul et al. 1990), and MTP (Hagan et al. 1994)]. The analysis of FGA (α-fibrinogen) is presented in Figure 4C. There is a strong conservation signal in the documented regulatory region around 13,000, as detected with the LMM.

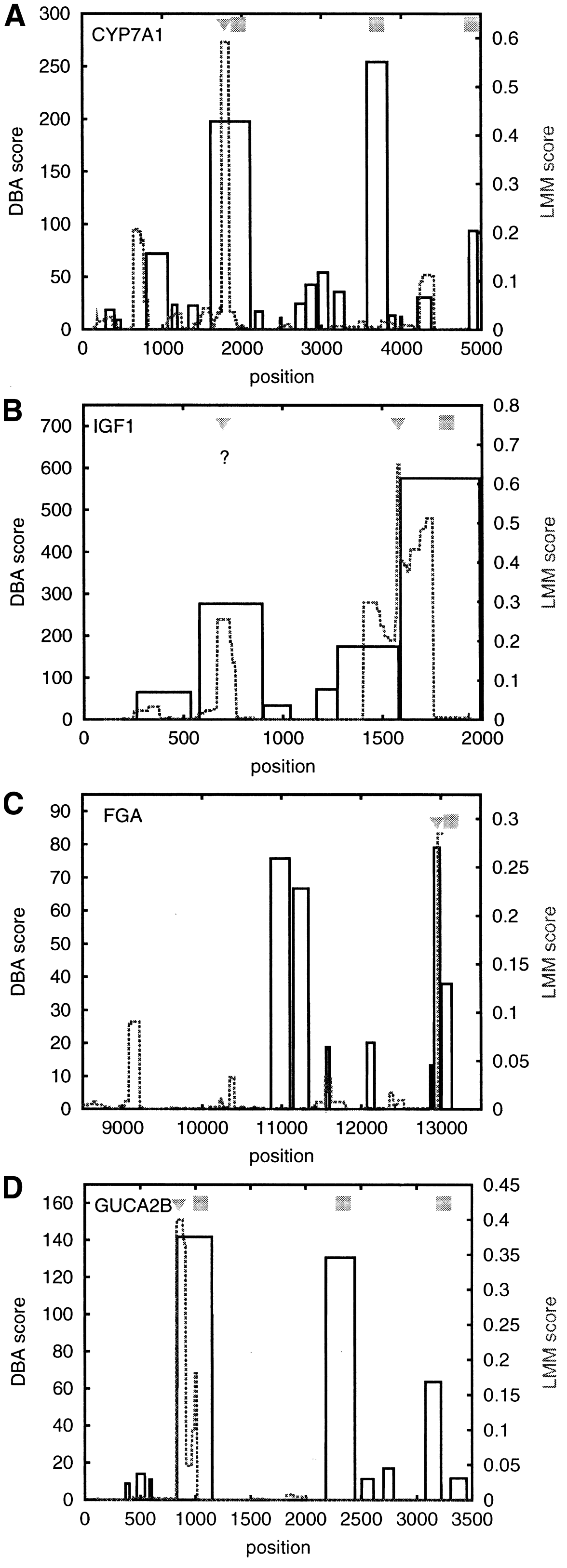

Figure 4 Combining the liver model with phylogenetic footprinting for selected sequences. The sequence similarity between human and rodent sequences as determined with DBA (Jareborg et al. 1999) is shown as a solid black line. The position-dependent LMM score of the human sequence centered with respect to a 200-bp window is shown as a broken gray line. Documented regulatory regions, depicted by triangles, are characterized by strong score for both DBA and LMM. Boxes show annotated exons that possess a high level of cross species conservation, reflected by a high DBA score. The position with respect to the human sequence is shown along the abscissa. Fig. 4A addresses CYP7A1 (human accession L13460, rat U01962). Fig. 4B addresses IGF1 (human S85346 [identical with M12659], mouse Y18062).

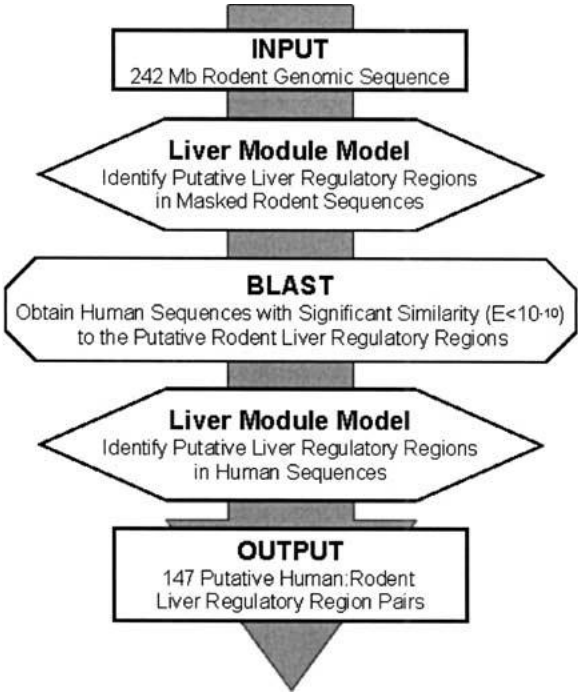
INPUT
242 Mb Rodent Genomic Sequence
Liver Module Model
Identify Putative Liver Regulatory Regions in Masked Rodent Sequences
BLAST
Obtain Human Sequences with Significant Similarity (E<10-10) to the Putative Rodent Liver Regulatory Regions
Liver Module Model
Identify Putative Liver Regulatory Regions in Human Sequences
OUTPUT
147 Putative Human:Rodent Liver Regulatory Region Pairs

Figure 5 Combining the liver model with phylogenetic footprinting on a quasi-genomic scale. See text for explanation.

Two regions (Table 2C) are present in genes known to be selectively expressed in liver, although their role in liver-specific expression has not been experimentally proven. The detected regulatory region of the murine complement factor B gene (BF) is documented (Garnier et al. 1996), but liver specifcity has not been examined. The *MIG6* gene (also called Gene 33) listed in Table 2 is a broadly expressed gene, however, the gene is specifically up-regulated in liver cells by insulin (Melendez et al. 2000).

Four examples for regions adjacent to genes not believed to be selectively expressed in liver are shown in Table 2D. The gene coding for the fatty acid binding protein (FABP2) is specifically expressed in the intestine. The captured segments are located in the promoter regions. The result for *GUCA2B* (uroguanylin), a gene primarily expressed in the intestine, is illustrated in Figure 4D. There is a documented regulatory region in the proximal promoter—approximately at position 800 in the plot—that contains a verified HNF-1 binding site (Whitaker et al. 1997). This serves as an example that the liver TF can play functional roles in other tissues in partnership with other TFs (Cereghini 1996). Functional roles of TFs in multiple contexts can lead to spurious predictions of our model. The region detected in the human HOXA4 sequence coincides with an annotated upstream regulatory region (Doerksen et al. 1996). HOXA4 is specifically expressed in the embryonic nervous system. *SLC34A1* is the human solute carrier family 34 (sodium phosphate), member 1 gene and is specifically expressed in the kidney. For both human and mouse, the detected modules lie within the promoter of SLC34A1.

Most of the remaining sequences were identified in unannotated genomic sequences and have not been mapped to specific genes. The positions of these regions have been posted on the internet site for this paper. Regions mapping onto known coding exons (nine occurrences) have been excluded, as the conservation measure is uninformative for these cases.

DISCUSSION

Accurate detection of regulatory sequences is a difficult challenge requiring a combination of methods. We have constructed a predictive model for liver gene regulatory regions to detect clusters of binding sites for TFs associated with liver-specific transcription. By combining our liver model with phylogenetic footprinting, we have performed a successful genome-wide screen for regulatory regions.

A limitation of the current model is the focus on genes that are mainly expressed in the mature liver, thus, future work must address modeling the regulatory mechanisms of

Table 2. Putative Liver Modules

		LMM score			Human		Rodent		
	gene	human	rodent	BLAST *E*-value	acc. #	begin/end	species	acc. #	begin/end
(a)	CYP7A1	0.59	0.35	5.1×10^{-14}	L04629	278/577	cg	L04690	1303/1602
	G6PC	0.97	0.63	2.4×10^{-16}	AF051355	3530/3829	mm	U91573	293/592
	G6PC	0.97	0.33	1.1×10^{-18}	AF051355	3506/3805	rn	U57552	1180/1479
	IGF1	0.30	0.39	1.3×10^{-14}	M12659	1266/1565	mm	Y18062	1223/1522
	PAH	0.30	0.90	2.0×10^{-20}	AF033857	1293/1592	mm	X97253	726/1025
(b)	FGA	0.28	0.37	5.0×10^{-14}	U36478	12780/13079	rn	X86561	1480/1779
	HNF1	0.31	0.31	4.5×10^{-70}	U73499	294/593	rr	X63959	14/313
	HNF1	0.31	0.31	8.4×10^{-65}	U73499	294/593	rn	X67649	3282/3581
	IGFBP1	0.41	0.45	4.3×10^{-18}	M23592	64/363	mm	X67493	493/792
	MTP	0.50	0.62	5.2×10^{-35}	S71339	509/808	cg	S74104	0/299
(c)	MIG6	0.49	0.50	2.0×10^{-20}	AL034417	76528/76827	rr	X07267	0/299
	BF	0.75	0.56	9.3×10^{-23}	AF019413	79680/79979	mm	AF109906	10767/11066
(d)	FABP2	0.29	0.58	4.3×10^{-18}	M18079	588/887	mm	M65033	285/584
	GUCA2B	0.40	0.49	2.4×10^{-16}	Z70295	768/1067	mm	AF006668	362/661
	HOXA4	0.48	0.48	6.4×10^{-96}	U41755	814/1113	mm	U75897	734/1033
	SLC34A1	0.27	0.39	2.0×10^{-20}	AF058289	1629/1928	mm	AF071876	3978/4277

Abbreviations: cg, *Cricetulus griseus;* mm, *Mus musculus;* rn, *Rattus norvegicus;* rr, *Rattus rattus.*

genes that are selectively expressed in the early stages of liver development.

In addition to the limitations in the available experimental data for other systems, the low-binding specificity of a subset of transcription factors poses a challenge for the construction of weight matrices, which form the basis of the computational approach described in this work. As more experimental data and more sensitive multiple local alignment algorithms become available (M. Tompa, pers. comm.; C. Workman, pers. comm.), this situation is expected to improve.

A more fundamental limitation lies in the fact that the model takes into account only the linear structure of the DNA, whereas properties related to protein–protein interactions and the chromatin structure are not addressed. The spacing between sites and the interaction between TF binding to adjacent sites can be incorporated in future models. Early efforts to incorporate such information have addressed pairs of well-studied TFs (Fickett 1996b) or have been restricted to evolutionarily related sets of genes or homologous retroviral regulatory regions (Frech et al. 1997). As more knowledge is accumulated, it may become possible to include information about chromatin structure.

Extending the current model to more tissues will require additional data. For both liver and skeletal muscle (Wasserman and Fickett 1998), two somewhat homogeneous tissues, abundant published experimental data was available. Very few other expression contexts offer such rich data resources, but it is possible to identify binding models for the critical TFs given a set of coregulated genes associated with a target context (Wasserman et al. 2000), and additional refinements of phylogenetic footprinting may allow the identification of likely modules to serve as training data.

In light of the recent progress in the study of transcriptional regulation, the abundant gene expression data, and ever-accumulating genomic sequences, sophisticated computational techniques are becoming indispensable for the elucidation of regulatory mechanisms.

METHODS

Position Weight Matrices

The coefficients $m_{\beta,i}$ of the PWMs were computed from the count matrices by

$$m_{\beta,i} = \log_2 \frac{\dfrac{n_{\beta,i} + \sqrt{N}/4}{N + \sqrt{N}}}{p_\beta}, \qquad (1)$$

in which $n_{\beta,i}$ is the count of nucleotide β at position *i, N* is the number of samples, and p_β denotes the background frequency of base β (assumed to be 0.25 for all four bases).

For our comprehensive jackknife test, we elected to use the Gibbs sampling program available at http://bayesweb.wadsworth.org/gibbs/gibbs.html to facilitate reproducibility of our procedure.

Logistic Regression

Standard commercial statistics packages are available for the use of LRA. We used a modified version of the logistic regression routine of the software package octave (a mathematical software package freely available from http://www.che.wisc.edu/octave/) to perform multivariate logistic regression. The octave logistic regression routine including documentation is available at http://www.cgb.ki.se/krivan/liver/liver.html.

Using a maximum likelihood procedure, the function

$$\pi(x) = \frac{e^{\beta_0 + \sum_{i=1}^{4} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{4} \beta_i x_i}} \qquad (2)$$

is fit to the 4 + 1 dimensional training data given by $\{(\mathbf{x}_j, y_j), j = 1, \ldots, N\}$, in which N is the number of positive and negative training sequences (N = 1820). For clarity, the logistic regression procedure is applied to a set of vectors containing the best score obtained with each transcription factor model on the collection of 200-bp sequences (positive and negative training sequences).

Phylogenetic Footprinting

We used the program DBA (DNA Block Aligner, Jareborg et al. 1999) with the default parameter settings to determine local sequence similarity between human and rodent DNA sequences in single-gene comparisons. The screening of sequences on a genomic scale was performed with BLAST (version 1.4.7), as the use of DBA would have been computationally impractical. The cutoff for the E value was set to $E = 10^{-10}$, and we used the parameters $M = 1$, $N = -2$.

Screening of Genomic Sequence

The pool of available genomic sequence from rodents (242 Mbp, minimum length 300 bp) was screened for putative liver regulatory modules (8492 detected). The rodent positives were masked for repetitive sequences and rescreened with the liver model (7192 modules retained). The putative modules were compared against the human genomic sequences available in the EMBL database to determine whether similar sequences were present (revealing 11,131 human entries similar to the rodent queries). The human sequences were screened with the liver model to restrict the set to regions that are conserved in putative function as well as sequence (314 human entries). In the final step, the set of human–rodent sequence pairs was consolidated (by intrinsic BLAST comparisons) to remove redundancy (147 final sequences).

Access to Data

Access to data as well as the submission of sequences for analysis with liver and muscle regulatory region models is possible via http://www.cgb.ki.se/krivan/liver/liver.html.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Grant No. PHY99-07949 and by funds from Pharmacia Corporation. W.K. acknowledges hospitality from the Institute for Theoretical Physics, University of California at Santa Barbara. We thank François Tronche and Joseph Locker for kindly providing manuscripts (Johnson and Williams 1994; Locker 2000). We thank James Fickett and Niclas Jareborg for helpful comments and suggestions. We also thank the anonymous referees for their helpful comments and valid criticism.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Bernard, P., Goudonnet, H., Artur, Y., Desvergne, B., and Wahli, W. 1999. Activation of the mouse TATA-less and human TATA-containing UDP-glucuronosyltransferase 1A1 promoters by hepatocyte nuclear factor 1. *Mol. Pharmacol.* **56:** 526–536.

Blackwood, E.M. and Kadonaga, J.T. 1998. Going the distance: A current view of enhancer action. *Science* **281:** 61–63.

Cereghini, S. 1996. Liver-enriched transcription factors and hepatocyte differentiation. *FASEB J.* **97:** 267–282.

Claverie, J.M. 2000. From bioinformatics to computational biology. *Genome Res.* **10:** 1277–1279.

Cooper, A.D., Chen, J., Botelho-Yetkinler, M.J., Cao, Y., Taniguchi, T., and Levy-Wilson, B. 1997. Characterization of hepatic-specific regulatory elements in the promoter region of the human cholesterol 7α-hydroxylase gene. *J. Biol. Chem.* **272:** 3444–3452.

Darlington, G.J. 1999. Molecular mechanisms of liver development and differentiation. *Curr. Opin. Cell Biol.* **11:** 678–682.

Doerksen, L.F., Bhattacharya, A., Kannan, P., Pratt, D., and Tainsky, M.A. 1996. Functional interaction between a RARE and an AP-2 binding site in the regulation of the human HOX A4 gene promoter. *Nucleic Acids Res.* **24:** 2849–2856.

Fickett, J.W. 1996a. Quantitative discrimination of MEF2 sites. *Mol. Cell. Biol.* **16:** 437–441.

———. 1996b. Coordinate positioning of MEF2 and myogenin binding sites. *Gene* **172:** GC19–GC32.

Fickett, J.W. and Wasserman, W.W. 2000. Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotechnol.* **11:** 19–24.

Frech, K., Danescu-Mayer, J., and Werner, T. 1997. A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. *J. Mol. Biol.* **270:** 674–687.

Garnier, G., Circolo, A., and Colten, H.R. 1996. Constitutive expression of murine complement factor B gene is regulated by the interaction of its upstream promoter with hepatocyte nuclear factor 4. *J. Biol. Chem.* **271:** 30205–30211.

Hagan, D.L., Kienzle, B., Jamil, H., and Hariharan, N. 1994. Transcriptional regulation of human and hamster microsomal triglyceride transfer protein genes. Cell type-specific expression and response to metabolic regulators. *J. Biol. Chem.* **269:** 28737–28744.

Hosmer, D.W. and Lemeshow, S. 1989. *Applied logistic regression.* John Wiley and Sons, New York, NY.

Hu, C.H., Harris, J.E., Davie, E.W., and Chung, D.W. 1995. Characterization of the 5′-flanking region of the gene for the α chain of human fibrinogen. *J. Biol. Chem.* **270:** 28342–28249.

Jareborg, N., Birney, E., and Durbin, R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9:** 815–824.

Johnson, P.F. and Williams, S.C. 1994. CCAAT/enhancer binding (C/EBP) proteins. In *Liver gene expression* (ed. F. Tronche and M. Yaniv), pp. 231–258. R.G. Landes Comp., Austin, TX.

Kaisaki, P.J., Menzel, S., Lindner, T., Oda, N., Rjasanowski, I., Sahm, J., Meincke, G., Schulze, J., Schmechel, H., Petzold, C., et al. 1997. Mutations in the hepatocyte nuclear factor-1α gene in MODY and early-onset NIDDM: Evidence for a mutational hotspot in exon 4. *Diabetes* **46:** 528–535.

Ktistaki, E. and Talianidis, I. 1997. Modulation of hepatic gene expression by hepatocyte nuclear factor 1. *Science* **277:** 109–112.

Locker, J. 2000. Tissue-specific regulation by transcription factors. In *Transcription factors* (ed. J. Locker), chapter 10, BIOS, London, UK.

Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., and Frazer, K.A. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288:** 136–140.

Melendez, P.A., Longo, N., Jimenez, B.D., and Cadilla, C.L. 2000. Insulin-induced gene 33 mRNA expression in Chinese hamster ovary cells is insulin receptor dependent. *J. Cell. Biochem.* **77:** 432–444.

Nolten, L.A., Steenbergh, P.H., and Sussenbach, J.S. 1995. Hepatocyte nuclear factor 1 α activates promoter 1 of the human insulin-like growth factor I gene via two distinct binding sites. *Mol. Endocrinol.* **9:** 1488–1499.

Roeder, R.G. 1996. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* **21:** 327–335.

Schneider, T.D., Stormo, G.D., Gold, L., and Ehrenfeucht, A. 1986. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188:** 415–431.

Stojanovic, N., Florea, L., Riemer, C., Gumucio, D., Slightom, J., Goodman, M., Miller, W., and Hardison, R. 1999. Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucleic Acids Res.* **27:** 3899–3910.

Stormo, G.D. and Fields, D.S. 1998. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.* **23:** 109–113.

Suwanichkul, A., Cubbage, M.L., and Powell, D.R. 1990. The promoter of the human gene for insulin-like growth factor binding protein-1. Basal promoter activity in HEP G2 cells depends upon liver factor B1. *J. Biol. Chem.* **265:** 21185–21193.

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.

Wasserman, W.W. and Fickett, J.W. 1998. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **278:** 167–181.

Wasserman, W.W., Palumbo, M., Thompson, W., Fickett, J.W., and Lawrence, C.E. 2000. Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet.* **26:** 225–228.

Whitaker, T.L., Steinbrecher, K.A., Copeland, N.G., Gilbert, D.J., Jenkins, N.A., and Cohen, M.B. 1997. The uroguanylin gene (Guca1b) is linked to guanylin (Guca2) on mouse chromosome 4. *Genomics* **45:** 348–354.

Yuh, C.H., Bolouri, H., and Davidson, E.H. 1998. Genomic cis-regulatory logic: Experimental and computational analysis of a sea urchin gene. *Science* **279:** 1896–1902.

Received January 18, 2001; accepted in revised form June 13, 2001.